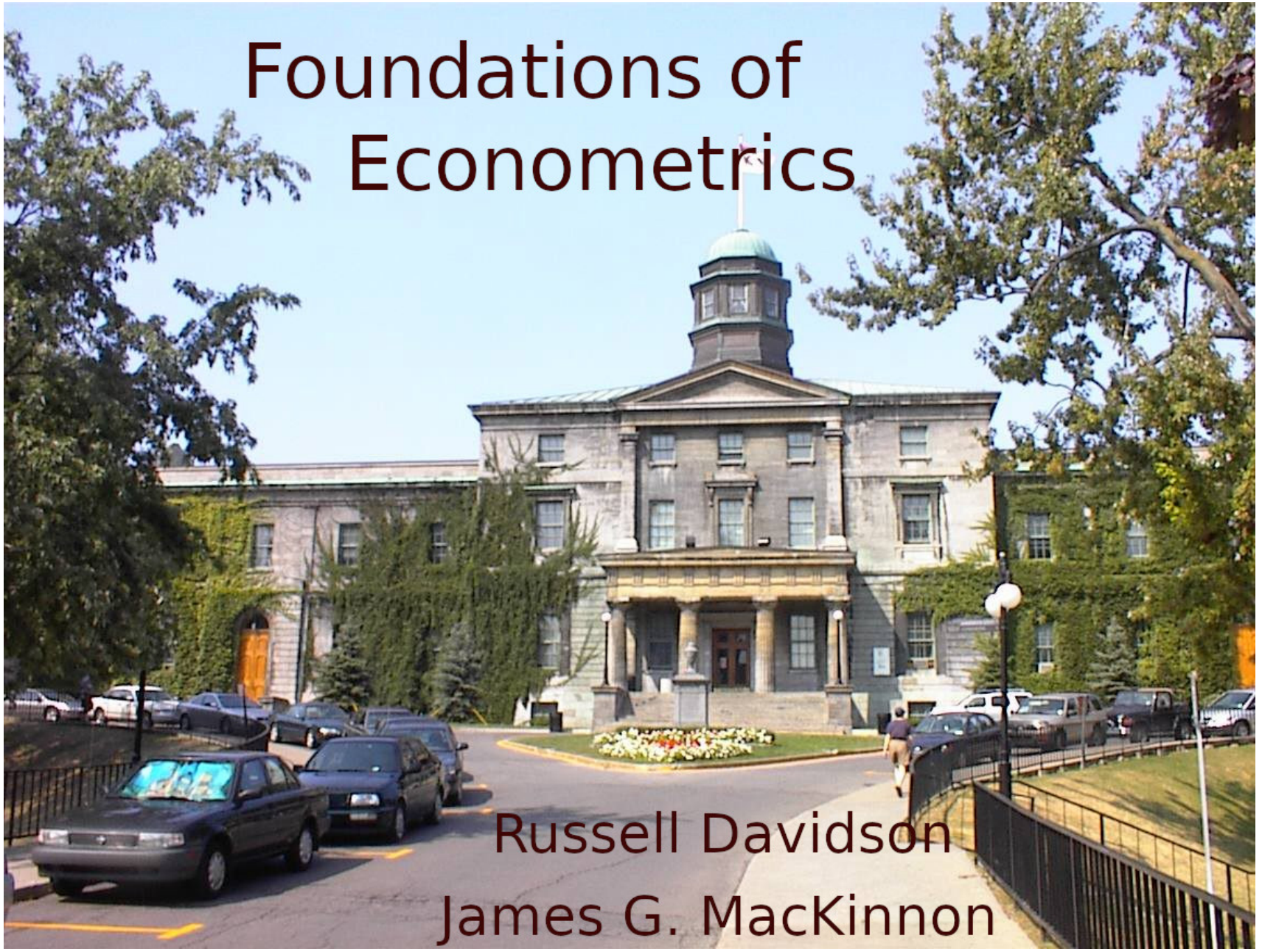

Foundations of
Econometrics
Russell Davidson
James G. MacKinnon

Contents

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Preface

This book is an updated and drastically shortened version of our 2004 textbook *Econometric Theory and Methods*. A plan to create a full second edition of that book never came to fruition, but the first several chapters of the book have served both of us well, not only as a text for a first, one-term, graduate course, but also for the Honours course in econometrics at McGill University. But even in those early chapters, there have been more and more things, over the years since the book was published, that we wish to update and change. It seemed quite feasible to do so if we limited ourselves to the chapters actually used in our one-term courses, and this book is the result.

How to Use This Book

This book is intended for a one-term course at either the Master's or Ph.D. level in a good graduate program, or even for serious final-year undergraduates, as at McGill, provided that the students have sufficient background, motivation, and ability.

Some of the exercises provided at the end of each chapter are really quite challenging, as we discovered many years ago while preparing solutions to them. These exercises are starred, as are a number of other exercises for which we think that the solutions are particularly illuminating, even if they are not especially difficult. In some cases, these starred exercises allow us to present important results without proving them in the text. In other cases, they are designed to allow instructors to cover advanced material that is not in the text itself. Because the solutions to the starred exercises should be of considerable value to students, they are available from the website for *Econometric Theory and Methods*, (ETM). All the data needed for the exercises are also available from the website, although these are necessarily not at all recent. Instructors might prefer to ask students to go online themselves to find more recent data that they can use instead of the older data on the website.

There are several types of exercises, intended for different purposes. Some of the exercises are empirical, designed to give students the opportunity to become familiar with a variety of practical econometric methods. Others involve simulation, including some that ask students to conduct small Monte Carlo experiments. Many are fairly straightforward theoretical exercises that good students should find illuminating and, we hope, not too difficult. Although most of the exercises have been taken over unchanged from ETM, some have been modified, and a fair number of new exercises introduced.

An instructor's manual was provided for ETM, with solutions to all the exercises of that book. It can be found online by anyone willing to spend a little time with Google. In a sense this is a shame, as it means that instructors

can no longer safely use exercises given here for exams or assignments, since some students may be tempted to copy the solutions from the manual. We fear that this is likely to be a problem that university instructors will have to face more and more frequently, at a time when a lot of instruction is being given online. However, for our purposes here, the most important point is that solutions for the starred exercises are readily available without access to Google or any other search engine.

Background

Simulation-based methods greatly enhance the asymptotic theory that has been at the heart of econometrics for many decades. Problems that are intractable analytically are often simple to handle by simulation, and many new techniques that exploit this fact have been proposed during the last three or four decades. Of these techniques, the one that seems most general in its application is the bootstrap, and for this short book we have written a new chapter dedicated to this important topic.

Estimating functions and estimating equations are topics that are not terribly familiar to most econometricians. We ourselves became aware of them only in the mid-1990s, when the late V. P. Godambe, then of the University of Waterloo in Ontario, prodded us to look more closely at a theme that he had himself pioneered back in the 1960s. Estimating equations provide a unified method for studying all of the estimation techniques discussed in this book, and many more besides, including the widely-used generalized method of moments (GMM).

We have tried hard to present material in a logical way, introducing new ideas as they are needed and building on the material developed earlier. This applies also to mathematical and statistical techniques that, in many cases, students may already be somewhat familiar with. Instead of treating these in appendices, we discuss them when they are first used, in the context of their applications to econometrics. We have found that this approach generally works very well in the real or virtual classroom. The sections on mathematics or statistics are never too long, and we make every effort to motivate them by indicating their relevance to econometrics. This probably means that the book is not appropriate for students with a really weak mathematical or statistical background.

While it is generally not hard to develop a consistent and appropriate notation for an individual topic, we found it exceedingly hard to maintain notation consistent across all the chapters of a book of the length of ETM. We do not claim to have succeeded there or here, but we have made strenuous efforts in that direction. We have been influenced by suggestions on many vexed points of notation from Karim Abadir, of Imperial College London in England, and Jan Magnus, of Tilburg University and the Free University of Amsterdam in the Netherlands. Although we have not followed their counsel in all cases,

Copyright © 2021, Russell Davidson and James G. MacKinnon

we wholeheartedly support their efforts to develop a useful and consistent notation for modern econometrics.

Organization

The book covers a number of fundamental concepts of estimation and statistical inference, beginning with ordinary least squares (OLS). Subsequently, we introduce the extremely important instrumental variables (IV) estimator, and the generalized least squares (GLS) estimator. For anything to do with nonlinear estimation, including maximum likelihood or GMM, none of which is treated here, the reader may have recourse to the full ETM book.

The first chapter is new. It contains a somewhat philosophical discussion, and it may well be that some people will disagree with the point of view adopted there. Textbooks and scientific papers normally do not broach philosophical questions, and some scientists have been known to express the opinion that philosophy is a waste of time for practicing scientists. Obviously we disagree, but here we give fair warning that people with a different cast of mind may omit this preliminary chapter, and lose nothing of conventional econometrics by so doing.

Most of Chapter 2 is fairly elementary, and much of the material in it should already be familiar to students who have a good background in statistics, econometrics, and matrix algebra at the undergraduate level. The discussion of how to simulate a regression model in Section 2.3 introduces some concepts that are not often taught in undergraduate econometrics courses but are crucial to understanding bootstrap methods. Section 2.5 is even more important, because it treats linear regression using estimating equations, a topic that is probably quite new to most students.

Chapter 3, a fundamental chapter, deals with the geometry of least squares in some detail, and relates it to the algebraic account that is probably more familiar to many students and instructors. Not all instructors find the geometrical approach quite as intuitive as we do. However, our experience is that many students do find it extremely helpful. The chapter introduces a number of fundamental concepts that reappear many times in various places. These include the application of Pythagoras' Theorem to ordinary least squares, the subspace spanned by the columns of a matrix of regressors and its orthogonal complement, orthogonal projection matrices, and the Frisch-Waugh-Lovell (FWL) Theorem. Some applications of the theorem are introduced here, including a section on the important topic of leverage and influential observations.

Chapter 4 is also a fundamental chapter. It deals with the statistical properties of the ordinary least squares (OLS) estimator and introduces such important concepts as unbiasedness, probability limits, consistency, covariance matrices, efficiency, the Gauss-Markov Theorem, the properties of residuals, and the consequences of model misspecification. Students with a strong background in statistics should be at least somewhat familiar with much of this material.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Statistical inference is first dealt with in Chapter 5. The nature of hypothesis testing is laid out, and the most commonly used tests are discussed. For the special case of the classical normal linear model, there exist exact results about the distribution of the test statistics, but, more generally, it is necessary to have recourse to asymptotic theory. This chapter contains two sections that are new in this book, one on performing tests of several hypotheses simultaneously, and one on pretesting.

Chapter 6 continues the story of statistical inference. The first three sections cover confidence intervals and confidence regions; the next three deal with covariance matrix estimation in some circumstances in which the disturbances are not independent and identically distributed. The robust covariance matrix estimators discussed are the HCCME (heteroskedasticity-consistent covariance estimator), HAC (heteroskedasticity and autocorrelation consistent) estimators, and the CRVE (cluster-robust variance estimator). The next section deals with the important difference-in differences technique of estimation, and the final section of this chapter explains the delta method, as a way to estimate the covariance matrix of a set of nonlinear functions of parameter estimates.

Chapter 7, on the bootstrap, is not entirely new, since the bootstrap is mentioned in numerous places in ETM. Here, however, we have collected the material on bootstrapping in ETM, and added a good deal more. The bootstrap is mentioned again in the following two chapters, but it seemed important not to postpone consideration of the bootstrap until after treating all the other topics in the book. In this chapter, we start with bootstrap hypothesis tests, along with bootstrap P values, and proceed to the study of bootstrap confidence sets. Bootstrapping is also discussed in cases in which the robust covariance matrix estimators of the previous chapter should be used.

Chapter 8 introduces estimation by instrumental variables (IV). When instrumental variables are used, techniques of inference are considerably different from those previously presented for least squares. The special cases of tests for over-identifying restrictions, and Durbin-Wu-Hausman (DWH) tests, each receive a section, as does bootstrapping models estimated by IV.

The final chapter, Chapter 9, embarks on the study of generalized least squares (GLS). Again, techniques of estimation and inference have to be adapted to this context. There is discussion of heteroskedasticity, both testing for its absence, and efficient estimation when the pattern of heteroskedasticity is known. It is natural that this should be followed by discussion of autocorrelation. Tests, both old and relatively new, with a null hypothesis of no autocorrelation are presented, and then techniques of estimating models with autocorrelated disturbances. A final section applies the ideas of GLS to an introductory treatment of panel data models.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Acknowledgements

It took us nearly six years to write ETM, the principal source for this book, and during that time we received assistance and encouragement from a large number of people. Bruce McCullough, of Drexel University, read every chapter, generally at an early stage, and made a great many valuable comments and suggestions. Thanasis Stengos, of the University of Guelph, also read the entire book and made several suggestions that have materially improved the final version. Richard Parks, of the University of Washington, taught out of the unfinished manuscript and found an alarming number of errors that were subsequently corrected. Others who read at least part of the book with care and provided us with useful feedback include John Galbraith (McGill University), the late Ramazan Gencay (University of Windsor), Sílvia Gonçalves (McGill University), Manfred Jäger (Martin Luther University), Richard Startz (University of Washington), Arthur Sweetman (McMaster University), and Tony Wirjanto (University of Waterloo). There are numerous other colleagues who also deserve our thanks, but the list is much too long to include here.

We made use of draft chapters of ETM in courses at both the Master's and Ph.D. levels at Queen's University, McGill University, and the University of Toronto in Canada, as well as at the Université de la Méditerranée and GREQAM in France. A great many students found errors or pointed out aspects of the exposition that were unclear. The book has been improved enormously by addressing their comments, and we are very grateful to all of them. We are also grateful to the many students at the above institutions, and also at the University of Washington, who expressed enthusiasm for the book when it was still a long way from being finished and thereby encouraged us to finish a task that, at times, seemed almost incapable of completion.

This book has seen much less use than has ETM as of this writing (December 2020), but is currently in heavy use for online teaching at McGill. This experience has led to numerous improvements in the exposition, for many of which we thank the students who were taking the course online.

We also owe a debt of gratitude to the thousands of talented programmers who have contributed to the development of free, open-source, software, without which it would have been much more difficult to write this book. The book was typeset entirely by us using the TEX LIVE distribution of TEX running on the Debian distribution of the Linux operating system. We also made extensive use of the gcc and g77 compilers, and of many other excellent free programs that run on Linux.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Notation

We have tried our best to use a consistent set of notation throughout the book. It has not always been possible to do so, but below we list most of the notational conventions used in the book.

n	sample size; number of observations
k	number of regressors
l	number of instrumental variables
$\boldsymbol{A}$ (upper-case bold letter)	a matrix or row vector
$\boldsymbol{a}$ (lower-case bold letter)	a column vector
$\boldsymbol{y}$	dependent variable; regressand
$\boldsymbol{X}$	matrix of explanatory variables; regressors
$\boldsymbol{W}$	matrix of instrumental variables
$\boldsymbol{\beta}$	vector of regression parameters
$\hat{\boldsymbol{\beta}}$	vector of estimated parameters
$\tilde{\boldsymbol{\beta}}$	vector of parameters estimated under restrictions
$\boldsymbol{u}$	vector of disturbances
σ^2	variance (usually of disturbances)
F	cumulative distribution function (CDF)
f	probability density function
Φ	CDF of standard normal distribution N(0,1)
ϕ	density of standard normal distribution
$\mathbf{I}$	identity matrix
$\mathbf{0}$	column vector of zeros
$\mathbf{O}$	matrix of zeros
$\boldsymbol{\iota}$	column vector of ones
$\mathcal{S}(\boldsymbol{X})$	linear span of the columns of $\boldsymbol{X}$
$\mathcal{S}^{\perp}(\boldsymbol{X})$	orthogonal complement of $\mathcal{S}(\boldsymbol{X})$
$\boldsymbol{P_X}$	orthogonal projection matrix on to $\mathcal{S}(\boldsymbol{X})$
$\boldsymbol{M_X}$	complementary orthogonal projection; $\boldsymbol{M_X} = \mathbf{I} - \boldsymbol{P_X}$
μ	a data-generating process (DGP)
$\mathbb{M}$	model, a set of DGPs
$\mathbb{R}$	the real line
$\mathbb{R}^n$	set of n–vectors
E^n	n–dimensional Euclidean space
E	expectation operator
Var	a variance or a covariance matrix
$\boldsymbol{\Omega}$	a covariance matrix
$\boldsymbol{\Gamma}(j)$	autocovariance matrix at lag j
L	lag operator
$\overset{a}{=}$	asymptotic equality
$\overset{a}{\sim}$	asymptotically distributed as
$\xrightarrow{d}$	convergence in distribution

Data, Solutions, and Corrections

The website for ETM, and for this book, is located at

http://qed.econ.queensu.ca/ETM/

This website provides all the data needed for the exercises, solutions to the starred exercises, and corrections made since the book was printed. The solutions and corrections are provided as PDF files, which almost all modern computers should have the software to view and print.

The website provides solutions only to the starred exercises. In addition, there is an instructor's manual, available to instructors only, which comes on a CD-ROM and contains solutions to all the exercises. For information about how to obtain it, please contact your local Oxford University Press sales representative or visit the Oxford higher education website at

http://www.oup-usa.org/highered

The authors are happy to hear from readers who have found errors that should be corrected. Their affiliations are given below:

Russell Davidson
<russell.davidson@mcgill.ca>

Aix-Marseille Université
CNRS, EHESS, AMSE
13205 Marseille cedex 01
France

Department of Economics
McGill University
855 Sherbrooke St. West
Montreal, Quebec, H3A 2T7
Canada

James G. MacKinnon
<jgm@econ.queensu.ca>

Department of Economics
Queen's University
Kingston, Ontario, K7L 3N6
Canada

Chapter 1

Models in Science

1.1 Introduction

A major contention in this chapter is that scientific models can be viewed as virtual realities, implemented, or rendered, by mathematical equations or by computer simulations. Their purpose is to explain, and help us understand the external reality that they model.

In economics, particularly in econometrics, models make use of random elements, so as to provide quantitatively for phenomena that we cannot or do not wish to model explicitly. By varying the realisations of the random elements in a simulation, it is possible to study counterfactual outcomes, which are necessary for any discussion of causality.

1.2 Scientific Models as Virtual Reality

The capacities of modern computers have made **virtual reality** something that we can experience in new ways, enabled by new technology. We hear of flight simulators, and the younger generation seems to spend a lot of time in the virtual reality of computer games. But people have been inventing virtual realities for as long as there have been scientists.

In most scientific disciplines, **models** play an essential role. Scientific models are often mathematical, but they need not be so. A mathematical model does, however, make clear the sense in which a model is a sort of virtual reality. Mathematics is nothing if not an abstract discipline; so much so that some have claimed that mathematics, pure mathematics anyway, has no meaning or substantive content. What is true, though, is that we can *give* mathematical constructions interpretations that imply much substantive content. This is just as true in economics as it is in the physical sciences.

Why is this? The aim of science is not only to acquire knowledge of the world, although the etymology of the word "science" implies only knowledge, but to *understand* the world. Science provides such understanding by *explaining* our experiences. Science advances when it provides better explanations. What constitutes an explanation? Well, a theory. That's just terminology, and so the question has merely been reformulated as: What constitutes a theory?

A theory is embodied in a model, and the model constitutes a virtual reality. But not all models count as theories, as we will explain later. However, we can conclude at present that virtual realities can give us understanding of the world, through the explanations that they may embody. Of course, some models mimic external reality, as we observe it, better than others, and so they provide better explanations. Scientific controversies are about which explanations are better.

What is there about a theory that provides an explanation? Think, if you will, of Keynes's celebrated *General Theory*. The theory implies a model of the macroeconomy, the macroeconomy in virtual reality, and within this model, there are relationships among the macroeconomic variables – relations that can be expressed mathematically, and are justified by the arguments that Keynes makes, showing that these relations mimic what we observe of the macroeconomy. When we observe that interest rates fall, the Keynesian model explains the economic mechanisms that led to this fall.

Not every economist is convinced by Keynesian explanations! The opponents of Keynes's model, or his view of the world, if we are to pay any attention to them, must construct *rival* virtual realities, and argue that the relations that these entail describe external reality better than the Keynesian ones.

The point about virtual reality is made very clearly in the sentences below, taken from Deutsch (1997).

> The fact that virtual reality is possible is an important fact about the fabric of reality. It is the basis not only of computation, but of human imagination and external experience, science and mathematics, art and fiction.
>
> David Deutsch, *The Fabric of Reality*
>
> The heart of a virtual-reality generator is its computer.
>
> *ibid.*

Deutsch goes on to make a different point about virtual reality, namely the possibility of rendering it physically, and this takes us too far from our concerns as econometricians. However, mathematics can constitute virtual reality as well as computers can. But, as our computers have become more powerful, so our models depend more and more on computer implementations. There are deep philosophical questions concerning whether we, as humans, can really understand something produced by computation rather than logical and mathematical reasoning, especially if one looks forward to what quantum computers may one day be able to do, and can do, in principle, according to the physical theories we have today.

But one thing we can easily say about models implemented on the computer is that everything must be *digital*, and so also *discrete*. That this is no real problem for practical things is evident from the extent we use digital sound recording, digital cameras and so on, and especially digital typography, one of the greatest boons for anyone writing books or papers, like this one. We

Copyright © 2021, Russell Davidson and James G. MacKinnon

speak of presenting a "paper", although it need never be printed on paper at all. What exists in the real world is a representation in virtual reality of a hardcopy paper. Similarly, we often speak of the "slides" for a presentation, although they are just as virtual.

There is in fact no consensus at the present time among theoretical physicists as to whether space-time is continuous, although this is assumed by most current physical models, or rather discrete - quantized, like everything else in quantum mechanics. It follows that there is no harm in letting our virtual realities be discrete – whether or not they are digital – and there may be considerable benefits from doing so.

Models in Economics

Just as in physics, many economic models assume that space and time are continuous, although in econometrics, for obvious reasons, time, at least, is usually treated as a discrete variable. Unlike many physical models however, econometric models invariably incorporate **random elements**.

There is a considerable philosophical difficulty that arises when we wish to impart any substantive meaning to the mathematics of probability and random variables, *if* we also wish to adhere to a deterministic world view. This is so because, in conventional interpretations of probability, events that have occurred, are occurring, or will (certainly) occur have a probability of one, and events that never occur have a probability of zero. If, as follows from a deterministic view, any event at all either does occur or does not, the mathematics of probability becomes trivial.

But we use probabilistic notions all the time, and not trivial ones either. What in the external world is it that we want to mimic by using randomness? We can all agree that many things in our lives appear to us to be random, but there are many philosophers who, while granting this *appearance* of randomness, still think that, at some fundamental level, the world is deterministic. This leads to a somewhat deeper question. *Why* are there such seemingly random events? To that question, a possible answer is that we model such events as realisations of random variables because we do not and cannot know everything. Even more to the point, we cannot *explain* everything. Whenever we cannot, or do not wish to, explain events which have an impact on the main objects of interest in our models, we model them as realizations of random variables. That at least is our view of what we do as econometricians, although many other econometricians may well either disagree or else express things quite differently.

If we adopt this point of view about why there are random elements in economic models, then we see why it is of interest to perform simulations with random numbers. Yes, the goal of our models is to understand through explanation, and calling things random explains nothing, but, even so, models with random elements can help us understand economic phenomena by giving

Copyright © 2021, Russell Davidson and James G. MacKinnon

partial explanations of economic mechanisms. Another conclusion from this reasoning is that some virtual realities may be quite imperfect renderings of the real world. Maybe flight simulators are pretty good these days, but they weren't always, and video games don't even try to mimic the real world.

It is not enough to wave our hands and say that we use random elements in our models. We need more than that if we want to consider a model as a virtual reality, probably one to be rendered by the computer. The best way to formulate this is to define a **model** as a set of **data-generating processes**, or DGPs, each of which constitutes a *unique* virtual reality. We can go further, and specify that a DGP is something that can be simulated on the computer, or that provides a **unique recipe** for simulation. In this way, we tie the virtual realities of economic models more closely to the computer, just as Deutsch would have it.

What has been missing and now must be introduced is the **distribution** of the random elements. Computers have **random-number generators**, or RNGs, and what they generate are sequences of **random numbers**. Such a sequence has most of the mathematical properties of a sequence of mutually independent realizations from the uniform distribution on the interval $[0, 1]$. See Knuth (1998), Chapter 3 for a very thorough discussion of this point. One property not shared by a sequence generated by a computer and a sequence that satisfies the mathematical requirements of realizations from the uniform distribution is that the elements of a computer-generated sequence are rational numbers that can be expressed with a finite number of bits, whereas a realisation of the uniform distribution may be any real number in the $[0, 1]$ interval. However, this and all other differences between what the computer generates and the mathematical ideal have no bad consequences for the simulations needed in econometrics.

Random numbers can be transformed into realizations from distributions other than the uniform. A valuable reference for many of these transformations is Devroye (1986). Thus we can incorporate any desired form of randomness that we can specify into the DGPs of a model.

Another feature of most economic models is that they involve **parameters**. A model normally does not specify the numerical values of these parameters; indeed a purely parametric model is a set rather than a singleton because the DGPs that it contains may differ in the values of their parameters. Models that are not purely parametric allow the DGPs that they contain to differ also in the **stochastic specification**, that is, the distribution of the random elements.

1.3 Causal Explanations

Suppose that we have a model of an economic phenomenon that we wish to study. Suppose, too, that it seems to correspond well to what we observe

Copyright © 2021, Russell Davidson and James G. MacKinnon

in external reality. Does that mean that we have *explanations*, complete or partial, of what we are studying? Not necessarily. Some models are purely descriptive. A statistical model, for instance, might specify the probabilistic properties of a set of variables, and nothing more. But that may be enough for us to do forecasting, even if our forecasts are not based on any profound understanding. Half a century ago, most physicists thought of quantum mechanics that way, as a mathematical recipe that could be used to predict experimental results. The "interpretations" of quantum mechanics that were then current were very counter-intuitive, and today physicists still argue not only about what interpretation is to be preferred, but about whether *any* interpretation meaningful to the human brain is possible.

However, the positivist approach that has held sway in physics for so long is finally giving way to a thirst for explanations. Perhaps theoretical physics does give better agreement with experimental data than any other discipline, but, some physicists are now asking, does it constitute a true *theory*? A theory must explain, by proposing a mechanism, or in other words a *causal* chain.

What is a cause?

This subsection draws heavily on the insights in Chapter 3 of Dennett (2003). Consider two events, A and B. An intuitive definition of the proposition that A causes B is:

(i) A and B are real, or true;

(ii) If A is not real or true, then neither is B; and

(iii) A precedes B in time.

This definition raises a number of issues. What do we mean by an "event"? There are several admissible answers: an action, a fact of nature, among others. A fact is true or not, and action is performed (it is real) or not. Our tentative definition is general enough to allow for various different possibilities.

In order to steer clear of some trivial cases, we want to suppose that the events A and B are logically *independent*. Thus we don't want to say that the conclusion of a mathematical theorem is *caused* by the premisses of the theorem.

It is important to distinguish between causal **necessity** and causal **sufficiency**. Necessity means that:

not A (written as $\neg A$) implies $\neg B$.

In words, without A, there can be no B. Logically, the condition is equivalent to the condition that B implies A; that is, A is a *necessary* condition for B. This is our condition (ii).

Sufficiency means that:

A implies B, or $\neg B$ implies $\neg A$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

In words, every time that A holds, unavoidably B holds as well; that is, A is a *sufficient* condition for B. Sufficiency is logically quite distinct from necessity. Necessity leaves open the possibility that A holds without B. Sufficiency leaves open the possibility that B holds without A.

It is easy enough to see how we might study these two types of causality when the events A and B are repeated, as with coin tosses or the roulette wheel, where we don't *a priori* expect to find any causality at all, or when an experiment is undertaken in which both A and $\neg A$ can occur, and possibly also B and $\neg B$.

But if A and B are unique, not repeated, events, what sense can we make of the assertion that A caused B? I suppose here that condition (i) is satisfied, so that A and B both occurred. In order to make any sense of the statement about causality, we have to admit to our discussion *imaginary worlds* or even *universes*. We call such worlds or universes **counterfactual**. Without considering them, it is impossible to know what *might* have occurred if A did not, or if B did not occur.

But this remark gives rise to as many problems as answers. What is the set of universes that these counterfactual universes inhabit? How can we delimit this set? Let's denote the set by $\mathcal{X}$. Then we have a number of reasonable choices:

(a) $\mathcal{X}$ is the set of *logically* possible universes, that is, all universes that are not logically self-contradictory;

(b) $\mathcal{X}$ is the set of universes compatible with the laws of physics, as we know them;

(c) $\mathcal{X}$ is the set of logically and physically admissible universes that are sufficiently *similar* or *close* to the real world.

The last choice is no doubt the best, but, in order to implement it, what can we mean by saying that a counterfactual universe is in a *neighbourhood* of the real one?

Counterfactual econometrics

In biostatistics and medicine, emphasis is often put on **randomized trials**, in which two groups of subjects are treated differently. One usually speaks of a control group, the members of which are not **treated**, and a treatment group, for which a particular treatment is prescribed. After some definite period, the members of both groups are examined for some particular property, which is thought of as the effect of being treated or not. Clearly, the idea is to be able to see whether the treatment causes the effect, and, perhaps, to reject the hypothesis that it does so. Here, if one can select the members of the two groups quite randomly, in a way totally unrelated to the treatment or

Copyright © 2021, Russell Davidson and James G. MacKinnon

the effect, then the distribution of effects within each group serves as the counterfactual distribution for the other.

Even in medicine, a truly randomized trial can be difficult to achieve, for both practical and ethical reasons. In econometrics, it is even more difficult, although not completely impossible. However, "natural experiments" can arise for which an econometrician may be able to identify two groups that are "treated" differently, perhaps by being subject to some government programme, and to measure some effect, such as wages, that might be affected by the treatment. This can be fruitful, but, naturally enough, it requires the use of sophisticated statistical and econometric techniques.

In a polemical essay, Heckman (2001) maintains that econometrics has suffered as a result of too great an application of the methodology of mathematical statistics. He says that

> *Statistics is strong in producing sampling theorems and in devising ways to describe data. But the field is not rooted in science, or in formal causal models of phenomena, and models of behavior of the sort that are central to economics are not a part of that field and are alien to most statisticians.*

This is a strong statement of what we call the preference of econometricians for **structural models**.

Whether or not they go along completely with Heckman on this point, econometricians, even sometimes in company with statisticians, have developed techniques for getting indirectly at information about counterfactual worlds. Of these, the method called **difference in differences** is probably the best known and the most used; see Section 6.7. Since counterfactual worlds are never realized, *some* assumptions must *always* be made in order to invent a virtual reality in which they can be rendered. Often, an assumption is made implying constancy in time of some relations; other times the assumption might be, as with randomized trials, that two or more groups are homogeneous. To say that we always need some assumption(s) is to say that there must always be a model, rich enough in its explanatory power to render credible counterfactual, and so virtual, realities.

One development of this sort is found in Athey and Imbens (2006). They extend the idea behind the difference-in-differences method to a method called change-in-changes. The name does not make clear what we regard as the chief virtue of their method, namely that, instead of limiting attention to *average* treatment effects, it considers the entire distribution of these effects. Average effects may be enough for biostatisticians; not for econometricians.

Statistical Models

Even if one is content with models that are purely descriptive rather than structural, an important point is that *any* statistical model must have some

Copyright © 2021, Russell Davidson and James G. MacKinnon

structure, that is, it must make some assumptions whereby the DGPs in the model are restricted, before any sort of statistically valid inference is possible. Even proponents of structural models try to avoid making any more assumptions than they must, because restrictions that do not correspond to external reality lead to models that cannot hope to mimic it at all closely. But in a celebrated paper, Bahadur and Savage (1956) show that even as simple a task as estimating a population mean by use of a sample mean drawn as an independent and identically distributed sample from the population, is impossible without further structure on the set of probability distributions allowed by the model to describe the distribution of the population. The restrictive assumption that the population mean exists as a mathematical expectation is not enough by itself.

In fact, those necessary restrictions are surprisingly strong. That is not to say that they are hard to satisfy in empirical work, just that it is rather counter-intuitive that they should be logically necessary for any valid statistical inference. It is in fact rather easy to give a mathematical construction of DGPs for which the usual statistical techniques of estimation and inference fail completely. Fortunately, these DGPs are "pathological", in the sense that, while they are quite coherent mathematically (in the virtual reality of mathematics), they do not correspond in any way to our "common-sense" view of how the world works.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 2

Regression Models

2.1 Introduction

Regression models form the core of the discipline of econometrics. Although econometricians routinely estimate a wide variety of statistical models, using many different types of data, the vast majority of these are either regression models or close relatives of them. In this chapter, we introduce the concept of a regression model, discuss several varieties of them, and introduce the estimation method that is most commonly used with regression models, namely, **ordinary least squares**. We derive this method in two different ways, first, by use of **estimating functions**, and then, in the traditional way, by least squares. Both are very general principles of estimation, which have many applications in econometrics.

The most elementary type of regression model is the **simple linear regression model**, which can be expressed by the following equation:

$$y_t = \beta_1 + \beta_2 x_t + u_t. \tag{2.01}$$

The subscript t is used to index the **observations** of a **sample**. The total number of observations, also called the **sample size**, will be denoted by n. Thus, for a sample of size n, the subscript t runs from 1 to n. Each observation comprises an observation on a **dependent variable**, written as y_t for observation t, and an observation on a single **explanatory variable**, or **independent variable**, written as x_t.

The relation (2.01) links the observations on the dependent and the explanatory variables for each observation in terms of two unknown **parameters**, β_1 and β_2, and an unobserved **disturbance**, or **error term**, u_t. Thus, of the five quantities that appear in (2.01), two, y_t and x_t, are observed, and three, β_1, β_2, and u_t, are not. Three of them, y_t, x_t, and u_t, are specific to observation t, while the other two, the parameters, are common to all n observations.

Here is a simple example of how a regression model like (2.01) could arise in economics. Suppose that the index t is a time index, as the notation suggests. Each value of t could represent a year, for instance. Then y_t could be household consumption as measured in year t, and x_t could be measured disposable income of households in the same year. In that case, (2.01) would represent what in elementary macroeconomics is called a **consumption function**.

If for the moment we ignore the presence of the disturbances, β_2 is the **marginal propensity to consume** out of disposable income, and β_1 is what is sometimes called **autonomous consumption**. As is true of a great many econometric models, the parameters in this example can be seen to have a direct interpretation in terms of economic theory. The variables, income and consumption, do indeed vary in value from year to year, as the term "variables" suggests. In contrast, the parameters reflect aspects of the economy that do not vary, but take on the same values each year.

The purpose of formulating the model (2.01) is to try to explain the observed values of the dependent variable in terms of those of the explanatory variable. According to equation (2.01), the value of y_t is given for each t by a linear function of x_t and the unobserved disturbance u_t. The linear (strictly speaking, affine[1]) function, which in this case is $\beta_1 + \beta_2 x_t$, is called the **regression function**. At this stage we should note that, as long as we say nothing about the unobserved disturbance u_t, equation (2.01) does not tell us anything. In fact, we can allow the parameters β_1 and β_2 to be quite arbitrary, since, for any given β_1 and β_2, the model (2.01) can always be made to be true by defining u_t suitably.

If we wish to make sense of the regression model (2.01), then, we must make some assumptions about the properties of the disturbance u_t. Precisely what those assumptions are will vary from case to case. In all cases, though, it is assumed that u_t is a **random variable**. Most commonly, it is assumed that, whatever the value of x_t, the expectation of the random variable u_t is zero. This assumption usually serves to **identify** the unknown parameters β_1 and β_2, in the sense that, under the assumption, equation (2.01) can be true only for specific values of those parameters.

The presence of disturbances in regression models means that the explanations these models provide are at best partial. This would not be so if the disturbances could be directly observed as economic variables, for then u_t could be treated as a further explanatory variable. In that case, (2.01) would be a relation linking y_t to x_t and u_t in a completely unambiguous fashion. Given x_t and u_t, y_t would be completely explained without error.

Of course, disturbances are not observed in the real world. They are included in regression models because we are not able to specify all of the real-world factors that determine the value of y_t. When we set up a model like (2.01) with u_t as a random variable, what we are really doing is using the mathematical concept of randomness to model our *ignorance* of the details of economic mechanisms. When we suppose that the expectation of a disturbance is zero, we are implicitly assuming that the factors determining y_t that we ignore are just as likely to make y_t bigger than it would have been if those factors were absent as they are to make y_t smaller. Thus we are assuming that, on

[1] A function $g(x)$ is said to be **affine** if it takes the form $g(x) = a + bx$ for two real numbers a and b.

Copyright © 2021, Russell Davidson and James G. MacKinnon

average, the effects of the neglected determining factors tend to cancel out. This does not mean that those effects are necessarily small. The proportion of the variation in y_t that is accounted for by the disturbances will depend on the nature of the data and the extent of our ignorance. Even if this proportion is large, and it can be very large indeed in some cases, regression models like (2.01) can be useful if they allow us to see how y_t is related to the variables, like x_t, that we can actually observe.

Much of the literature in econometrics, and therefore much of this book, is concerned with how to estimate, and test hypotheses about, the parameters of regression models. In the case of (2.01), these parameters are the **constant term**, or **intercept**, β_1, and the **slope coefficient**, β_2. Although we will begin our discussion of estimation in this chapter, most of it will be postponed until later chapters. In this chapter, we are primarily concerned with understanding regression models as statistical models, rather than with estimating them or testing hypotheses about them.

In the next section, we review some elementary concepts from probability theory, including random variables and their expectations. Many readers will already be familiar with these concepts. They will be useful in Section 2.3, where we discuss the meaning of regression models and some of the forms that such models can take. In Section 2.4, we review some topics from matrix algebra and show how multiple regression models can be written using matrix notation. Finally, in Section 2.5, we introduce the method of moments and show how it leads to ordinary least squares as a way of estimating regression models.

2.2 Distributions, Densities, and Moments

The variables that appear in an econometric model are treated as what mathematical probabilists call **random variables**. In order to characterize a random variable, we must first specify the set of all the possible values that the random variable can take on. The simplest case is a **scalar random variable**, or **scalar r.v.** The set of possible values for a scalar r.v. may be the real line or a subset of the real line, such as the set of nonnegative real numbers. It may also be the set of integers or a subset of the set of integers, such as the numbers 1, 2, and 3.

Since a random variable is a collection of possibilities, random variables cannot be observed as such. What we do observe are **realizations** of random variables, a realization being one value out of the set of possible values. For a scalar random variable, each realization is therefore a single real value.

If X is any random variable, **probabilities** can be assigned to subsets of the full set of possibilities of values for X, in some cases to each point in that set. Such subsets are called **events**, and their probabilities are assigned by a **probability distribution**, according to a few general rules.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Discrete and Continuous Random Variables

The easiest sort of probability distribution to consider arises when X is a **discrete random variable**, which can take on a finite, or perhaps a countably infinite, number of values, which we may denote as $x_1, x_2, \ldots$. The probability distribution simply assigns probabilities, that is, numbers between 0 and 1, to each of these values, in such a way that the probabilities sum to 1:

$$\sum_{i=1}^{\infty} p(x_i) = 1,$$

where $p(x_i)$ is the probability assigned to x_i. Any assignment of nonnegative probabilities that sum to one automatically respects all the general rules alluded to above.

In the context of econometrics, the most commonly encountered discrete random variables occur in the context of **binary data**, which can take on the values 0 and 1, and in the context of **count data**, which can take on the values 0, 1, 2,....

Another possibility is that X may be a **continuous random variable**, which, for the case of a scalar r.v., can take on any value in some continuous subset of the real line, or possibly the whole real line. The dependent variable in a regression model is normally a continuous r.v. For a continuous r.v., the probability distribution can be represented by a **cumulative distribution function**, or **CDF**. This function, which is often denoted $F(x)$, is defined on the real line. Its value is $\Pr(X \leq x)$, the probability of the event that X is equal to or less than some value x. In general, the notation $\Pr(A)$ signifies the probability assigned to the event A, a subset of the full set of possibilities. Since X is continuous, it does not really matter whether we define the CDF as $\Pr(X \leq x)$ or as $\Pr(X < x)$ here, but it is conventional to use the former definition.

Notice that, in the preceding paragraph, we used X to denote a random variable and x to denote a realization of X, that is, a particular value that the random variable X may take on. This distinction is important when discussing the meaning of a probability distribution, but it will rarely be necessary in most of this book.

Probability Distributions

We may now make explicit the general rules that must be obeyed by probability distributions in assigning probabilities to events. There are just three of these rules:

(i) All probabilities lie between 0 and 1;

(ii) The null set is assigned probability 0, and the full set of possibilities is assigned probability 1;

Copyright © 2021, Russell Davidson and James G. MacKinnon

(iii) The probability assigned to an event that is the union of two disjoint events is the sum of the probabilities assigned to those disjoint events.

We will not often need to make explicit use of these rules, but we can use them now in order to derive some properties of any well-defined CDF for a scalar r.v. First, a CDF $F(x)$ tends to 0 as $x \to -\infty$. This follows because the event $(X \leq x)$ tends to the null set as $x \to -\infty$, and the null set has probability 0. By similar reasoning, $F(x)$ tends to 1 when $x \to +\infty$, because then the event $(X \leq x)$ tends to the entire real line. Further, $F(x)$ must be a weakly increasing function of x. This is true because, if $x_1 < x_2$, we have

$$(X \leq x_2) = (X \leq x_1) \cup (x_1 < X \leq x_2), \tag{2.02}$$

where $\cup$ is the symbol for set union. The two subsets on the right-hand side of (2.02) are clearly disjoint, and so

$$\Pr(X \leq x_2) = \Pr(X \leq x_1) + \Pr(x_1 < X \leq x_2).$$

Since all probabilities are nonnegative, it follows that the probability that $(X \leq x_2)$ must be no smaller than the probability that $(X \leq x_1)$.

For a continuous r.v., the CDF assigns probabilities to every interval on the real line. However, if we try to assign a probability to a single point, the result is always just zero. Suppose that X is a scalar r.v. with CDF $F(x)$. For any interval $[a, b]$ of the real line, the fact that $F(x)$ is weakly increasing allows us to compute the probability that $X \in [a, b]$. If $a < b$,

$$\Pr(X \leq b) = \Pr(X \leq a) + \Pr(a < X \leq b),$$

whence it follows directly from the definition of a CDF that

$$\Pr(a \leq X \leq b) = F(b) - F(a), \tag{2.03}$$

since, for a continuous r.v., we make no distinction between $\Pr(a < X \leq b)$ and $\Pr(a \leq X \leq b)$. If we set $b = a$, in the hope of obtaining the probability that $X = a$, then we get $F(a) - F(a) = 0$.

Probability Density Functions

For continuous random variables, the concept of the **probability density function**, or **PDF**, more frequently referred to as just the **density**, is very closely related to that of a CDF. Whereas a distribution function exists for any well-defined random variable, a density exists only when the random variable is continuous, and when its CDF is differentiable. For a scalar r.v., the density function, often denoted by f, is just the derivative of the CDF:

$$f(x) \equiv F'(x).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Because $F(-\infty) = 0$ and $F(\infty) = 1$, every density must be **normalized** to integrate to unity. By the Fundamental Theorem of Calculus,

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} F'(x)\,\mathrm{d}x = F(\infty) - F(-\infty) = 1. \tag{2.04}$$

It is obvious that a density is nonnegative, since it is the derivative of a weakly increasing function.

Probabilities can be computed in terms of the density as well as the CDF. Note that, by (2.03) and the Fundamental Theorem of Calculus once more,

$$\Pr(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x. \tag{2.05}$$

Since (2.05) must hold for arbitrary a and b, it is clear why $f(x)$ must always be nonnegative. However, it is important to remember that $f(x)$ is not bounded above by unity, because the value of a density at a point x is not a probability. Only when a density is integrated over some interval, as in (2.05), does it yield a probability.

The most common example of a continuous distribution is provided by the **normal distribution**. This is the distribution that generates the famous or infamous "bell curve" sometimes thought to influence students' grade distributions. The fundamental member of the normal family of distributions is the **standard normal distribution**. It is a continuous scalar distribution, defined on the entire real line. The density of the standard normal distribution is often denoted $\phi(\cdot)$. Its explicit expression, which we will need later in the book, is

$$\phi(x) = (2\pi)^{-1/2} \exp\bigl(-\tfrac{1}{2}x^2\bigr). \tag{2.06}$$

Unlike $\phi(\cdot)$, the CDF, usually denoted $\Phi(\cdot)$, has no elementary closed-form expression. However, by (2.05) with $a = -\infty$ and $b = x$, we have

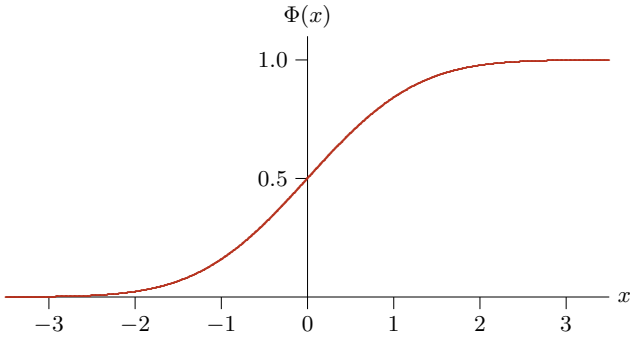
$$\Phi(x) = \int_{-\infty}^{x} \phi(y)\,\mathrm{d}y.$$

The functions $\Phi(\cdot)$ and $\phi(\cdot)$ are graphed in Figure 2.1. Since the density is the derivative of the CDF, it achieves a maximum at $x = 0$, where the CDF is rising most steeply. As the CDF approaches both 0 and 1, and consequently, becomes very flat, the density approaches 0.

Although it may not be obvious at once, discrete random variables can be characterized by a CDF just as well as continuous ones can be. Consider a binary r.v. X that can take on only two values, 0 and 1, and let the probability that $X = 0$ be p. It follfollows that the probability that $X = 1$ is $1 - p$. Then the CDF of X, according to the definition of $F(x)$ as $\Pr(X \le x)$, is the following discontinuous, "staircase" function:

Copyright © 2021, Russell Davidson and James G. MacKinnon

Standard Normal CDF:



Standard Normal PDF:

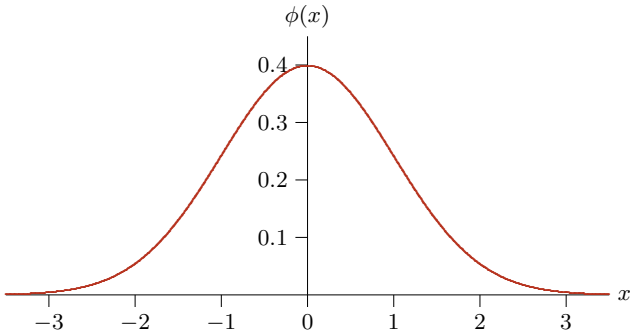


Figure 2.1 The CDF and PDF of the standard normal distribution

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ p & \text{for } 0 \le x < 1 \\ 1 & \text{for } x \ge 1. \end{cases}$$

Such a CDF, for $p = 0.7$, is graphed in Figure 2.2. Obviously, we cannot graph a corresponding density, for it does not exist. For general discrete random variables, the discontinuities of the CDF occur at the discrete permitted values of X, and the jump at each discontinuity is equal to the probability of the corresponding value. Since the sum of the jumps must therefore equal 1, the limiting value of F, to the right of all permitted values, is also 1.

Using a CDF is a reasonable way to deal with random variables that are neither completely discrete nor completely continuous. Such hybrid variables can be produced by the phenomenon of **censoring**. A random variable is said to be censored if not all of its potential values can actually be observed. For instance, in some data sets, a household's measured income is set equal to 0 if

Copyright © 2021, Russell Davidson and James G. MacKinnon

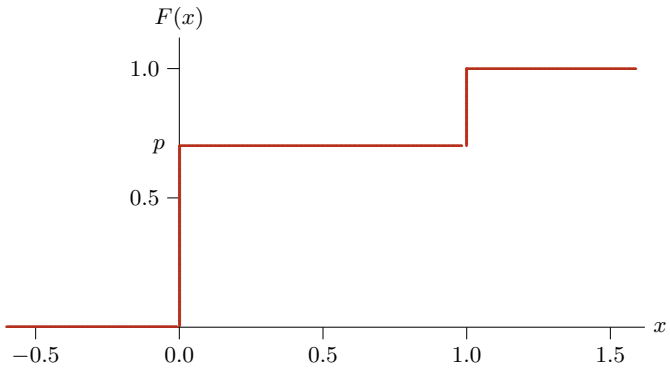


Figure 2.2 The CDF of a binary random variable

it is actually negative. It might be negative if, for instance, the household lost more on the stock market than it earned from other sources in a given year. Even if the true income variable is continuously distributed over the positive and negative real line, the observed, censored, variable has an **atom**, or bump, at 0, since the single value of 0 now has a nonzero probability attached to it, namely, the probability that an individual's income is nonpositive. As with a purely discrete random variable, the CDF has a discontinuity at 0, with a jump equal to the probability of a negative or zero income.

Moments of Random Variables

A fundamental property of a random variable is its **expectation**. For a discrete r.v. that can take on m possible finite values $x_1, x_2, \ldots, x_m$, the expectation is simply

$$\mathrm{E}(X) \equiv \sum_{i=1}^{m} p(x_i)\, x_i. \tag{2.07}$$

Thus each possible value x_i is multiplied by the probability associated with it. If m is infinite, the sum above has an infinite number of terms.

For a continuous r.v., the expectation is defined analogously using the density:

$$\mathrm{E}(X) \equiv \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x. \tag{2.08}$$

Not every r.v. has an expectation, however. The integral of a density function always exists and equals 1. But since X can range from $-\infty$ to ∞, the integral (2.08) may well diverge at either limit of integration, or both, if the density f does not tend to zero fast enough. Similarly, if m in (2.07) is infinite, the sum may diverge. The expectation of a random variable is sometimes called

Copyright © 2021, Russell Davidson and James G. MacKinnon

the **mean** or, to prevent confusion with the usual meaning of the word as the mean of a sample, the **population mean**. A common notation for it is μ.

The expectation of a random variable is often referred to as its **first moment**. The so-called **higher moments**, if they exist, are the expectations of the r.v. raised to a power. Thus the **second moment** of a random variable X is the expectation of X^2, the **third moment** is the expectation of X^3, and so on. In general, the k^{th} moment of a continuous random variable X is

$$m_k(X) \equiv \int_{-\infty}^{\infty} x^k f(x)\, \mathrm{d}x.$$

Observe that the value of any moment depends only on the probability distribution of the r.v. in question. For this reason, we often speak of the moments of the distribution rather than the moments of a specific random variable. If a distribution possesses a k^{th} moment, it also possesses all moments of order less than k.

The higher moments just defined are called the **uncentered moments** of a distribution, because, in general, X does not have mean zero. It is often more useful to work with the **central moments**, which are defined as the ordinary moments of the difference between the random variable and its expectation. Thus the k^{th} central moment of the distribution of a continuous r.v. X is

$$\mu_k \equiv \mathrm{E}\big(X - \mathrm{E}(X)\big)^k = \int_{-\infty}^{\infty} (x - \mu)^k f(x)\, \mathrm{d}x,$$

where $\mu \equiv \mathrm{E}(X)$. For a discrete X, the k^{th} central moment is

$$\mu_k \equiv \mathrm{E}\big(X - \mathrm{E}(X)\big)^k = \sum_{i=1}^{m} p(x_i)(x_i - \mu)^k.$$

By far the most important central moment is the second. It is called the **variance** of the random variable and is frequently written as $\mathrm{Var}(X)$. Another common notation for a variance is σ^2. This notation underlines the important fact that a variance cannot be negative. The positive square root of the variance, σ, is called the **standard deviation** of the distribution. Estimates of standard deviations are often referred to as **standard errors**, especially when the random variable in question is a parameter estimator.

Multivariate Distributions

A **vector-valued random variable** takes on values that are vectors. It can be thought of as several scalar random variables that have a single, joint distribution. For simplicity, we will focus on the case of **bivariate random variables**, where the vector has two elements. A continuous, bivariate random variable (X_1, X_2) has a distribution function

$$F(x_1, x_2) = \Pr\big((X_1 \leq x_1) \cap (X_2 \leq x_2)\big),$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $\cap$ is the symbol for set intersection. Thus $F(x_1, x_2)$ is the joint probability that both $X_1 \leq x_1$ and $X_2 \leq x_2$. For continuous variables, the density, if it exists, is the **joint density function**[2]

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}. \tag{2.09}$$

This function has exactly the same properties as an ordinary density. In particular, as in (2.04),

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_1 \mathrm{d}x_2 = 1.$$

More generally, the probability that X_1 and X_2 jointly lie in any region is the integral of $f(x_1, x_2)$ over that region. A case of particular interest is

$$\begin{aligned} F(x_1, x_2) &= \Pr\big((X_1 \leq x_1) \cap (X_2 \leq x_2)\big) \\ &= \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(y_1, y_2)\, \mathrm{d}y_1 \mathrm{d}y_2, \end{aligned} \tag{2.10}$$

which shows how to compute the CDF given the density.

The concept of joint probability distributions leads naturally to the important notion of **statistical independence**. Let (X_1, X_2) be a bivariate random variable. Then X_1 and X_2 are said to be **statistically independent**, or often just **independent**, if the joint CDF of (X_1, X_2) is the product of the CDFs of X_1 and X_2. In straightforward notation, this means that

$$F(x_1, x_2) = F(x_1, \infty) F(\infty, x_2). \tag{2.11}$$

The first factor here is the joint probability that $X_1 \leq x_1$ and $X_2 \leq \infty$. Since the second inequality imposes no constraint, this factor is just the probability that $X_1 \leq x_1$. The function $F(x_1, \infty)$, which is called the **marginal CDF** of X_1, is thus just the CDF of X_1 considered by itself. Similarly, the second factor on the right-hand side of (2.11) is the marginal CDF of X_2.

It is also possible to express statistical independence in terms of the **marginal density** of X_1 and the marginal density of X_2. The marginal density of X_1 is, as one would expect, the derivative of the marginal CDF of X_1,

$$f(x_1) \equiv F_1(x_1, \infty),$$

[2] Here we are using what computer scientists would call "overloaded function" notation. This means that $F(\cdot)$ and $f(\cdot)$ denote, respectively, the CDF and the density of whatever their argument(s) happen to be. This practice is harmless provided there is no ambiguity.

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $F_1(\cdot)$ denotes the partial derivative of $F(\cdot)$ with respect to its first argument. It can be shown from (2.10) that the marginal density can also be expressed in terms of the joint density, as follows:

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, \mathrm{d}x_2. \tag{2.12}$$

Thus $f(x_1)$ is obtained by integrating X_2 out of the joint density. Similarly, the marginal density of X_2 is obtained by integrating X_1 out of the joint density. From (2.09), it can be shown that, if X_1 and X_2 are independent, so that (2.11) holds, then

$$f(x_1, x_2) = f(x_1) f(x_2). \tag{2.13}$$

Thus, when densities exist, statistical independence means that the joint density factorizes as the product of the marginal densities, just as the joint CDF factorizes as the product of the marginal CDFs.

Conditional Probabilities

Suppose that A and B are any two events. Then the probability of event A **conditional** on B, or given B, is denoted as $\Pr(A \,|\, B)$ and is defined implicitly by the equation

$$\Pr(A \cap B) = \Pr(B) \Pr(A \,|\, B). \tag{2.14}$$

For this equation to make sense as a definition of $\Pr(A \,|\, B)$, it is necessary that $\Pr(B) \neq 0$. The idea underlying the definition is that, if we know somehow that the event B has been realized, this knowledge can provide information about whether event A has also been realized. For instance, if A and B are disjoint, and B is realized, then it is certain that A has not been. As we would wish, this does indeed follow from the definition (2.14), since $A \cap B$ is the null set, of zero probability, if A and B are disjoint. Similarly, if B is a subset of A, knowing that B has been realized means that A must have been realized as well. Since in this case $\Pr(A \cap B) = \Pr(B)$, (2.14) tells us that $\Pr(A \,|\, B) = 1$, as required.

To gain a better understanding of (2.14), consider Figure 2.3. The bounding rectangle represents the full set of possibilities, and events A and B are subsets of the rectangle that overlap as shown. Suppose that the figure has been drawn in such a way that probabilities of subsets are proportional to their areas. Thus the probabilities of A and B are the ratios of the areas of the corresponding circles to the area of the bounding rectangle, and the probability of the intersection $A \cap B$ is the ratio of its area to that of the rectangle.

Suppose now that it is known that B has been realized. This fact leads us to redefine the probabilities so that everything outside B now has zero probability, while, inside B, probabilities remain proportional to areas. Event B now has probability 1, in order to keep the total probability equal to 1. Event A

Copyright © 2021, Russell Davidson and James G. MacKinnon

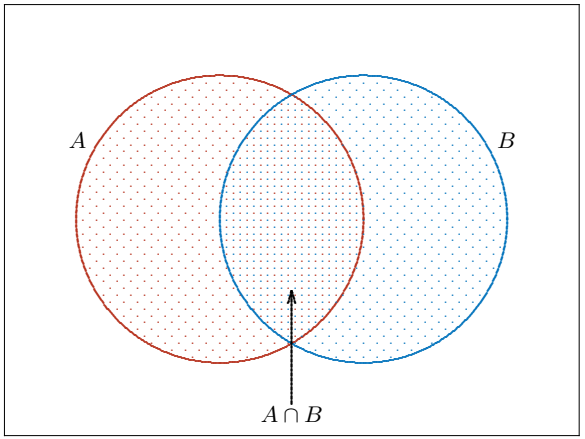


Figure 2.3 Conditional probability

can be realized only if the realized point is in the intersection $A \cap B$, since the set of all points of A outside this intersection has zero probability. The probability of A, conditional on knowing that B has been realized, is thus the ratio of the area of $A \cap B$ to that of B. This construction leads directly to equation (2.14).

There are many ways to associate a random variable X with the rectangle shown in Figure 2.3. Such a random variable could be any function of the two coordinates that define a point in the rectangle. For example, it could be the horizontal coordinate of the point measured from the origin at the lower left-hand corner of the rectangle, or its vertical coordinate, or the Euclidean distance of the point from the origin. The realization of X is the value of the function it corresponds to at the realized point in the rectangle.

For concreteness, let us assume that the function is simply the horizontal coordinate, and let the width of the rectangle be equal to 1. Then, since all values of the horizontal coordinate between 0 and 1 are equally probable, the random variable X has what is called the **uniform distribution** on the interval $[0, 1]$. The CDF of this distribution is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1. \end{cases}$$

Because $F(x)$ is not differentiable at $x = 0$ and $x = 1$, the density of the uniform distribution does not exist at those points. Elsewhere, the derivative of $F(x)$ is 0 outside $[0, 1]$ and 1 inside. The CDF and density are illustrated in Figure 2.4. This special case of the uniform distribution is often denoted the $\mathrm{U}(0, 1)$ distribution.

Copyright © 2021, Russell Davidson and James G. MacKinnon

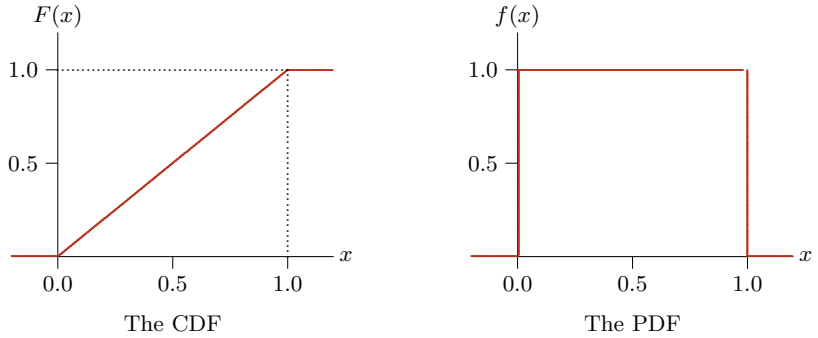


Figure 2.4 The CDF and PDF of the uniform distribution on $[0, 1]$

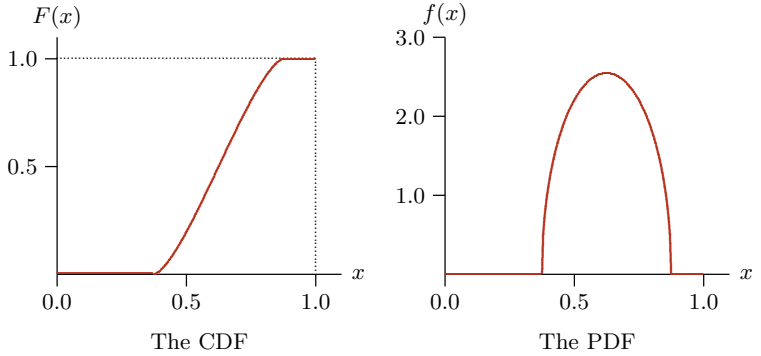


Figure 2.5 The CDF and PDF conditional on event B

If the information were available that B had been realized, then the distribution of X conditional on this information would be very different from the $\mathrm{U}(0, 1)$ distribution. Now only values between the extreme horizontal limits of the circle of B are allowed. If one computes the area of the part of the circle to the left of a given vertical line, then for each event $a \equiv (X \leq x)$ the probability of this event conditional on B can be worked out. The result is just the CDF of X conditional on the event B. Its derivative is the density of X conditional on B. These are shown in Figure 2.5.

The concept of conditional probability can be extended beyond probability conditional on an event to probability conditional on a random variable. Suppose that X_1 is a r.v. and X_2 is a discrete r.v. with permitted values $z_1, \ldots, z_m$. For each $i = 1, \ldots, m$, the CDF of X_1, and, if X_1 is continuous, its density, can be computed conditional on the event $(X_2 = z_i)$. If X_2 is also a continuous r.v., then things are a little more complicated, because events like $(X_2 = x_2)$

Copyright © 2021, Russell Davidson and James G. MacKinnon

for some real x_2 have zero probability, and so cannot be conditioned on in the manner of (2.14).

On the other hand, it makes perfect intuitive sense to think of the distribution of X_1 conditional on some specific realized value of X_2. This conditional distribution gives us the probabilities of events concerning X_1 when we know that the realization of X_2 was actually x_2. We therefore make use of the **conditional density** of X_1 for a given value x_2 of X_2. This **conditional density**, or conditional PDF, is defined as

$$f(x_1 \,|\, x_2) = \frac{f(x_1, x_2)}{f(x_2)}. \tag{2.15}$$

Thus, for a given value x_2 of X_2, the conditional density is proportional to the joint density of X_1 and X_2. Of course, (2.15) is well defined only if $f(x_2) > 0$. In some cases, more sophisticated definitions can be found that would allow $f(x_1 \,|\, x_2)$ to be defined for all x_2 even if $f(x_2) = 0$, but we will not need these in this book. See, among others, Billingsley (1995).

Conditional Expectations

Whenever we can describe the distribution of a random variable, X_1, conditional on another, X_2, either by a conditional CDF or a conditional density, we can consider the **conditional expectation** of X_1. If it exists, this conditional expectation is just the ordinary expectation computed using the conditional distribution. If x_2 is a possible value for X_2, then this conditional expectation is written as $\mathrm{E}(X_1 \,|\, x_2)$.

For a given value x_2, the conditional expectation $\mathrm{E}(X_1 \,|\, x_2)$ is, like any other ordinary expectation, a deterministic, that is, nonrandom, quantity. But we can consider the expectation of X_1 conditional on *every* possible realization of X_2. In this way, we can construct a new random variable, which we denote by $\mathrm{E}(X_1 \,|\, X_2)$, the realization of which is $\mathrm{E}(X_1 \,|\, x_2)$ when the realization of X_2 is x_2. We can call $\mathrm{E}(X_1 \,|\, X_2)$ a deterministic function of the random variable X_2, because the realization of $\mathrm{E}(X_1 \,|\, X_2)$ is unambiguously determined by the realization of X_2.

Conditional expectations defined as random variables in this way have a number of interesting and useful properties. The first, called the **Law of Iterated Expectations**, can be expressed as follows:

$$\mathrm{E}\big(\mathrm{E}(X_1 \,|\, X_2)\big) = \mathrm{E}(X_1). \tag{2.16}$$

If a conditional expectation of X_1 can be treated as a random variable, then the conditional expectation itself may have an expectation. According to (2.16), this expectation is just the ordinary expectation of X_1.

Another property of conditional expectations is that any deterministic function of a conditioning variable X_2 is its own conditional expectation. Thus,

Copyright © 2021, Russell Davidson and James G. MacKinnon

for example, $\mathrm{E}(X_2 \,|\, X_2) = X_2$, and $\mathrm{E}(X_2^2 \,|\, X_2) = X_2^2$. A result, sometimes referred to as **taking out what is known**, says that, conditional on X_2, the expectation of a product of another random variable X_1 and a deterministic function of X_2 is the product of that deterministic function and the expectation of X_1 conditional on X_2:

$$\mathrm{E}\big(X_1 h(X_2) \,|\, X_2\big) = h(X_2)\mathrm{E}(X_1 \,|\, X_2), \tag{2.17}$$

for any deterministic function $h(\cdot)$. An important special case of this, which we will make use of in Section 2.5, arises when $\mathrm{E}(X_1 \,|\, X_2) = 0$. In that case, for any function $h(\cdot)$, $\mathrm{E}(X_1 h(X_2)) = 0$, because

$$\begin{aligned} \mathrm{E}\big(X_1 h(X_2)\big) &= \mathrm{E}\big(\mathrm{E}(X_1 h(X_2) \,|\, X_2)\big) \\ &= \mathrm{E}\big(h(X_2)\mathrm{E}(X_1 \,|\, X_2)\big) \\ &= \mathrm{E}(0) = 0. \end{aligned}$$

The first equality here follows from the Law of Iterated Expectations, (2.16). The second follows from (2.17). Since $\mathrm{E}(X_1 \,|\, X_2) = 0$, the third line then follows immediately. We will present other properties of conditional expectations as the need arises.

2.3 The Specification of Regression Models

At this point, it is appropriate to formalize the idea of a **model** in the context of econometrics. We begin by recalling the notion of a **data-generating process**, or **DGP**. In Section 1.2, we proposed a definition of a DGP as something that can be simulated on a computer, and that constitutes a unique recipe for simulation. This definition is fine for virtual reality, but, despite some claims to the contrary, we do not think that we are living in a simulation! What do we mean, then, in speaking of a DGP in the real world?

The difficulty here is that the real world is messy. Statistical agencies wrestle with this problem all the time, but succeed nonetheless in generating the data sets used by econometricians. Our explanations of the economy, and our understanding of economic mechanisms, are based on these data sets. In medicine, our understanding of biological mechanisms is based on data collected in hospitals, clinics, and labs. In the physical sciences, we base our theories on experimental and observational data. In all these disciplines, we can talk about data-generating processes, without entering into the details of just how data are collected.

It wouldn't make a lot of sense to say that a statistical agency, for instance, is a real-world DGP, although such agencies certainly do generate much of the data we use in empirical work. In the end, it is probably better to say as little

Copyright © 2021, Russell Davidson and James G. MacKinnon

as possible about a real-world DGP, and instead speak of external reality, or simply of the real world.

We now return our attention to the regression model (2.01) and revert to the notation of Section 2.1 in which y_t and x_t denote, respectively, the dependent and independent variables. The model (2.01) can be interpreted as a model for the expectation of y_t conditional on x_t. Let us assume that the disturbance u_t has expectation 0 conditional on x_t. Then, taking conditional expectations of both sides of (2.01), we see that

$$\mathrm{E}(y_t \,|\, x_t) = \beta_1 + \beta_2 x_t + \mathrm{E}(u_t \,|\, x_t) = \beta_1 + \beta_2 x_t.$$

Without the key assumption that $\mathrm{E}(u_t \,|\, x_t) = 0$, the second equality here would not hold. As we pointed out in Section 2.1, it is impossible to make any sense of a regression model unless we make strong assumptions about the disturbances. Of course, we could define u_t as the difference between y_t and $\mathrm{E}(y_t \,|\, x_t)$, which would give $\mathrm{E}(u_t \,|\, x_t) = 0$ by definition. But if we require that $\mathrm{E}(u_t \,|\, x_t) = 0$ and also specify (2.01), we must necessarily have $\mathrm{E}(y_t \,|\, x_t) = \beta_1 + \beta_2 x_t$.

As an example, suppose that we estimate the model (2.01) when in fact

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_t^2 + v_t \tag{2.18}$$

with $\beta_3 \neq 0$ and a disturbance v_t such that $\mathrm{E}(v_t \,|\, x_t) = 0$. If the data were generated by (2.18), the disturbance u_t in (2.01) would be equal to $\beta_3 x_t^2 + v_t$. By the results on conditional expectations in Section 2.2, we see that

$$\mathrm{E}(u_t \,|\, x_t) = \mathrm{E}\big(\beta_3 x_t^2 + v_t \,|\, x_t\big) = \beta_3 x_t^2,$$

which must be nonzero unless $x_t = 0$. This example shows the force of the assumption that the expectation of the disturbance is zero conditional on x_t. Unless the expectation of y_t conditional on x_t really is a linear function of x_t, the regression function in (2.01) is not **correctly specified**, in the precise sense that (2.01) cannot hold with a disturbance that has expectation zero conditional on x_t. It will become clear in later chapters that estimating incorrectly specified models usually leads to results that are meaningless or, at best, seriously misleading.

Information Sets

In a more general setting, what we are interested in is usually not the expectation of y_t conditional on a single explanatory variable x_t but its expectation conditional on a set of potential explanatory variables. This set is often called an **information set**, and it is denoted Ω_t. Typically, the information set contains more variables than would actually be used in a regression model. For example, it might consist of all the variables observed by the economic agents

Copyright © 2021, Russell Davidson and James G. MacKinnon

whose actions determine y_t at the time they make the decisions that cause them to perform those actions. Such an information set could be very large. As a consequence, much of the art of constructing, or specifying, a regression model is deciding which of the variables that belong to Ω_t should be included in the model and which of the variables should be excluded.

In some cases, economic theory makes it fairly clear what the information set Ω_t should consist of, and sometimes also which variables in Ω_t should make their way into a regression model. In many others, however, it may not be at all clear how to specify Ω_t. In general, we want to condition on **exogenous** variables but not on **endogenous** ones. These terms refer to the *origin* or *genesis* of the variables: An exogenous variable has its origins *outside* the model under consideration, while the mechanism generating an endogenous variable is *inside* the model. When we write a single equation like (2.01), the only endogenous variable allowed is the dependent variable, y_t.

Recall the example of the consumption function that we looked at in Section 2.1. That model seeks to explain household consumption in terms of disposable income, but it makes no claim to explain disposable income, which is simply taken as given. The consumption function model can be correctly specified only if two conditions hold:

(i) The expectation of consumption conditional on disposable income is an affine function of the latter.

(ii) Consumption is *not* a variable that contributes to the determination of disposable income.

The second condition means that the origin of disposable income, that is, the mechanism by which disposable income is generated, lies outside the model for consumption. In other words, disposable income is exogenous in that model. If the simple consumption model we have presented is correctly specified, the two conditions above must be satisfied. Needless to say, we do not claim that this model is in fact correctly specified.

It is not always easy to decide just what information set to condition on. As the above example shows, it is often not clear whether or not a variable is exogenous. Moreover, even if a variable clearly is exogenous, we may not want to include it in Ω_t. For example, if the ultimate purpose of estimating a regression model is to use it for forecasting, there may be no point in conditioning on information that will not be available at the time the forecast is to be made.

Disturbances

Whenever we specify a regression model, it is essential to make assumptions about the properties of the disturbances. The simplest set of assumptions is that all of them have expectation 0, come from the same distribution, and are independent of each other. Although this is a rather strong set of assumptions, it is very commonly adopted in practice.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Mutual independence of the disturbances, when coupled with the assumption that $\mathrm{E}(u_t) = 0$, implies that the expectation of u_t is 0 conditional on all of the other disturbances u_s, $s \neq t$. However, the implication does not work in the other direction, because the assumption of mutual independence is stronger than the assumption about the conditional expectations. A very strong assumption which is often made is that the disturbances are **independently and identically distributed**, or **IID**. According to this assumption, the disturbances are mutually independent, and they are in addition realizations from the same, identical, probability distribution.

When the successive observations are ordered by time, it often seems plausible that a disturbance is correlated with neighboring disturbances. Thus u_t might well be correlated with u_s when the value of $|t - s|$ is small. This could occur, for example, if there is correlation across time periods of random factors that influence the dependent variable but are not explicitly accounted for in the regression function. This phenomenon is called **serial correlation**, and it often appears to be observed in practice. When there is serial correlation, the disturbances cannot be IID because they are not independent.

Another possibility is that the variance of the disturbances may be systematically larger for some observations than for others. This happens if the conditional variance of y_t depends on some of the same variables as the conditional expectation. This phenomenon is called **heteroskedasticity**, and it also often appears to be observed in practice. For example, in the case of the consumption function, the variance of consumption may well be higher for households with high incomes than for households with low incomes. When there is heteroskedasticity, the disturbances cannot be IID, because they are not identically distributed. It is perfectly possible to take explicit account of both serial correlation and heteroskedasticity, but doing so would take us outside the context of regression models like (2.01).

It may sometimes be desirable to write a regression model like the one we have been studying as

$$\mathrm{E}(y_t \,|\, \Omega_t) = \beta_1 + \beta_2 x_t, \tag{2.19}$$

in order to stress the fact that this is a model for the expectation of y_t conditional on a certain information set. However, by itself, (2.19) is just as incomplete a specification as (2.01). In order to see this point, we must now state what we mean by a **complete specification** of a regression model. Probably the best way to do this is to say that a complete specification of any econometric model is one that provides an unambiguous recipe for simulating the model on a computer. After all, if we can use the model to generate simulated data, it must be completely specified.

Simulating Econometric Models

Consider equation (2.01). When we say that we **simulate** this model, we mean that we generate numbers for the dependent variable, y_t, according

Copyright © 2021, Russell Davidson and James G. MacKinnon

to equation (2.01). Obviously, one of the first things we must fix for the simulation is the sample size, n. That done, we can generate each of the y_t, for $t = 1, \ldots, n$, by evaluating the right-hand side of the equation n times. For this to be possible, we need to know the value of each variable and each parameter that appears on the right-hand side.

If we suppose that the explanatory variable x_t is exogenous, then we simply take it as given. In the context of the consumption function example, if we had data on the disposable income of households in some country every year for a period of n years, we could just use those data. Our simulation would then be specific to the country in question and to the time period of the data. Alternatively, it could be that we or some other econometricians had previously specified another model, for the explanatory variable this time, and we could then use simulated data provided by that model.

Besides the explanatory variable, the other elements of the right-hand side of (2.01) are the parameters, β_1 and β_2, and the disturbance u_t. A key feature of the parameters is that we do not know their true values. However, for purposes of simulation, we could use either values suggested by economic theory or values obtained by estimating the model. Evidently, the simulation results will depend on precisely what values we use.

Unlike the parameters, the disturbances cannot be taken as given; instead, we wish to treat them as random. Luckily, it is easy to use a computer to generate "random" numbers by using a program called a **random number generator**; we mentioned these already in Section 1.2, and will discuss them again in more detail in Chapter 7. The "random" numbers generated by computers are not random according to some meanings of the word. For instance, a computer can be made to spit out exactly the same sequence of supposedly random numbers more than once. In addition, a digital computer is a perfectly deterministic device. Therefore, if random means the opposite of deterministic, only computers that are not functioning properly would be capable of generating truly random numbers. Because of this, some people prefer to speak of computer-generated random numbers as **pseudo-random**. However, for the purposes of simulations, the numbers computers provide have all the properties of random numbers that we need, and so we will call them simply random rather than pseudo-random.

Computer-generated random numbers are mutually independent **drawings**, or realizations, from specific probability distributions, usually the uniform $\mathrm{U}(0, 1)$ distribution or the standard normal distribution, both of which were defined in Section 2.2. Of course, techniques exist for generating drawings from many other distributions as well, as do techniques for generating drawings that are not independent. For the moment, the essential point is that we must always specify the probability distribution of the random numbers we use in a simulation. It is important to note that specifying the expectation of a distribution, or even the expectation conditional on some other variables, is not enough to specify the distribution in full.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Let us now summarize the various steps in performing a simulation by giving a sort of generic algorithm for simulations of regression models. In the model specification, it is convenient to distinguish between the **deterministic specification** and the **stochastic specification**. In model (2.01), the deterministic specification consists of the regression function, of which the ingredients are the explanatory variable and the parameters. The stochastic specification ("stochastic" is another word for "random") consists of the probability distribution of the disturbances, and the requirement that the disturbances should be IID drawings from this distribution. Then, in order to simulate the dependent variable y_t in (2.01), we do as follows:

- Fix the sample size, n;
- Choose the parameters (here β_1 and β_2) of the deterministic specification;
- Obtain the n successive values x_t, $t = 1, \ldots, n$, of the explanatory variable. As explained above, these values may be real-world data or the output of another simulation;
- Evaluate the n successive values of the regression function $\beta_1 + \beta_2 x_t$, for $t = 1, \ldots, n$;
- Choose the probability distribution of the disturbances, if necessary specifying parameters such as its expectation and variance;
- Use a random-number generator to generate the n successive and mutually independent values u_t of the disturbances;
- Form the n successive values y_t of the dependent variable by adding the disturbances to the values of the regression function.

The n values y_t, $t = 1, \ldots, n$, thus generated are the output of the simulation; they are the **simulated values** of the dependent variable.

The chief interest of such a simulation is that, if the model we simulate is correctly specified and thus reflects the real-world generating process for the dependent variable, our simulation mimics the real world accurately, because it makes use of the same data-generating mechanism as that in operation in the real world.

A complete specification, then, is anything that leads unambiguously to a recipe like the one given above. We will define a **fully specified parametric model** as a model for which it is possible to simulate the dependent variable once the values of the parameters are known. A **partially specified parametric model** is one for which more information, over and above the parameter values, must be supplied before simulation is possible. Both sorts of models are frequently encountered in econometrics.

To conclude this discussion of simulations, let us return to the specifications (2.01) and (2.19). Both are obviously incomplete as they stand. In order to complete either one, it is necessary to specify the information set Ω_t and the distribution of u_t conditional on Ω_t. In particular, it is necessary to know

Copyright © 2021, Russell Davidson and James G. MacKinnon

whether the disturbances u_s with $s \neq t$ belong to Ω_t. In (2.19), one aspect of the conditional distribution is given, namely, the conditional expectation. Unfortunately, because (2.19) contains no explicit disturbance, it is easy to forget that it is there. Perhaps as a result, it is more common to write regression models in the form of (2.01) than in the form of (2.19). However, writing a model in the form of (2.01) does have the disadvantage that it obscures both the dependence of the model on the choice of an information set and the fact that the distribution of the disturbances must be specified conditional on that information set.

Linear and Nonlinear Regression Models

The simple linear regression model (2.01) is by no means the only reasonable model for the expectation of y_t conditional on x_t. Consider, for example, the models

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_t^2 + u_t \tag{2.20}$$

$$y_t = \gamma_1 + \gamma_2 \log x_t + u_t, \quad \text{and} \tag{2.21}$$

$$y_t = \delta_1 + \delta_2 \frac{1}{x_t} + u_t. \tag{2.22}$$

These are all models that might be plausible in some circumstances.[3] In equation (2.20), there is an extra parameter, β_3, which allows $\mathrm{E}(y_t \mid x_t)$ to vary quadratically with x_t whenever β_3 is nonzero. In effect, x_t and x_t^2 are being treated as separate explanatory variables. Thus (2.20) is the first example we have seen of a **multiple linear regression model**. It reduces to the simple linear regression model (2.01) when $\beta_3 = 0$.

In the models (2.21) and (2.22), on the other hand, there are no extra parameters. Instead, a nonlinear transformation of x_t is used in place of x_t itself. As a consequence, the relationship between x_t and $\mathrm{E}(y_t \mid x_t)$ in these two models is necessarily nonlinear. Nevertheless, (2.20), (2.21), and (2.22) are all said to be linear regression models, because, even though the expectation of y_t may depend nonlinearly on x_t, it always depends linearly on the unknown parameters of the regression function. As we will see in Section 2.5, it is quite easy to estimate a linear regression model. In contrast, genuinely nonlinear models, in which the regression function depends nonlinearly on the parameters, are somewhat harder to estimate, and are not treated in this book.

Because it is very easy to estimate linear regression models, a great deal of applied work in econometrics makes use of them. It may seem that the

[3] In this book, all logarithms are natural logarithms. Thus $a = \log x$ implies that $x = \mathrm{e}^a$. Some authors use "ln" to denote natural logarithms and "log" to denote base 10 logarithms. Since econometricians should never have any use for base 10 logarithms, we avoid this aesthetically displeasing notation.

Copyright © 2021, Russell Davidson and James G. MacKinnon

linearity assumption is very restrictive. However, as the examples (2.20), (2.21), and (2.22) illustrate, this assumption need not be unduly restrictive in practice, at least not if the econometrician is at all creative. If we are willing to transform the dependent variable as well as the independent ones, the linearity assumption can be made even less restrictive. As an example, consider the nonlinear regression model

$$y_t = \mathrm{e}^{\beta_1} x_{t2}^{\beta_2} x_{t3}^{\beta_3} + u_t, \tag{2.23}$$

in which there are two explanatory variables, x_{t2} and x_{t3}, and the regression function is multiplicative. If the notation seems odd, suppose that there is implicitly a third explanatory variable, x_{t1}, which is constant and always equal to e. Notice that the regression function in (2.23) can be evaluated only when x_{t2} and x_{t3} are positive for all t.[4] It is a genuinely nonlinear regression function, because it is clearly linear neither in parameters nor in variables. For reasons that will shortly become apparent, a nonlinear model like (2.23) is very rarely estimated in practice.

A model like (2.23) is not as outlandish as may appear at first glance. It could arise, for instance, if we wanted to estimate a Cobb-Douglas production function. In that case, y_t would be output for observation t, and x_{t2} and x_{t3} would be inputs, say labor and capital. Since e^{β_1} is just a positive constant, it plays the role of the scale factor that is present in every Cobb-Douglas production function.

As equation (2.23) is written, everything enters multiplicatively except the disturbance. But it is easy to modify (2.23) so that the disturbance also enters multiplicatively. One way to do this is to write

$$y_t = e^{\beta_1} x_{t2}^{\beta_2} x_{t3}^{\beta_3} + u_t \equiv \big(e^{\beta_1} x_{t2}^{\beta_2} x_{t3}^{\beta_3}\big)(1 + v_t), \tag{2.24}$$

where the disturbance factor $1 + v_t$ multiplies the regression function. If we now assume that the underlying disturbances v_t are IID, it follows that the additive disturbances u_t are proportional to the regression function. This may well be a more plausible specification than that in which the u_t are supposed to be IID, as was implicitly assumed in (2.23). To see this, notice first that the additive disturbance u_t has the same units of measurement as y_t. If (2.23) is interpreted as a production function, then u_t is measured in units of output. However, the multiplicative disturbance v_t is dimensionless. In other words, it is a pure number, like 0.02, which could be expressed as 2 per cent. If the u_t are assumed to be IID, then we are assuming that the random component of output is of the same order of magnitude regardless of the scale of production. If, on the other hand, the v_t are assumed to be IID, then the

[4] If x and a are real numbers, x^a is not usually a real number unless $x > 0$. Think of the square root of -1.

Copyright © 2021, Russell Davidson and James G. MacKinnon

random component is proportional to total output. This second assumption is almost always more reasonable than the first.

If the model (2.24) is a good one, then the v_t should be quite small, usually less than about 0.05. For small values of the argument w, a standard approximation to the exponential function gives us that $\mathrm{e}^w \cong 1 + w$. As a consequence, (2.24) is very similar to the model

$$y_t = \mathrm{e}^{\beta_1} x_{t2}^{\beta_2} x_{t3}^{\beta_3} \mathrm{e}^{v_t} \tag{2.25}$$

whenever the disturbances are reasonably small.

Now suppose we take logarithms of both sides of (2.25). The result is

$$\log y_t = \beta_1 + \beta_2 \log x_{t2} + \beta_3 \log x_{t3} + v_t, \tag{2.26}$$

which is a **loglinear regression model**. This model is linear in the parameters and in the logarithms of all the variables, and so it is very much easier to estimate than the nonlinear model (2.23). Since (2.25) is at least as plausible as (2.23), it is not surprising that loglinear regression models, like (2.26), are estimated very frequently in practice, while multiplicative models with additive disturbances, like (2.23), are very rarely estimated. Of course, it is important to remember that (2.26) is not a model for the expectation of y_t conditional on x_{t2} and x_{t3}. Instead, it is a model for the expectation of $\log y_t$ conditional on those variables. If it is really the conditional expectation of y_t that we are interested in, then we will not want to estimate a loglinear model like (2.26).

2.4 Matrix Algebra

It is impossible to study econometrics beyond the most elementary level without using matrix algebra. Most readers are probably already quite familiar with matrix algebra. This section reviews some basic results that will be used throughout the book. It also shows how regression models can be written very compactly using matrix notation. More advanced material will be discussed in later chapters, as it is needed.

An $n \times m$ **matrix** $\boldsymbol{A}$ is a rectangular array that consists of nm elements arranged in n rows and m columns. The name of the matrix is conventionally shown in boldface. A typical element of $\boldsymbol{A}$ might be denoted by either A_{ij} or a_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The first subscript always indicates the row, and the second always indicates the column. It is sometimes necessary to show the elements of a matrix explicitly, in which case they are arrayed in rows and columns and surrounded by large brackets, as in

$$\boldsymbol{B} = \begin{bmatrix} 2 & 3 & 6 \\ 4 & 5 & 8 \end{bmatrix}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Here $\boldsymbol{B}$ is a 2×3 matrix.

If a matrix has only one column or only one row, it is called a **vector**. There are two types of vectors, **column vectors** and **row vectors**. Since column vectors are more common than row vectors, a vector that is not specified to be a row vector is normally treated as a column vector. If a column vector has n elements, it may be referred to as an n–vector. Boldface is used to denote vectors as well as matrices. It is conventional to use uppercase letters for matrices and lowercase letters for column vectors. However, it is sometimes necessary to ignore this convention.

If a matrix has the same number of columns and rows, it is said to be **square**. A square matrix $\boldsymbol{A}$ is **symmetric** if $A_{ij} = A_{ji}$ for all i and j. Symmetric matrices occur very frequently in econometrics. A square matrix is said to be **diagonal** if $A_{ij} = 0$ for all $i \neq j$; in this case, the only nonzero entries are those on what is called the **principal diagonal**. Sometimes a square matrix has all zeros above or below the principal diagonal. Such a matrix is said to be **triangular**. If the nonzero elements are all above the diagonal, it is said to be **upper-triangular**; if the nonzero elements are all below the diagonal, it is said to be **lower-triangular**. Here are some examples:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 6 \\ 4 & 6 & 5 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 0 \\ 5 & 2 & 6 \end{bmatrix}.$$

In this case, $\boldsymbol{A}$ is symmetric, $\boldsymbol{B}$ is diagonal, and $\boldsymbol{C}$ is lower-triangular.

The **transpose** of a matrix is obtained by interchanging its row and column subscripts. Thus the ij^{th} element of $\boldsymbol{A}$ becomes the ji^{th} element of its transpose, which is denoted $\boldsymbol{A}^\top$. Note that many authors use $\boldsymbol{A}'$ rather than $\boldsymbol{A}^\top$ to denote the transpose of $\boldsymbol{A}$. The transpose of a symmetric matrix is equal to the matrix itself. The transpose of a column vector is a row vector, and vice versa. Here are some examples:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 5 & 7 \\ 3 & 8 & 4 \end{bmatrix} \quad \boldsymbol{A}^\top = \begin{bmatrix} 2 & 3 \\ 5 & 8 \\ 7 & 4 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \quad \boldsymbol{b}^\top = \begin{bmatrix} 2 & 4 & 6 \end{bmatrix}.$$

Note that a matrix $\boldsymbol{A}$ is symmetric if and only if $\boldsymbol{A} = \boldsymbol{A}^\top$.

Arithmetic Operations on Matrices

Addition and **subtraction** of matrices works exactly the way it does for scalars, with the proviso that matrices can be added or subtracted only if they are **conformable**. In the case of addition and subtraction, this just means that they must have the same dimensions, that is, the same number of rows and the same number of columns. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable, then a typical element of $\boldsymbol{A} + \boldsymbol{B}$ is simply $A_{ij} + B_{ij}$, and a typical element of $\boldsymbol{A} - \boldsymbol{B}$ is $A_{ij} - B_{ij}$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Matrix multiplication actually involves both additions and multiplications. It is based on what is called the **inner product**, or **scalar product**, or sometimes **dot product** of two vectors. Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ are n–vectors. Then their inner product is

$$\boldsymbol{a}^\top \boldsymbol{b} = \boldsymbol{b}^\top \boldsymbol{a} = \sum_{i=1}^{n} a_i b_i.$$

As the name suggests, this is just a scalar.

When two matrices are multiplied together, the ij^{th} element of the result is equal to the inner product of the i^{th} row of the first matrix with the j^{th} column of the second matrix. Thus, if $\boldsymbol{C} = \boldsymbol{AB}$,

$$C_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}. \tag{2.27}$$

For (2.27) to make sense, we must assume that $\boldsymbol{A}$ has m columns and that $\boldsymbol{B}$ has m rows. In general, if two matrices are to be conformable for multiplication, the first matrix must have as many columns as the second has rows. Further, as is clear from (2.27), the result has as many rows as the first matrix and as many columns as the second. One way to make this explicit is to write something like

$$\underset{n\times m}{\boldsymbol{A}} \quad \underset{m\times l}{\boldsymbol{B}} \quad = \quad \underset{n\times l}{\boldsymbol{C}}.$$

One rarely sees this type of notation in a book or journal article. However, it is often useful to employ it when doing calculations, in order to verify that the matrices being multiplied are indeed conformable and to derive the dimensions of their product.

The rules for multiplying matrices and vectors together are the same as the rules for multiplying matrices with each other; vectors are simply treated as matrices that have only one column or only one row. For instance, if we multiply an n–vector $\boldsymbol{a}$ by the transpose of an n–vector $\boldsymbol{b}$, we obtain what is called the **outer product** of the two vectors. The result, written as $\boldsymbol{ab}^\top$, is an $n \times n$ matrix with typical element $a_i b_j$.

Matrix multiplication is, in general, not commutative. The fact that it is possible to **premultiply** $\boldsymbol{B}$ by $\boldsymbol{A}$ does not imply that it is possible to **postmultiply** $\boldsymbol{B}$ by $\boldsymbol{A}$. In fact, it is easy to see that both operations are possible if and only if one of the matrix products is square, in which case the other matrix product is square also, although generally with different dimensions. Even when both operations are possible, $\boldsymbol{AB} \neq \boldsymbol{BA}$ except in special cases.

A special matrix that econometricians frequently make use of is $\mathbf{I}$, which denotes the **identity matrix**. It is a diagonal matrix with every diagonal element equal to 1. A subscript is sometimes used to indicate the number of rows and columns. Thus

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The identity matrix is so called because when it is either premultiplied or postmultiplied by any matrix, it leaves the latter unchanged. Thus, for any matrix $\boldsymbol{A}$, $\boldsymbol{AI} = \mathbf{I}\boldsymbol{A} = \boldsymbol{A}$, provided, of course, that the matrices are conformable for multiplication. It is easy to see why the identity matrix has this property. Recall that the only nonzero elements of $\mathbf{I}$ are equal to 1 and are on the principal diagonal. This fact can be expressed simply with the help of the symbol known as the **Kronecker delta**, written as δ_{ij}. The definition is

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.28}$$

The ij^{th} element of $\mathbf{I}$ is just δ_{ij}. By (2.27), the ij^{th} element of $\boldsymbol{AI}$ is

$$\sum_{k=1}^{m} A_{ik}\mathrm{I}_{kj} = \sum_{k=1}^{m} A_{ik}\delta_{kj} = A_{ij},$$

since all the terms in the sum over k vanish except that for which $k = j$.

A special vector that we frequently use in this book is $\boldsymbol{\iota}$. It denotes a column vector every element of which is 1. This special vector comes in handy whenever one wishes to sum the elements of another vector, because, for any n–vector $\boldsymbol{b}$,

$$\boldsymbol{\iota}^\top \boldsymbol{b} = \sum_{i=1}^{n} b_i. \tag{2.29}$$

Matrix multiplication and matrix addition interact in an intuitive way. It is easy to check from the definitions of the respective operations that the **distributive** properties hold. That is, assuming that the dimensions of the matrices are conformable for the various operations,

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) &= \boldsymbol{AB} + \boldsymbol{AC}, \quad \text{and} \\ (\boldsymbol{B} + \boldsymbol{C})\boldsymbol{A} &= \boldsymbol{BA} + \boldsymbol{CA}. \end{aligned}$$

In addition, both operations are **associative**, which means that

$$\begin{aligned} (\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} &= \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}), \quad \text{and} \\ (\boldsymbol{AB})\boldsymbol{C} &= \boldsymbol{A}(\boldsymbol{BC}). \end{aligned}$$

The transpose of the product of two matrices is the product of the transposes of the matrices with the order reversed. Thus

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top. \tag{2.30}$$

The reversal of the order is necessary if the transposed matrices are to be

Copyright © 2021, Russell Davidson and James G. MacKinnon

conformable for multiplication. The result (2.30) can be proved immediately by writing out the typical entries of both sides and checking that

$$(\boldsymbol{AB})^\top_{ij} = (\boldsymbol{AB})_{ji} = \sum_{k=1}^{m} A_{jk}B_{ki} = \sum_{k=1}^{m} (\boldsymbol{B}^\top)_{ik}(\boldsymbol{A}^\top)_{kj} = (\boldsymbol{B}^\top\boldsymbol{A}^\top)_{ij},$$

where m is the number of columns of $\boldsymbol{A}$ and the number of rows of $\boldsymbol{B}$. It is always possible to multiply a matrix by its own transpose: If $\boldsymbol{A}$ is $n \times m$, then $\boldsymbol{A}^\top$ is $m \times n$, $\boldsymbol{A}^\top\boldsymbol{A}$ is $m \times m$, and $\boldsymbol{AA}^\top$ is $n \times n$. It follows directly from (2.30) that both of these matrix products are symmetric:

$$\boldsymbol{A}^\top\boldsymbol{A} = (\boldsymbol{A}^\top\boldsymbol{A})^\top \quad \text{and} \quad \boldsymbol{AA}^\top = (\boldsymbol{AA}^\top)^\top.$$

It is frequently necessary to multiply a matrix, say $\boldsymbol{B}$, by a scalar, say α. **Multiplication by a scalar** works exactly the way one would expect: Every element of $\boldsymbol{B}$ is multiplied by α. Since multiplication by a scalar is commutative, we can write this either as $\alpha\boldsymbol{B}$ or as $\boldsymbol{B}\alpha$, but $\alpha\boldsymbol{B}$ is the more common notation.

Occasionally, it is necessary to multiply two matrices together element by element. The result is called the **direct product**, the **Hadamard product**, or the **Schur product** of the two matrices. The direct product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is denoted $\boldsymbol{A}*\boldsymbol{B}$, and a typical element of it is equal to $A_{ij}B_{ij}$.

A square matrix may or may not be **invertible**. If $\boldsymbol{A}$ is invertible, then it has an **inverse matrix** $\boldsymbol{A}^{-1}$ with the property that

$$\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \mathbf{I}.$$

If $\boldsymbol{A}$ is symmetric, then so is $\boldsymbol{A}^{-1}$. If $\boldsymbol{A}$ is triangular, then so is $\boldsymbol{A}^{-1}$. Except in certain special cases, it is not easy to calculate the inverse of a matrix by hand. One such special case is that of a diagonal matrix, say $\boldsymbol{D}$, with typical diagonal element D_{ii}. It is easy to verify that $\boldsymbol{D}^{-1}$ is also a diagonal matrix, with typical diagonal element D_{ii}^{-1}.

If an $n \times n$ square matrix $\boldsymbol{A}$ is invertible, then its **rank** is n. Such a matrix is said to have **full rank**. If a square matrix does not have full rank, and therefore is not invertible, it is said to be **singular**. If a square matrix is singular, its rank must be less than its dimension. If, by omitting j rows and j columns of $\boldsymbol{A}$, we can obtain a matrix $\boldsymbol{A}'$ that is invertible, and if j is the smallest number for which this is true, then the rank of $\boldsymbol{A}$ is $n - j$. More generally, for matrices that are not necessarily square, the rank is the largest number m for which an $m \times m$ nonsingular matrix can be constructed by omitting some rows and some columns from the original matrix. The rank of a matrix is closely related to the geometry of vector spaces, which will be discussed in Section 3.2.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Regression Models and Matrix Notation

The simple linear regression model (2.01) can easily be written in matrix notation. If we stack the model for all the observations, we obtain

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_1 + u_1 \\ y_2 &= \beta_1 + \beta_2 x_2 + u_2 \\ \vdots \quad & \quad \vdots \quad \quad \vdots \quad \quad \vdots \\ y_n &= \beta_1 + \beta_2 x_n + u_n \,. \end{aligned} \tag{2.31}$$

Let $\boldsymbol{y}$ denote an n–vector with typical element y_t, $\boldsymbol{u}$ an n–vector with typical element u_t, $\boldsymbol{X}$ an $n \times 2$ matrix that consists of a column of 1s and a column with typical element x_t, and $\boldsymbol{\beta}$ a 2–vector with typical element β_i, $i = 1, 2$. Thus we have

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Equations (2.31) can now be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{2.32}$$

It is easy to verify from the rules of matrix multiplication that a typical row of (2.32) is a typical row of (2.31). When we postmultiply the matrix $\boldsymbol{X}$ by the vector $\boldsymbol{\beta}$, we obtain a vector $\boldsymbol{X}\boldsymbol{\beta}$ with typical element $\beta_1 + \beta_2 x_t$.

When a regression model is written in the form (2.32), the separate columns of the matrix $\boldsymbol{X}$ are called **regressors**, and the column vector $\boldsymbol{y}$ is called the **regressand**. In (2.31), there are just two regressors, corresponding to the constant and one explanatory variable. One advantage of writing the regression model in the form (2.32) is that we are not restricted to just one or two regressors. Suppose that we have k regressors, one of which may or may not correspond to a constant, and the others to a number of explanatory variables. Then the matrix $\boldsymbol{X}$ becomes

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \tag{2.33}$$

where x_{ti} denotes the t^{th} observation on the i^{th} regressor, and the vector $\boldsymbol{\beta}$ now has k elements, β_1 through β_k. Equation (2.32) remains perfectly valid

Copyright © 2021, Russell Davidson and James G. MacKinnon

when $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are redefined in this way. A typical row of this equation is

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t = \sum_{i=1}^{k} \beta_i x_{ti} + u_t, \tag{2.34}$$

where we have used $\boldsymbol{X}_t$ to denote row t of $\boldsymbol{X}$.

In equation (2.32), we used the rules of matrix multiplication to write the regression function, for the entire sample, in a very simple form. These rules make it possible to find equally convenient expressions for other aspects of regression models. The key fact is that every element of the product of two matrices is a summation. Thus it is often very convenient to use matrix algebra when dealing with summations. Consider, for example, the matrix of sums of squares and cross-products of the $\boldsymbol{X}$ matrix. This is a $k \times k$ symmetric matrix, of which a typical element is either

$$\sum_{t=1}^{n} x_{ti}^2 \quad \text{or} \quad \sum_{t=1}^{n} x_{ti} x_{tj},$$

the former being a typical diagonal element and the latter a typical off-diagonal one. This entire matrix can be written very compactly as $\boldsymbol{X}^\top\boldsymbol{X}$. Similarly, the vector with typical element

$$\sum_{t=1}^{n} x_{ti} y_t$$

can be written as $\boldsymbol{X}^\top\boldsymbol{y}$. As we will see in the next section, the least-squares estimates of $\boldsymbol{\beta}$ depend only on the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ and the vector $\boldsymbol{X}^\top\boldsymbol{y}$.

Partitioned Matrices

There are many ways of writing an $n \times k$ matrix $\boldsymbol{X}$ that are intermediate between the straightforward notation $\boldsymbol{X}$ and the full element-by-element decomposition of $\boldsymbol{X}$ given in (2.33). We might wish to separate the columns while grouping the rows, as

$$\underset{n \times k}{\boldsymbol{X}} = \underset{n \times 1}{\begin{bmatrix} \boldsymbol{x}_1 \end{bmatrix}} \underset{n \times 1}{\begin{bmatrix} \boldsymbol{x}_2 \end{bmatrix}} \cdots \underset{n \times 1}{\begin{bmatrix} \boldsymbol{x}_k \end{bmatrix}},$$

or we might wish to separate the rows but not the columns, as

$$\underset{n \times k}{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} \begin{matrix} 1 \times k \\ 1 \times k \\ \\ 1 \times k \end{matrix}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

To save space, we can also write this as $\boldsymbol{X} = \left[\boldsymbol{X}_1 \vdots \boldsymbol{X}_2 \vdots \ldots \vdots \boldsymbol{X}_n\right]$. There is no restriction on how a matrix can be partitioned, so long as all the **submatrices** or **blocks** fit together correctly. Thus we might have

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \end{bmatrix} \begin{matrix} n_1 \\ n_2 \end{matrix} \\ \begin{matrix} k_1 & k_2 \end{matrix}$$

or, equivalently, $\boldsymbol{X} = [\boldsymbol{X}_{11} \vdots \boldsymbol{X}_{21} \quad \boldsymbol{X}_{12} \vdots \boldsymbol{X}_{22}]$. Here the submatrix $\boldsymbol{X}_{11}$ has dimensions $n_1 \times k_1$, $\boldsymbol{X}_{12}$ has dimensions $n_1 \times k_2$, $\boldsymbol{X}_{21}$ has dimensions $n_2 \times k_1$, and $\boldsymbol{X}_{22}$ has dimensions $n_2 \times k_2$, with $n_1 + n_2 = n$ and $k_1 + k_2 = k$. Thus $\boldsymbol{X}_{11}$ and $\boldsymbol{X}_{12}$ have the same number of rows, and also $\boldsymbol{X}_{21}$ and $\boldsymbol{X}_{22}$, as required for the submatrices to fit together horizontally. Similarly, $\boldsymbol{X}_{11}$ and $\boldsymbol{X}_{21}$ have the same number of columns, and also $\boldsymbol{X}_{12}$ and $\boldsymbol{X}_{22}$, as required for the submatrices to fit together vertically as well.

If two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimensions are partitioned in exactly the same way, they can be added or subtracted block by block. A simple example is

$$\boldsymbol{A} + \boldsymbol{B} = [\,\boldsymbol{A}_1 \quad \boldsymbol{A}_2\,] + [\,\boldsymbol{B}_1 \quad \boldsymbol{B}_2\,] = [\,\boldsymbol{A}_1 + \boldsymbol{B}_1 \quad \boldsymbol{A}_2 + \boldsymbol{B}_2\,],$$

where $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ have the same dimensions, as do $\boldsymbol{A}_2$ and $\boldsymbol{B}_2$.

More interestingly, as we now explain, matrix multiplication can sometimes be performed block by block on partitioned matrices. If the product $\boldsymbol{AB}$ exists, then $\boldsymbol{A}$ has as many columns as $\boldsymbol{B}$ has rows. Now suppose that the columns of $\boldsymbol{A}$ are partitioned in the same way as the rows of $\boldsymbol{B}$. Then

$$\boldsymbol{AB} = [\,\boldsymbol{A}_1 \quad \boldsymbol{A}_2 \quad \cdots \quad \boldsymbol{A}_p\,] \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_p \end{bmatrix}.$$

Here each $\boldsymbol{A}_i$, $i = 1, \ldots, p$, has as many columns as the corresponding $\boldsymbol{B}_i$ has rows. The product can be computed following the usual rules for matrix multiplication just as though the blocks were scalars, yielding the result

$$\boldsymbol{AB} = \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{B}_i. \tag{2.35}$$

To see this, it is enough to compute the typical element of each side of equation (2.35) directly and observe that they are the same. Matrix multiplication can also be performed block by block on matrices that are partitioned both horizontally and vertically, provided all the submatrices are conformable; see Exercise 2.19.

Copyright © 2021, Russell Davidson and James G. MacKinnon

These results on multiplying partitioned matrices lead to a useful corollary. Suppose that we are interested only in the first m rows of a product $\boldsymbol{AB}$, where $\boldsymbol{A}$ has more than m rows. Then we can partition the rows of $\boldsymbol{A}$ into two blocks, the first with m rows, the second with all the rest. We need not partition $\boldsymbol{B}$ at all. Then

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_1 \boldsymbol{B} \\ \boldsymbol{A}_2 \boldsymbol{B} \end{bmatrix}. \tag{2.36}$$

This works because $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ both have the full number of columns of $\boldsymbol{A}$, which must be the same as the number of rows of $\boldsymbol{B}$, since $\boldsymbol{AB}$ exists. It is clear from the rightmost expression in (2.36) that the first m rows of $\boldsymbol{AB}$ are given by $\boldsymbol{A}_1\boldsymbol{B}$. In order to obtain any subset of the rows of a matrix product of arbitrarily many factors, the rule is that we take the submatrix of the leftmost factor that contains just the rows we want, and then multiply it by all the other factors unchanged. Similarly, if we want to select a subset of columns of a matrix product, we can just select them from the rightmost factor, leaving all the factors to the left unchanged.

2.5 Techniques of Estimation

Almost all econometric models contain unknown parameters. For most of the uses to which such models can be put, it is necessary to have **estimates** of these parameters. To compute parameter estimates, we need both a model containing the parameters and a sample made up of observed data. If the model is correctly specified, it describes the real-world mechanism which generated the data in our sample.

It is common in statistics to speak of the "population" from which a sample is drawn. Recall the use of the term "population mean" as a synonym for the mathematical term "expectation"; see Section 2.2. The expression is a holdover from the time when statistics was biostatistics, and the object of study was the human population, usually that of a specific town or country, from which random samples were drawn by statisticians for study. The average weight of all members of the population, for instance, would then be estimated by the mean of the weights of the individuals in the sample, that is, by the **sample mean** of individuals' weights. The sample mean was thus an estimate of the **population mean**. The underlying idea is just that the sample *represents* the population from which it has been drawn.

In econometrics, the use of the term population is simply a metaphor. A better concept is that of a **data-generating process**, or **DGP**. By this term, we mean whatever mechanism is at work in the real world of economic activity giving rise to the numbers in our samples, that is, precisely the mechanism that our econometric model is supposed to describe. A data-generating process is thus the analog in econometrics of a population in biostatistics. Samples may be

Copyright © 2021, Russell Davidson and James G. MacKinnon

drawn from a DGP just as they may be drawn from a population. In both cases, the samples are assumed to be representative of the DGP or population from which they are drawn.

A very natural way to estimate parameters is to replace population means by sample means. This technique is called the **method of moments**, and it is one of the most widely-used estimation methods in statistics. As the name implies, it can be used with moments other than the expectation. In general, the method of moments estimates population moments by the corresponding sample moments. However, we cannot apply the method of moments directly to regression models, because, except in one trivial case that we discuss first, the parameters we wish to estimate are not population means.

The simplest possible linear regression has only one regressor, namely the constant. For each observation in the sample, the model gives just $y_t = \beta + u_t$, from which we see that $\mathrm{E}(y_t) = \beta$ for all t. The method of moments applies directly, and we define the **estimator** of β by the sample mean of the y_t:

$$\hat{\beta} = \frac{1}{n}\sum_{t=1}^{n} y_t. \tag{2.37}$$

Econometricians generally make a distinction between an **estimate**, which is simply a number used to estimate some parameter, normally based on a particular data set, and an **estimator**, which is a rule, such as (2.45), for obtaining estimates from any set of data. More formally, an estimator is a random variable, and an estimate is a realization of the random variable.

Here is a slightly less direct way of arriving at the estimator (2.37). The disturbance for observation t is $u_t = y_t - \beta$, and, according to our model, the expectation of this disturbance is zero. Since we have n disturbances for a sample of size n, we can consider their sample mean:

$$\frac{1}{n}\sum_{t=1}^{n} u_t = \frac{1}{n}\sum_{t=1}^{n}(y_t - \beta)$$

It seems natural to seek a parameter estimator which ensures that this sample mean is equal to zero, the population mean. The equation that defines this estimator is therefore

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \beta) = 0. \tag{2.38}$$

Since β is common to all the observations and thus does not depend on the index t, equation (2.38) can be written as

$$\frac{1}{n}\sum_{t=1}^{n} y_t - \beta = 0.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

It is clear that the solution for β to this equation is (2.37).

Now consider the simple model (2.01). It is not obvious how to use the method of moments when we have two parameters, β_1 and β_2. Equation (2.38) would become

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \beta_1 - \beta_2 x_t) = 0, \tag{2.39}$$

but this is just one equation, and there are two unknowns. In order to obtain another equation, we can use the fact that our model specifies that the expectation of u_t is 0 *conditional* on the explanatory variable x_t. Actually, it may well specify that the expectation of u_t is 0 conditional on many other things as well, depending on our choice of the information set Ω_t, but we will ignore this for now. The conditional expectation assumption implies not only that $\mathrm{E}(u_t) = 0$, but also that $\mathrm{E}(x_t u_t) = 0$, since, by (2.16) and (2.17),

$$\mathrm{E}(x_t u_t) = \mathrm{E}\big(\mathrm{E}(x_t u_t \,|\, x_t)\big) = \mathrm{E}\big(x_t \mathrm{E}(u_t \,|\, x_t)\big) = 0. \tag{2.40}$$

Thus we can supplement (2.39) by the following equation, which replaces the population mean in (2.40) by the corresponding sample mean,

$$\frac{1}{n}\sum_{t=1}^{n} x_t(y_t - \beta_1 - \beta_2 x_t) = 0. \tag{2.41}$$

The equations (2.39) and (2.41) are two linear equations in two unknowns, β_1 and β_2. Except in rare conditions, which can easily be ruled out, they have a unique solution that is not difficult to calculate. Solving these equations yields an estimator.

We could just solve (2.39) and (2.41) directly, but it is far more illuminating to rewrite them in matrix form. Since β_1 and β_2 do not depend on t, these two equations can be written as

$$\begin{aligned}
\beta_1 + \left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} y_t \\
\left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)\beta_1 + \left(\frac{1}{n}\sum_{t=1}^{n} x_t^2\right)\beta_2 &= \frac{1}{n}\sum_{t=1}^{n} x_t y_t.
\end{aligned}$$

Multiplying both equations by n and using the rules of matrix multiplication that were discussed in Section 3.4, we can also write them as

$$\begin{bmatrix} n & \sum_{t=1}^{n} x_t \\ \sum_{t=1}^{n} x_t & \sum_{t=1}^{n} x_t^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} x_t y_t \end{bmatrix}. \tag{2.42}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Equations (2.42) can be rewritten much more compactly. As we saw in Section 2.3, the model (2.01) is simply a special case of the **multiple linear regression model**

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{2.43}$$

where the n–vector $\boldsymbol{y}$ has typical element y_t, the k–vector $\boldsymbol{\beta}$ has typical element β_i, and, in general, the matrix $\boldsymbol{X}$ is $n \times k$. In this case, $\boldsymbol{X}$ is $n \times 2$; it can be written as $\boldsymbol{X} = [\boldsymbol{\iota} \;\; \boldsymbol{x}]$, where $\boldsymbol{\iota}$ denotes a column of 1s, and $\boldsymbol{x}$ denotes a column with typical element x_t. Thus, recalling (2.29), we see that

$$\boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} \sum_{t=1}^n y_t \\ \sum_{t=1}^n x_t y_t \end{bmatrix}$$

and

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} n & \sum_{t=1}^n x_t \\ \sum_{t=1}^n x_t & \sum_{t=1}^n x_t^2 \end{bmatrix}.$$

These are the principal quantities that appear in the equations (2.42). Thus it is clear that we can rewrite those equations as

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.44}$$

The estimator $\hat{\boldsymbol{\beta}}$ that solves these equations can be written as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.45}$$

It is generally called the **ordinary least squares**, or **OLS**, estimator for the linear regression model, for a reason explained shortly. It is clear that $\hat{\boldsymbol{\beta}}$ is well defined only if $\boldsymbol{X}^\top \boldsymbol{X}$ can be inverted.

The formula (2.45) gives us the OLS estimator for the simple linear regression model (2.01), but in fact it does far more than that. As we now show, it also gives us an estimator for the multiple linear regression model (2.43). Since each of the explanatory variables is required to be in the information set Ω_t, we have, for $i = 1, \ldots, k$,

$$\mathrm{E}(x_{ti}\, u_t) = 0.$$

In the corresponding sample mean form, this yields

$$\frac{1}{n} \sum_{t=1}^n x_{ti}(y_t - \boldsymbol{X}_t \boldsymbol{\beta}) = 0; \tag{2.46}$$

recall from equation (2.34) that $\boldsymbol{X}_t$ denotes the t^{th} row of $\boldsymbol{X}$. As i varies from 1 to k, equation (2.46) yields k equations for the k unknown components of $\boldsymbol{\beta}$. In most cases, there is a constant, which we may take to be the first regressor. If so, $x_{t1} = 1$, and the first of these equations simply says that the sample mean of the disturbances is 0.

Copyright © 2021, Russell Davidson and James G. MacKinnon

In matrix form, after multiplying them by n, the k equations of (2.46) can be written as

$$\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}. \tag{2.47}$$

The notation $\boldsymbol{0}$ in bold-face type is used to signify a **zero vector**, here a k–vector, each element of which is zero. Equations (2.47) are clearly equivalent to equations (2.44). Thus solving them yields the estimator (2.45), which applies no matter what the number of regressors.

It is easy to see that the OLS estimator (2.45) depends on $\boldsymbol{y}$ and $\boldsymbol{X}$ exclusively through a number of scalar products. Each column $\boldsymbol{x}_i$ of the matrix $\boldsymbol{X}$ corresponds to one of the regressors, as does each row $\boldsymbol{x}_i^\top$ of the transposed matrix $\boldsymbol{X}^\top$. Thus we can write $\boldsymbol{X}^\top \boldsymbol{y}$ as

$$\boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_k^\top \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} \boldsymbol{x}_1^\top \boldsymbol{y} \\ \boldsymbol{x}_2^\top \boldsymbol{y} \\ \vdots \\ \boldsymbol{x}_k^\top \boldsymbol{y} \end{bmatrix}.$$

The elements of the rightmost expression here are just the scalar products of the regressors $\boldsymbol{x}_i$ with the regressand $\boldsymbol{y}$. Similarly, we can write $\boldsymbol{X}^\top \boldsymbol{X}$ as

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_k^\top \end{bmatrix} [\, \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_k \,] = \begin{bmatrix} \boldsymbol{x}_1^\top \boldsymbol{x}_1 & \boldsymbol{x}_1^\top \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_1^\top \boldsymbol{x}_k \\ \boldsymbol{x}_2^\top \boldsymbol{x}_1 & \boldsymbol{x}_2^\top \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_2^\top \boldsymbol{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_k^\top \boldsymbol{x}_1 & \boldsymbol{x}_k^\top \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_k^\top \boldsymbol{x}_k \end{bmatrix}.$$

Once more, all the elements of the rightmost expression are scalar products of pairs of regressors. Since $\boldsymbol{X}^\top \boldsymbol{X}$ can be expressed exclusively in terms of scalar products of the variables of the regression, the same is true of its inverse, the elements of which are in general complicated functions of those scalar products. Thus $\hat{\boldsymbol{\beta}}$ is a function solely of scalar products of pairs of variables.

Estimating Functions

The technique we used to derive an estimator for the model (2.43) can be generalized. Consider a model the DGPs of which are characterised, either completely or partially, by a vector $\boldsymbol{\beta}$ of parameters. Suppose that, in order to estimate $\boldsymbol{\beta}$, there are data available. For instance, for the multiple linear regression model, there would be an n-vector $\boldsymbol{y}$ with observations on the dependent variable, and an $n \times k$ matrix $\boldsymbol{X}$ with the regressors. A function $f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})$ of the data and the model parameters is called a **zero function** (for the given model) if, for each DGP contained in the model, μ say, characterised by parameters $\boldsymbol{\beta}_\mu$, the expectation $\mathrm{E}_\mu\big(f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}_\mu)\big) = 0$ when $\boldsymbol{y}$ and also possibly $\boldsymbol{X}$, are generated by the DGP μ, as indicated by the notation E_μ.

Often, one can define a separate zero function for each observation in the sample. The example of this we looked at is the **residual** for that observation.

Copyright © 2021, Russell Davidson and James G. MacKinnon

For the multiple linear regression model, the residual for observation t is $u_t(y_t, \boldsymbol{X}_t, \boldsymbol{\beta}) = y_t - \boldsymbol{X}_t\boldsymbol{\beta}$, and, if for a given DGP μ the true parameter vector is $\boldsymbol{\beta}_\mu$, this means that $y_t = \boldsymbol{X}_t\boldsymbol{\beta}_\mu + u_t$, where u_t is the disturbance associated with observation t. Since u_t has expectation zero, it follows that $\mathrm{E}_\mu\big(u_t(y_t, \boldsymbol{X}_t, \boldsymbol{\beta}_\mu)\big) = \mathrm{E}(u_t) = 0$. In such a case, the residual $u_t(y_t, \boldsymbol{X}_t, \boldsymbol{\beta})$ is called an **elementary zero function**, the word elementary signifying that it is specific to the single observation t.

A powerful estimation method makes use of zero functions, usually linear combinations of the elementary zero functions of the model. The OLS estimator uses the linear combinations of the residuals defined with the regressors as coefficients, that is, the k components of the vector

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{t=1}^{n} \boldsymbol{X}_t{}^\top(y_t - \boldsymbol{X}_t\boldsymbol{\beta}) = \sum_{t=1}^{n} \boldsymbol{X}_t u_t(y_t, \boldsymbol{X}_t, \boldsymbol{\beta}).$$

Such linear combinations of the elementary zero functions are called **estimating functions**, and of course they too are zero functions. We get an **estimating equation** by setting an estimating function equal to zero. An estimator is thereby defined when there are as many estimating equations as there are parameters to estimate. The equations implicit in (2.47) are the **OLS estimating equations**.

Least-Squares Estimation

We have derived the OLS estimator (2.45) by using the method of estimating functions. Deriving it in this way has at least two major advantages. Firstly, this method is a very general and very powerful principle of estimation, one that we will encounter again and again throughout this book. Secondly, by using estimating functions, we were able to obtain (2.45) without making any use of calculus. However, as we have already remarked, (2.45) is generally referred to as the ordinary least-squares estimator. It is interesting to see why this is so.

For the multiple linear regression model (2.43), the expression $y_t - \boldsymbol{X}_t\boldsymbol{\beta}$ is equal to the disturbance for the t^{th} observation, but only if the correct value of the parameter vector $\boldsymbol{\beta}$ is used. If the same expression is thought of as a function of $\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ allowed to vary arbitrarily, then above we called it the residual associated with the t^{th} observation. Similarly, the n–vector $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ is called the **residual vector**. The sum of the squares of the components of that vector is called the **sum of squared residuals**, or **SSR**. Since this sum is a scalar, the sum of squared residuals is a scalar-valued function of the k–vector $\boldsymbol{\beta}$:

$$\mathrm{SSR}(\boldsymbol{\beta}) = \sum_{t=1}^{n} (y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2. \tag{2.48}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The notation here emphasizes the fact that this function can be computed for arbitrary values of the argument $\boldsymbol{\beta}$ purely in terms of the observed data $\boldsymbol{y}$ and $\boldsymbol{X}$.

The idea of **least-squares** estimation is to minimize the sum of squared residuals associated with a regression model. At this point, it may not be at all clear why we would wish to do such a thing. However, it can be shown that the parameter vector $\hat{\boldsymbol{\beta}}$ which minimizes (2.48) is the same as the estimator (2.45). This being so, we will regularly use the traditional terminology associated with linear regressions, based on least squares. Thus, the parameter estimates which are the components of the vector $\hat{\boldsymbol{\beta}}$ that minimizes the SSR (2.48) are called the **least-squares estimates**, and the corresponding vector of residuals is called the vector of **least-squares residuals**. When least squares is used to estimate a linear regression model like (2.01), it is called **ordinary least-squares**, or **OLS**, to distinguish it from other varieties of least squares that we will encounter later, such as generalized least squares (Chapter 8).

Consider briefly the simplest case of (2.01), in which $\beta_2 = 0$ and the model contains only a constant term. Expression (2.48) becomes

$$\mathrm{SSR}(\beta_1) = \sum_{t=1}^{n} (y_t - \beta_1)^2 = \sum_{t=1}^{n} y_t^2 + n\beta_1^2 - 2\beta_1 \sum_{t=1}^{n} y_t. \tag{2.49}$$

Differentiating the rightmost expression in equations (2.49) with respect to β_1 and setting the derivative equal to zero gives the following first-order condition for a minimum of the sum of squared residuals:

$$\frac{\partial \mathrm{SSR}}{\partial \beta_1} = 2\beta_1 n - 2\sum_{t=1}^{n} y_t = 0. \tag{2.50}$$

For this simple model, the matrix $\boldsymbol{X}$ consists solely of the constant vector, $\boldsymbol{\iota}$. Therefore, by (2.29), $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{\iota}^\top\boldsymbol{\iota} = n$, and $\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{\iota}^\top\boldsymbol{y} = \sum_{t=1}^{n} y_t$. Thus, if the first-order condition (2.50) is multiplied by one-half, it can be rewritten as $\boldsymbol{\iota}^\top\boldsymbol{\iota}\beta_1 = \boldsymbol{\iota}^\top\boldsymbol{y}$, which is clearly just a special case of (2.44). Solving (2.50) for β_1 yields the sample mean of the y_t,

$$\hat{\beta}_1 = \frac{1}{n}\sum_{t=1}^{n} y_t = (\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^\top\boldsymbol{y}. \tag{2.51}$$

We already saw, in equation (2.37), that this is the MM estimator for the model with $\beta_2 = 0$. The rightmost expression in (2.51) makes it clear that the sample mean is just a special case of the famous formula (2.45).

Not surprisingly, the OLS estimator is equivalent to the estimating-function estimator (2.45) for the multiple linear regression model as well. For this model,

$$\mathrm{SSR}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.52}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

If this inner product is written out in terms of the scalar components of $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{\beta}$, then the first-order conditions for minimizing the SSR (2.52) can be written as (2.44); see Exercise 2.22. Thus (2.45) provides a general formula for the OLS estimator $\hat{\boldsymbol{\beta}}$ in the multiple linear regression model.

Estimators like that obtained by minimizing a sum of squared residuals are called the **M-estimators**, which are defined as the parameter values that maximize or minimize (hence the 'M') a **criterion function**. The OLS estimator has traditionally been defined as an M-estimator with as criterion function the sum of squared residuals. But, as we have seen, it can also be defined using estimation equations, in which estimating functions are set equal to zero. Such estimators are called **Z-estimators**, with 'Z' for zero.

Final Remarks

We have seen that it is perfectly easy to obtain an algebraic expression, (2.45), for the OLS estimator $\hat{\boldsymbol{\beta}}$. With modern computers and appropriate software, it is also easy to obtain OLS estimates numerically, even for regressions with millions of observations and dozens of explanatory variables; the time-honored term for doing so is "running a regression." What is not so easy, and will occupy us for most of the next four chapters, is to understand the properties of these estimates.

We will be concerned with two types of properties. The first type, **numerical properties**, arise as a consequence of the way that OLS estimates are obtained. These properties hold for every set of OLS estimates, no matter how the data were generated. That they hold for any data set can easily be verified by direct calculation. The numerical properties of OLS will be discussed in Chapter 3. The second type, **statistical properties**, depend on the way in which the data were generated. They can be verified theoretically, under certain assumptions, and they can be illustrated by simulation, but we can never prove that they are true for any given data set. The statistical properties of OLS will be discussed in detail in Chapters 4, 5, and 6.

Readers who seek a deeper treatment of the topics dealt with in the first two sections may wish to consult a text on mathematical statistics, such as Mittelhammer (2013), Hogg, McKean, and Craig (2007), Shao (2007), Schervish (1996), or Gallant (1997).

2.6 Notes on the Exercises

Each chapter of this book is followed by a set of exercises. These exercises are of various sorts, and they have various intended functions. Some are, quite simply, just for practice. Others have a tidying-up function. Details left out of the discussions in the main text are taken up, and conscientious readers can check that unproved claims made in the text are in fact justified. Some of

Copyright © 2021, Russell Davidson and James G. MacKinnon

these exercises are particularly challenging. They are starred, and solutions to them are available on the book's website.

A number of exercises serve chiefly to extend the material presented in the chapter. In many cases, the new material in such exercises recurs later in the book, and it is hoped that readers who have worked through them will follow later discussions more easily. A case in point concerns the **bootstrap**. Some of the exercises in this chapter and the next two are designed to familiarize readers with the tools that are used to implement the bootstrap, so that, when it is introduced formally in Chapter 7, the bootstrap will appear as a natural development.

Many of the exercises require the reader to make use of a computer, sometimes to compute estimates and test statistics using real or simulated data, and sometimes for the purpose of doing simulations. There are a great many computer packages that are capable of doing the things we ask for in the exercises, and it seems unnecessary to make any specific recommendations as to what software would be best. Besides, we expect that many readers will already have developed their own personal preferences for software packages, and we know better than to try to upset such preferences.

Some exercises require, not only a computer, but also actual (or simulated) economic data. It cannot be stressed enough that econometrics is an empirical discipline, and that the analysis of economic data is its *raison d'être*. All of the data files needed for the exercises are available from the website for this book. The address is

http://qed.econ.queensu.ca/ETM/

This website will ultimately contain corrections and updates to the book as well as the data and the solutions to the starred exercises.

2.7 Exercises

2.1 Consider a sample of n observations, $y_1, y_2, \ldots, y_n$, on some random variable Y. The **empirical distribution function**, or **EDF**, of this sample is a discrete distribution with n possible points. These points are just the n observed points, $y_1, y_2, \ldots, y_n$. Each point is assigned the same probability, which is just $1/n$, in order to ensure that all the probabilities sum to 1.

Compute the expectation of the discrete distribution characterized by the EDF, and show that it is equal to the **sample mean**, that is, the unweighted average of the n sample points, $y_1, y_2, \ldots, y_n$.

2.2 A random variable computed as the ratio of two independent standard normal variables follows what is called the **Cauchy distribution**. It can be shown that the density of this distribution is

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Show that the Cauchy distribution has no first moment, which means that its expectation does not exist.

Use your favorite random number generator to generate samples of 10, 100, 1,000, and 10,000 drawings from the Cauchy distribution, and as many intermediate values of n as you have patience or computer time for. For each sample, compute the sample mean. Do these sample means seem to converge to zero as the sample size increases? Repeat the exercise with drawings from the standard normal density. Do these sample means tend to converge to zero as the sample size increases?

2.3 Consider two events A and B such that $A \subset B$. Compute $\Pr(A \mid B)$ in terms of $\Pr(A)$ and $\Pr(B)$. Interpret the result.

2.4 Prove **Bayes' Theorem**. This famous theorem states that, for any two events A and B with nonzero probabilities,

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}.$$

Another form of the theorem deals with two continuous random variables X_1 and X_2, which have a joint density $f(x_1, x_2)$. Show that, for any values x_1 and x_2 that are permissible for X_1 and X_2, respectively,

$$f(x_1 \mid x_2) = \frac{f(x_2 \mid x_1) f(x_1)}{f(x_2)}.$$

2.5 Suppose that X and Y are two binary random variables. Their joint distribution is given in the following table.

	$Y = 0$	$Y = 1$
$X = 0$	.16	.37
$X = 1$	.29	.18

What is the marginal distribution of Y? What is the distribution of Y conditional on $X = 0$? What is the distribution of Y conditional on $X = 1$?

Demonstrate the Law of Iterated Expectations explicitly by showing that $\mathrm{E}(\mathrm{E}(X \mid Y)) = \mathrm{E}(X)$. Let $h(Y) = Y^3$. Show explicitly that $\mathrm{E}(Xh(Y) \mid Y) = h(Y)\mathrm{E}(X \mid Y)$ in this case.

2.6 Using expression (2.06) for the density $\phi(x)$ of the standard normal distribution, show that the derivative of $\phi(x)$ is the function $-x\phi(x)$, and that the second derivative is $(x^2-1)\phi(x)$. Use these facts to show that the expectation of a standard normal random variable is 0, and that its variance is 1. These two properties account for the use of the term "standard."

2.7 A normally distributed random variable can have any expectation μ and any positive variance σ^2. Such a random variable is said to follow the $\mathrm{N}(\mu, \sigma^2)$ distribution. A standard normal variable therefore has the $\mathrm{N}(0, 1)$ distribution. Suppose that X has the standard normal distribution. Show that the random variable $Z \equiv \mu + \sigma X$ has expectation μ and variance σ^2.

2.8 Find the CDF of the $\mathrm{N}(\mu, \sigma^2)$ distribution in terms of $\Phi(\cdot)$, the CDF of the standard normal distribution. Differentiate your answer so as to obtain the density of $\mathrm{N}(\mu, \sigma^2)$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

2.9 If two random variables X_1 and X_2 are statistically independent, show that $\mathrm{E}(X_1 \mid X_2) = \mathrm{E}(X_1)$.

2.10 The **covariance** of two random variables X_1 and X_2, which is often written as $\mathrm{Cov}(X_1, X_2)$, is defined as the expectation of the product of $X_1 - \mathrm{E}(X_1)$ and $X_2 - \mathrm{E}(X_2)$. Consider a random variable X_1 with expectation zero. Show that the covariance of X_1 and any other random variable X_2, whether it has expectation zero or not, is just the expectation of the product of X_1 and X_2.

⋆**2.11** Show that the covariance of the random variable $\mathrm{E}(X_1 \mid X_2)$ and the random variable $X_1 - \mathrm{E}(X_1 \mid X_2)$ is zero. It is easiest to show this result by first showing that it is true when the covariance is computed conditional on X_2.

⋆**2.12** Show that the variance of the random variable $X_1 - \mathrm{E}(X_1 \mid X_2)$ cannot be greater than the variance of X_1, and that the two variances are equal if X_1 and X_2 are independent. This result shows how one random variable can be informative about another: Conditioning on it reduces variance unless the two variables are independent.

2.13 Prove that, if X_1 and X_2 are statistically independent, $\mathrm{Cov}(X_1, X_2) = 0$.

⋆**2.14** Let a random variable X_1 be distributed as $\mathrm{N}(0, 1)$. Now suppose that a second random variable, X_2, is constructed as the product of X_1 and an independent random variable Z, which equals 1 with probability $1/2$ and -1 with probability $1/2$.

What is the (marginal) distribution of X_2? What is the covariance between X_1 and X_2? What is the distribution of X_1 conditional on X_2?

2.15 Consider the linear regression models

$$H_1: \quad y_t = \beta_1 + \beta_2 x_t + u_t \quad \text{and}$$

$$H_2: \quad \log y_t = \gamma_1 + \gamma_2 \log x_t + u_t.$$

Suppose that the data are actually generated by H_2, with $\gamma_1 = 1.5$ and $\gamma_2 = 0.5$, and that the value of x_t varies from 10 to 110 with an average value of 60. Ignore the disturbances and consider the deterministic relations between y_t and x_t implied by the two models. Find the values of β_1 and β_2 that make the relation given by H_1 have the same level and the same value of $\mathrm{d}y_t/\mathrm{d}x_t$ as the level and value of $\mathrm{d}y_t/\mathrm{d}x_t$ implied by the relation given by H_2 when it is evaluated at the average value of the regressor.

Using the deterministic relations, plot y_t as a function of x_t for both models for $10 \leq x_t \leq 110$. Also plot $\log y_t$ as a function of $\log x_t$ for both models for the same range of x_t. How well do the two models approximate each other in each of the plots?

2.16 Consider two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of dimensions such that the product $\boldsymbol{AB}$ exists. Show that the i^{th} row of $\boldsymbol{AB}$ is the matrix product of the i^{th} row of $\boldsymbol{A}$ with the entire matrix $\boldsymbol{B}$. Show that this result implies that the i^{th} row of a product $\boldsymbol{ABC}\ldots$, with arbitrarily many factors, is the product of the i^{th} row of $\boldsymbol{A}$ with $\boldsymbol{BC}\ldots$.

What is the corresponding result for the columns of $\boldsymbol{AB}$? What is the corresponding result for the columns of $\boldsymbol{ABC}\ldots$?

2.17 Consider two invertible square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, of the same dimensions. Show that the inverse of the product $\boldsymbol{AB}$ exists and is given by the formula

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

This shows that there is a **reversal rule** for inverses as well as for transposes; see (2.30).

2.18 Show that the transpose of the product of an arbitrary number of factors is the product of the transposes of the individual factors in completely reversed order:

$$(\boldsymbol{ABC}\cdots)^\top = \cdots \boldsymbol{C}^\top \boldsymbol{B}^\top \boldsymbol{A}^\top.$$

Show also that an analogous result holds for the inverse of the product of an arbitrary number of factors.

2.19 Consider the following example of multiplying partitioned matrices:

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{bmatrix}.$$

Check all the expressions on the right-hand side, verifying that all products are well defined and that all sums are of matrices of the same dimensions.

2.20 Suppose that $\boldsymbol{X} = [\boldsymbol{\iota}\ \boldsymbol{X}_1\ \boldsymbol{X}_2]$, where $\boldsymbol{X}$ is $n \times k$, $\boldsymbol{\iota}$ is an n–vector of 1s, $\boldsymbol{X}_1$ is $n \times k_1$, and $\boldsymbol{X}_2$ is $n \times k_2$. What is the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ in terms of the components of $\boldsymbol{X}$? What are the dimensions of its component matrices? What is the element in the upper left-hand corner of $\boldsymbol{X}^\top\boldsymbol{X}$ equal to?

2.21 Fix a sample size of $n = 100$, and simulate the very simplest regression model, namely, $y_t = \beta + u_t$. Set $\beta = 1$, and let the disturbances u_t be drawings from the standard normal distribution. Compute the sample mean of the y_t,

$$\bar{y} \equiv \frac{1}{n}\sum_{t=1}^{n} y_t.$$

Use your favorite econometrics software package to run a regression with $\boldsymbol{y}$, the 100×1 vector with typical element y_t, as the dependent variable, and a constant as the sole explanatory variable. Show that the OLS estimate of the constant is equal to the sample mean. Why is this a necessary consequence of the formula (2.45)?

2.22 For the multiple linear regression model (2.43), the sum of squared residuals can be written as

$$\mathrm{SSR}(\boldsymbol{\beta}) = \sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Show that, if we minimize $\mathrm{SSR}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, the minimizing value of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$, the OLS estimator given by (2.45). The easiest way is to show that the first-order conditions for a minimum are exactly the equations (2.46), or (2.47), that arise from estimation using estimating functions. This can be done without using matrix calculus.

2.23 The file **house-price-data.txt** contains data on the sale prices and various characteristics of 546 houses sold in Windsor, Ontario, Canada in 1987, taken from Anglin and Gençay (1996). Regress the house price on a constant and the 11 explanatory variables.[5] By how much, on average, does the price of a house increase when the number of full bathrooms increases by one?

[5] This type of regression model is often called a **hedonic regression**, because it attempts to estimate the values that consumers place on various characteristics of a good, in this case a house.

Copyright © 2021, Russell Davidson and James G. MacKinnon

2.24 Plot the residuals from the regression of exercise 1.23 against the fitted values. Also, plot the squared residuals against the fitted values. These two plots should suggest that the regression is not well specified. Explain why, and estimate another hedonic regression that performs better. **Hint:** Perhaps it would make sense to transform one or more of the variables.

2.25 The file **consumption-data.txt** contains data on personal disposable income and consumption expenditures in the United States, seasonally adjusted in 2009 dollars, from the first quarter of 1947 until the last quarter of 2014. Regress the logarithm of consumption (c_t) on a constant, the logarithm of disposable income (y_t), and the logarithm of consumption lagged one quarter (c_{t-1}) for the period 1948:1 to 2014:4. This regression can be written as

$$c_t = \beta_1 + \beta_2 y_t + \beta_3 c_{t-1} + u_t. \tag{2.53}$$

Plot a graph of the OLS residuals for regression (2.53) against time. Does the appearance of the residuals suggest that this model of the consumption function is well specified?

2.26 Simulate the consumption function model (2.53) for the same sample period, using the actual data on disposable income. For the parameters, use the OLS estimates obtained in exercise 1.25. For the disturbances, use drawings from the $\mathrm{N}(0, s^2)$ distribution, where s is the standard error of the regression reported by the regression package. Use the actual value of c_t in 1947:4 for the initial value of c_{t-1}, but use simulated values after that.

Next, run regression (2.53) using the simulated consumption data and the actual disposable income data. How do the parameter estimates differ from the ones obtained using the real data?

Plot the residuals from the regression with the simulated data. Explain why the plot looks substantially different from the one obtained using the real data in exercise 1.25.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 3

The Geometry of Linear Regression

3.1 Introduction

In Chapter 2, we introduced regression models, both linear and nonlinear, and discussed how to estimate linear regression models by using estimating equations. We saw that all n observations of a linear regression model with k regressors can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{3.01}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are n–vectors, $\boldsymbol{X}$ is an $n \times k$ matrix, one column of which may be a constant term, and $\boldsymbol{\beta}$ is a k–vector. We also saw that the estimates of the vector $\boldsymbol{\beta}$, which are usually called the ordinary least-squares or OLS estimates, are

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{3.02}$$

In this chapter, we will be concerned with the **numerical properties** of these OLS estimates. We refer to certain properties of estimates as "numerical" if they have nothing to do with how the data were actually generated. Such properties hold for every set of data by virtue of the way in which $\hat{\boldsymbol{\beta}}$ is computed, and the fact that they hold can always be verified by direct calculation. In contrast, the **statistical properties** of OLS estimates, which will be discussed in Chapter 4, necessarily depend on unverifiable assumptions about how the data were generated, and they can never be verified for any actual data set.

In order to understand the numerical properties of OLS estimates, it is useful to look at them from the perspective of Euclidean geometry. This geometrical interpretation is remarkably simple. Essentially, it involves using Pythagoras' Theorem and a small amount of high-school trigonometry in the context of finite-dimensional vector spaces. Although this approach is simple, it is very powerful. Once one has a thorough grasp of the geometry involved in ordinary least squares, one can often save oneself many tedious lines of algebra by a simple geometrical argument. We will encounter many examples of this throughout the book.

In the next section, we review some relatively elementary material on the geometry of vector spaces and Pythagoras' Theorem. In Section 3.3, we then

discuss the most important numerical properties of OLS estimation from a geometrical perspective. In Section 3.4, we introduce an extremely useful result called the FWL Theorem, and in Section 3.5 we present a number of applications of this theorem. Finally, in Section 3.6, we discuss how and to what extent individual observations influence parameter estimates.

3.2 The Geometry of Vector Spaces

In Section 2.4, an n–vector was defined as a column vector with n elements, that is, an $n \times 1$ matrix. The elements of such a vector are real numbers. The usual notation for the **real line** is $\mathbb{R}$, and it is therefore natural to denote the set of n–vectors as $\mathbb{R}^n$. However, in order to use the insights of Euclidean geometry to enhance our understanding of the algebra of vectors and matrices, it is desirable to introduce the notion of a **Euclidean space** in n dimensions, which we will denote as E^n. The difference between $\mathbb{R}^n$ and E^n is not that they consist of different sorts of vectors, but rather that a wider set of operations is defined on E^n. A shorthand way of saying that a vector $\boldsymbol{x}$ belongs to an n–dimensional Euclidean space is to write $\boldsymbol{x} \in E^n$.

Addition and subtraction of vectors in E^n is no different from the addition and subtraction of $n \times 1$ matrices discussed in Section 2.4. The same thing is true of multiplication by a scalar in E^n. The final operation essential to E^n is that of the **scalar product**, **inner product**, or **dot product**. For any two vectors $\boldsymbol{x}, \boldsymbol{y} \in E^n$, their scalar product is

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \equiv \boldsymbol{x}^\top \boldsymbol{y}.$$

The notation on the left is generally used in the context of the geometry of vectors, while the notation on the right is generally used in the context of matrix algebra. Note that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$, since $\boldsymbol{x}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{x}$. Thus the scalar product is **commutative**.

The scalar product is what allows us to make a close connection between n–vectors considered as matrices and considered as geometrical objects. It allows us to define the **length** of any vector in E^n. The length, or **norm**, of a vector $\boldsymbol{x}$ is simply

$$\|\boldsymbol{x}\| \equiv (\boldsymbol{x}^\top \boldsymbol{x})^{1/2}.$$

This is just the square root of the inner product of $\boldsymbol{x}$ with itself. In scalar terms, it is

$$\|\boldsymbol{x}\| \equiv \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}. \tag{3.03}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

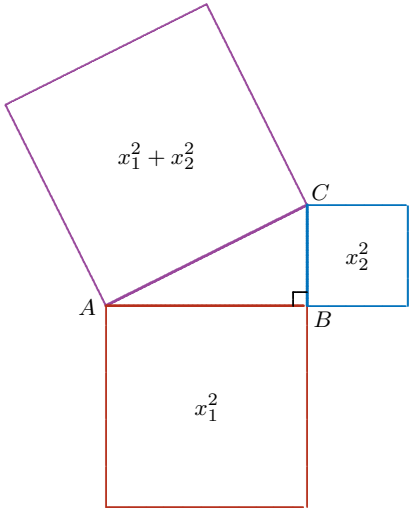


Figure 3.1 Pythagoras' Theorem

Pythagoras' Theorem

The definition (3.03) is inspired by the celebrated theorem of Pythagoras, which says that the square on the longest side of a right-angled triangle is equal to the sum of the squares on the other two sides. This longest side is called the **hypotenuse**. Pythagoras' Theorem is illustrated in Figure 3.1. The figure shows a right-angled triangle, ABC, with hypotenuse AC and two other sides, AB and BC, of lengths x_1 and x_2, respectively. The squares on each of the three sides of the triangle are drawn, and the area of the square on the hypotenuse is shown as $x_1^2 + x_2^2$, in accordance with the theorem.

A beautiful proof of Pythagoras' Theorem, not often found in geometry texts, is shown in Figure 3.2. Two squares of equal area are drawn. Each square contains four copies of the same right-angled triangle. The square on the left also contains the squares on the two shorter sides of the triangle, while the square on the right contains the square on the hypotenuse. The theorem follows at once.

Any vector $\boldsymbol{x} \in E^2$ has two components, usually denoted as x_1 and x_2. These two components can be interpreted as the **Cartesian coordinates** of the vector in the plane. The situation is illustrated in Figure 3.3. With O as the origin of the coordinates, a right-angled triangle is formed by the lines OA, AB, and OB. The length of the horizontal side of the triangle, OA, is the horizontal coordinate x_1. The length of the vertical side, AB, is the vertical coordinate x_2. Thus the point B has Cartesian coordinates (x_1, x_2). The vector $\boldsymbol{x}$ itself is usually represented as the hypotenuse of the triangle, OB, that

Copyright © 2021, Russell Davidson and James G. MacKinnon

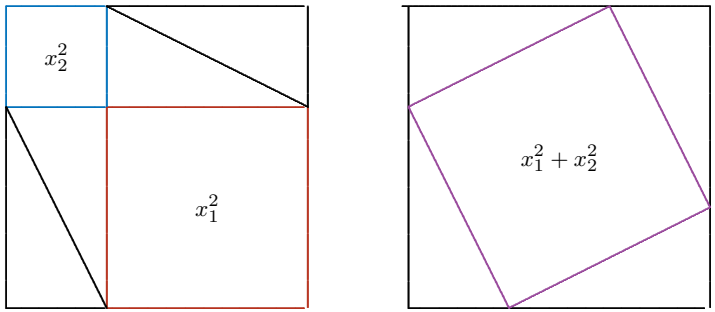


Figure 3.2 Proof of Pythagoras' Theorem

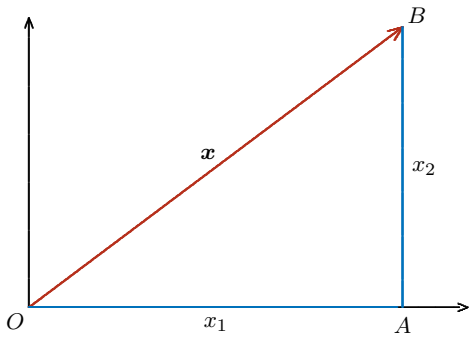


Figure 3.3 A vector $\boldsymbol{x}$ in E^2

is, the directed line (depicted as an arrow) joining the origin to the point B, with coordinates (x_1, x_2). Pythagoras' Theorem tells us that the length of the vector $\boldsymbol{x}$, which is the hypotenuse of the triangle, is $(x_1^2 + x_2^2)^{1/2}$. By equation (3.03), this is what $\|\boldsymbol{x}\|$ is equal to when $n = 2$.

Vector Geometry in Two Dimensions

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two vectors in E^2, with components (x_1, x_2) and (y_1, y_2), respectively. Then, by the rules of matrix addition, the components of $\boldsymbol{x} + \boldsymbol{y}$ are $(x_1 + y_1, x_2 + y_2)$. Figure 3.4 shows how the addition of $\boldsymbol{x}$ and $\boldsymbol{y}$ can be performed geometrically in two different ways. The vector $\boldsymbol{x}$ is drawn as the directed line segment, or arrow, from the origin O to the point A with coordinates (x_1, x_2). The vector $\boldsymbol{y}$ can be drawn similarly and represented by the arrow OB. However, we could also draw $\boldsymbol{y}$ starting, not at O, but at the point reached after drawing $\boldsymbol{x}$, namely A. The arrow AC has the same length and direction as OB, and we will see in general that arrows with the

Copyright © 2021, Russell Davidson and James G. MacKinnon

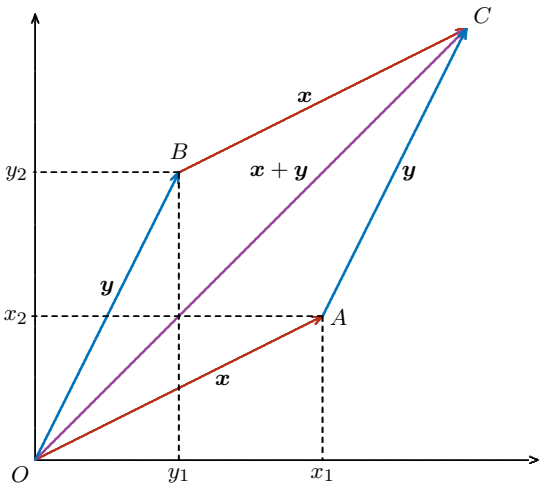


Figure 3.4 Addition of vectors

same length and direction can be taken to represent the same vector. It is clear by construction that the coordinates of C are $(x_1 + y_1, x_2 + y_2)$, that is, the coordinates of $\boldsymbol{x} + \boldsymbol{y}$. Thus the sum $\boldsymbol{x} + \boldsymbol{y}$ is represented geometrically by the arrow OC.

The classical way of adding vectors geometrically is to form a parallelogram using the line segments OA and OB that represent the two vectors as adjacent sides of the parallelogram. The sum of the two vectors is then the diagonal through O of the resulting parallelogram. It is easy to see that this classical method also gives the result that the sum of the two vectors is represented by the arrow OC, since the figure $OACB$ is just the parallelogram required by the construction, and OC is its diagonal through O. The parallelogram construction also shows clearly that vector addition is commutative, since $\boldsymbol{y} + \boldsymbol{x}$ is represented by OB, for $\boldsymbol{y}$, followed by BC, for $\boldsymbol{x}$. The end result is once more OC.

Multiplying a vector by a scalar is also very easy to represent geometrically. If a vector $\boldsymbol{x}$ with components (x_1, x_2) is multiplied by a scalar α, then $\alpha\boldsymbol{x}$ has components $(\alpha x_1, \alpha x_2)$. This is depicted in Figure 3.5, where $\alpha = 2$. The line segments OA and OB represent $\boldsymbol{x}$ and $\alpha\boldsymbol{x}$, respectively. It is clear that even if we move $\alpha\boldsymbol{x}$ so that it starts somewhere other than O, as with CD in the figure, the vectors $\boldsymbol{x}$ and $\alpha\boldsymbol{x}$ are always **parallel**. If α were negative, then $\alpha\boldsymbol{x}$ would simply point in the opposite direction. Thus, for $\alpha = -2$, $\alpha\boldsymbol{x}$ would be represented by DC, rather than CD.

Copyright © 2021, Russell Davidson and James G. MacKinnon

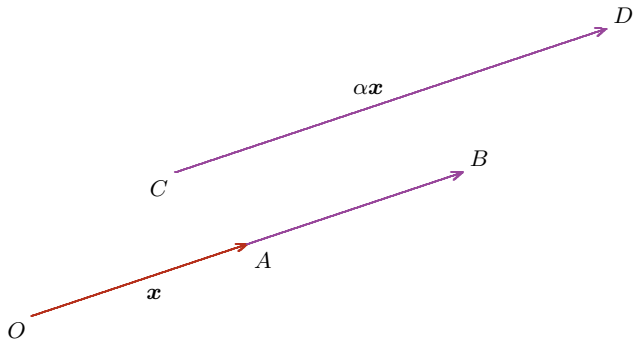


Figure 3.5 Multiplication by a scalar

Another property of multiplication by a scalar is clear from Figure 3.5. By direct calculation,

$$\|\alpha\boldsymbol{x}\| = \langle\alpha\boldsymbol{x}, \alpha\boldsymbol{x}\rangle^{1/2} = |\alpha|(\boldsymbol{x}^\top\boldsymbol{x})^{1/2} = |\alpha|\,\|\boldsymbol{x}\|. \tag{3.04}$$

Since $\alpha = 2$, OB and CD in the figure are twice as long as OA.

The Geometry of Scalar Products

The scalar product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, whether in E^2 or E^n, can be expressed geometrically in terms of the lengths of the two vectors and the **angle** between them, and this result will turn out to be very useful. In the case of E^2, it is natural to think of the angle between two vectors as the angle between the two line segments that represent them. As we will now show, it is also quite easy to define the angle between two vectors in E^n.

If the angle between two vectors is 0, they must be **parallel**. The vector $\boldsymbol{y}$ is parallel to the vector $\boldsymbol{x}$ if $\boldsymbol{y} = \alpha\boldsymbol{x}$ for some suitable α. In that event,

$$\langle\boldsymbol{x}, \boldsymbol{y}\rangle = \langle\boldsymbol{x}, \alpha\boldsymbol{x}\rangle = \alpha\boldsymbol{x}^\top\boldsymbol{x} = \alpha\|\boldsymbol{x}\|^2.$$

From (3.04), we know that $\|\boldsymbol{y}\| = |\alpha|\,\|\boldsymbol{x}\|$, and so, if $\alpha > 0$, it follows that

$$\langle\boldsymbol{x}, \boldsymbol{y}\rangle = \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|. \tag{3.05}$$

Of course, this result is true only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel and point in the same direction (rather than in opposite directions).

For simplicity, consider initially two vectors, $\boldsymbol{w}$ and $\boldsymbol{z}$, both of length 1, and let θ denote the angle between them. This is illustrated in Figure 3.6. Suppose that the first vector, $\boldsymbol{w}$, has coordinates $(1, 0)$. It is therefore represented by a horizontal line of length 1 in the figure. Suppose that the second vector, $\boldsymbol{z}$,

Copyright © 2021, Russell Davidson and James G. MacKinnon

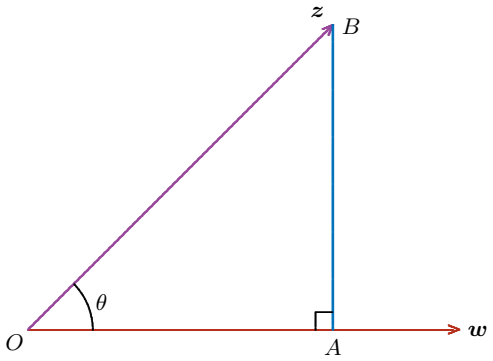


Figure 3.6 The angle between two vectors

is also of length 1, that is, $\|\boldsymbol{z}\| = 1$. Then, by elementary trigonometry, the coordinates of $\boldsymbol{z}$ must be $(\cos\theta, \sin\theta)$. To show this, note first that, if so,

$$\|\boldsymbol{z}\|^2 = \cos^2\theta + \sin^2\theta = 1, \tag{3.06}$$

as required. Next, consider the right-angled triangle OAB, in which the hypotenuse OB represents $\boldsymbol{z}$ and is of length 1, by (3.06). The length of the side AB opposite O is $\sin\theta$, the vertical coordinate of $\boldsymbol{z}$. Then the sine of the angle BOA is given, by the usual trigonometric rule, by the ratio of the length of the opposite side AB to that of the hypotenuse OB. This ratio is $\sin\theta/1 = \sin\theta$, and so the angle BOA is indeed equal to θ.

Now let us compute the scalar product of $\boldsymbol{w}$ and $\boldsymbol{z}$. It is

$$\langle \boldsymbol{w}, \boldsymbol{z} \rangle = \boldsymbol{w}^\top \boldsymbol{z} = w_1 z_1 + w_2 z_2 = z_1 = \cos\theta,$$

because $w_1 = 1$ and $w_2 = 0$. This result holds for vectors $\boldsymbol{w}$ and $\boldsymbol{z}$ of length 1. More generally, let $\boldsymbol{x} = \alpha\boldsymbol{w}$ and $\boldsymbol{y} = \gamma\boldsymbol{z}$, for positive scalars α and γ. Then $\|\boldsymbol{x}\| = \alpha$ and $\|\boldsymbol{y}\| = \gamma$. Thus we have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \alpha\gamma \boldsymbol{w}^\top \boldsymbol{z} = \alpha\gamma \langle \boldsymbol{w}, \boldsymbol{z} \rangle.$$

Because $\boldsymbol{x}$ is parallel to $\boldsymbol{w}$, and $\boldsymbol{y}$ is parallel to $\boldsymbol{z}$, the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$ is the same as that between $\boldsymbol{w}$ and $\boldsymbol{z}$, namely θ. Therefore,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\| \, \|\boldsymbol{y}\| \cos\theta. \tag{3.07}$$

This is the general expression, in geometrical terms, for the scalar product of two vectors. It is true in E^n just as it is in E^2, although we have not proved this. In fact, we have not quite proved (3.07) even for the two-dimensional case, because we made the simplifying assumption that the direction of $\boldsymbol{x}$

Copyright © 2021, Russell Davidson and James G. MacKinnon

and $\boldsymbol{w}$ is horizontal. In Exercise 3.1, we ask the reader to provide a more complete proof.

The cosine of the angle between two vectors provides a natural way to measure how close two vectors are in terms of their directions. Recall that $\cos\theta$ varies between -1 and 1; if we measure angles in radians, $\cos 0 = 1$, $\cos \pi/2 = 0$, and $\cos\pi = -1$. Thus $\cos\theta$ is 1 for vectors that are parallel, 0 for vectors that are at right angles to each other, and -1 for vectors that point in directly opposite directions. If the angle θ between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is a right angle, its cosine is 0, and so, from (3.07), the scalar product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is 0. Conversely, if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$, then $\cos\theta = 0$ unless $\boldsymbol{x}$ or $\boldsymbol{y}$ is a zero vector. If $\cos\theta = 0$, it follows that $\theta = \pi/2$. Thus, if two nonzero vectors have a zero scalar product, they are at right angles. Such vectors are often said to be **orthogonal**, or, less commonly, **perpendicular**. This definition implies that the zero vector is orthogonal to everything.

Since the cosine function can take on values only between -1 and 1, a consequence of (3.07) is that

$$|\boldsymbol{x}^\top \boldsymbol{y}| \leq \|\boldsymbol{x}\| \, \|\boldsymbol{y}\|. \tag{3.08}$$

This result, which is called the **Cauchy-Schwartz inequality**, says that the absolute value of the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$ can never be greater than the length of the vector $\boldsymbol{x}$ times the length of the vector $\boldsymbol{y}$. Only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel does the inequality in (3.08) become the equality (3.05). Readers are asked to prove this result in Exercise 3.2.

Subspaces of Euclidean Space

For arbitrary positive integers n, the elements of an n–vector can be thought of as the coordinates of a point in E^n. In particular, in the regression model (3.01), the regressand $\boldsymbol{y}$ and each column of the matrix of regressors $\boldsymbol{X}$ can be thought of as vectors in E^n. This makes it possible to represent a relationship like (3.01) geometrically.

It is obviously impossible to represent all n dimensions of E^n physically when $n > 3$. For the pages of a book, even three dimensions can be too many, although a proper use of perspective drawings can allow three dimensions to be shown. Fortunately, we can represent (3.01) without needing to draw in n dimensions. The key to this is that there are only three vectors in (3.01): $\boldsymbol{y}$, $\boldsymbol{X\beta}$, and $\boldsymbol{u}$. Since only two vectors, $\boldsymbol{X\beta}$ and $\boldsymbol{u}$, appear on the right-hand side of (3.01), only two dimensions are needed to represent it. Because $\boldsymbol{y}$ is equal to $\boldsymbol{X\beta} + \boldsymbol{u}$, these two dimensions suffice for $\boldsymbol{y}$ as well.

To see how this works, we need the concept of a **subspace** of a Euclidean space E^n. Normally, such a subspace has a dimension lower than n. The easiest way to define a subspace of E^n is in terms of a set of **basis vectors**. A subspace that is of particular interest to us is the one for which the columns of $\boldsymbol{X}$ provide the basis vectors. We may denote the k columns of $\boldsymbol{X}$ as $\boldsymbol{x}_1$, $\boldsymbol{x}_2, \ldots \boldsymbol{x}_k$. Then the subspace associated with these k basis vectors is denoted

Copyright © 2021, Russell Davidson and James G. MacKinnon

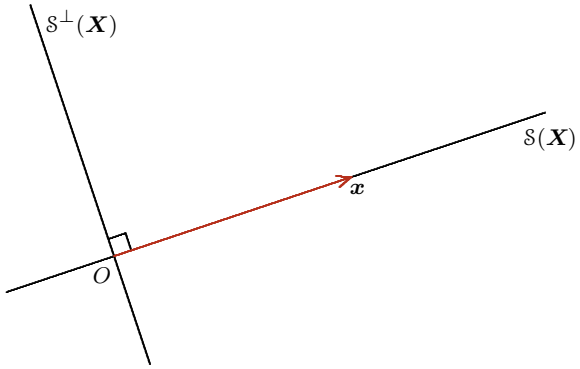


Figure 3.7 The spaces $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}^\perp(\boldsymbol{X})$

by $\mathcal{S}(\boldsymbol{X})$ or $\mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. The basis vectors are said to **span** this subspace, which in general is a k–dimensional subspace.

The subspace $\mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ consists of every vector that can be formed as a **linear combination** of the $\boldsymbol{x}_i$, $i = 1, \ldots, k$. Formally, it is defined as

$$\mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \equiv \Big\{ \boldsymbol{z} \in E^n \ : \ \boldsymbol{z} = \sum_{i=1}^{k} b_i \boldsymbol{x}_i, \quad b_i \in \mathbb{R} \Big\}. \tag{3.09}$$

The subspace defined in (3.09) is called the subspace spanned by the $\boldsymbol{x}_i$, $i = 1, \ldots, k$, or the **column space** of $\boldsymbol{X}$; less formally, it may simply be referred to as the **span** of $\boldsymbol{X}$, or the span of the $\boldsymbol{x}_i$.

The **orthogonal complement** of $\mathcal{S}(\boldsymbol{X})$ in E^n, which is denoted $\mathcal{S}^\perp(\boldsymbol{X})$, is the set of all vectors $\boldsymbol{w}$ in E^n that are orthogonal to everything in $\mathcal{S}(\boldsymbol{X})$. This means that, for every $\boldsymbol{z}$ in $\mathcal{S}(\boldsymbol{X})$, $\langle \boldsymbol{w}, \boldsymbol{z} \rangle = \boldsymbol{w}^\top \boldsymbol{z} = 0$. Formally,

$$\mathcal{S}^\perp(\boldsymbol{X}) \equiv \left\{ \boldsymbol{w} \in E^n \ : \ \boldsymbol{w}^\top \boldsymbol{z} = 0 \text{ for all } \boldsymbol{z} \in \mathcal{S}(\boldsymbol{X}) \right\}.$$

If the dimension of $\mathcal{S}(\boldsymbol{X})$ is k, then the dimension of $\mathcal{S}^\perp(\boldsymbol{X})$ is $n - k$.

Figure 3.7 illustrates the concepts of a subspace and its orthogonal complement for the simplest case, in which $n = 2$ and $k = 1$. The matrix $\boldsymbol{X}$ has only one column in this case, and it is therefore represented in the figure by a single vector, denoted $\boldsymbol{x}$. As a consequence, $\mathcal{S}(\boldsymbol{X})$ is 1–dimensional, and, since $n = 2$, $\mathcal{S}^\perp(\boldsymbol{X})$ is also 1–dimensional. Notice that $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}^\perp(\boldsymbol{X})$ would be the same if $\boldsymbol{x}$ were *any* vector, except for the origin, parallel to the straight line that represents $\mathcal{S}(\boldsymbol{X})$.

Now let us return to E^n. Suppose, to begin with, that $k = 2$. We have two vectors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which span a subspace of, at most, two dimensions. It is always possible to represent vectors in a 2–dimensional space on a piece of

Copyright © 2021, Russell Davidson and James G. MacKinnon

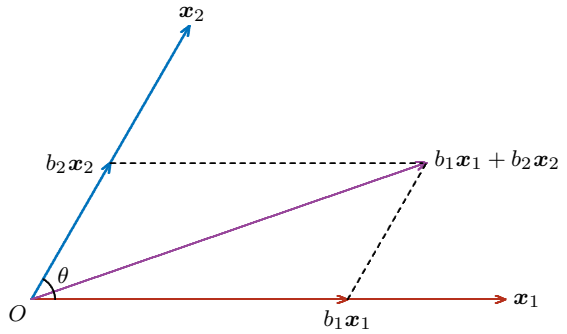


Figure 3.8 A 2-dimensional subspace

paper, whether that space is E^2 itself or, as in this case, the 2–dimensional subspace of E^n spanned by the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. To represent the first vector, $\boldsymbol{x}_1$, we choose an origin and a direction, both of which are entirely arbitrary, and draw an arrow of length $\|\boldsymbol{x}_1\|$ in that direction. Suppose that the origin is the point O in Figure 3.8, and that the direction is the horizontal direction in the plane of the page. Then an arrow to represent $\boldsymbol{x}_1$ can be drawn as shown in the figure. For $\boldsymbol{x}_2$, we compute its length, $\|\boldsymbol{x}_2\|$, and the angle, θ, that it makes with $\boldsymbol{x}_1$. Suppose for now that $\theta \neq 0$. Then we choose as our second dimension the vertical direction in the plane of the page, with the result that we can draw an arrow for $\boldsymbol{x}_2$, as shown.

Any vector in $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be drawn in the plane of Figure 3.8. Consider, for instance, the linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ given by the expression $\boldsymbol{z} \equiv b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2$. We could draw the vector $\boldsymbol{z}$ by computing its length and the angle that it makes with $\boldsymbol{x}_1$. Alternatively, we could apply the rules for adding vectors geometrically that were illustrated in Figure 3.4 to the vectors $b_1\boldsymbol{x}_1$ and $b_2\boldsymbol{x}_2$. This is illustrated in the figure for the case in which $b_1 = 2/3$ and $b_2 = 1/2$.

In precisely the same way, we can represent any three vectors by arrows in 3–dimensional space, but we leave this task to the reader. It will be easier to appreciate the renderings of vectors in three dimensions in perspective that appear later on if one has already tried to draw 3–dimensional pictures, or even to model relationships in three dimensions with the help of a computer.

We can finally represent the regression model (3.01) geometrically. This is done in Figure 3.9. The horizontal direction is chosen for the vector $\boldsymbol{X\beta}$, and then the other two vectors $\boldsymbol{y}$ and $\boldsymbol{u}$ are shown in the plane of the page. It is clear that, by construction, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$. Notice that $\boldsymbol{u}$, the vector of disturbances, is not orthogonal to $\boldsymbol{X\beta}$. The figure contains no reference to any system of axes, because there would be n of them, and we would not be able to avoid needing n dimensions to treat them all.

Copyright © 2021, Russell Davidson and James G. MacKinnon

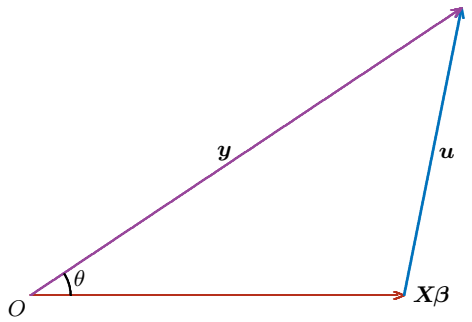


Figure 3.9 The geometry of the linear regression model

Linear Independence

In order to define the OLS estimator by the formula (2.45), it is necessary to assume that the $k \times k$ square matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible, or nonsingular. Equivalently, as we saw in Section 2.4, we may say that $\boldsymbol{X}^\top\boldsymbol{X}$ has full rank. This condition is equivalent to the condition that the columns of $\boldsymbol{X}$ should be **linearly independent**. This is a very important concept for econometrics. Note that the meaning of linear independence is quite different from the meaning of statistical independence, which we discussed in Section 2.2. It is important not to confuse these two concepts.

The vectors $\boldsymbol{x}_1$ through $\boldsymbol{x}_k$ are said to be **linearly dependent** if we can write one of them as a linear combination of the others. In other words, there is a vector $\boldsymbol{x}_j$, $1 \leq j \leq k$, and coefficients c_i such that

$$\boldsymbol{x}_j = \sum_{i \neq j} c_i \boldsymbol{x}_i. \tag{3.10}$$

Another, equivalent, definition is that there exist coefficients b_i, at least one of which is nonzero, such that

$$\sum_{i=1}^{k} b_i \boldsymbol{x}_i = \boldsymbol{0}. \tag{3.11}$$

Recall that $\boldsymbol{0}$ denotes the **zero vector**, every component of which is 0. It is clear from the definition (3.11) that, if any of the $\boldsymbol{x}_i$ is itself equal to the zero vector, then the $\boldsymbol{x}_i$ are linearly dependent. If $\boldsymbol{x}_j = \boldsymbol{0}$, for example, then equation (3.11) is satisfied if we make b_j nonzero and set $b_i = 0$ for all $i \neq j$.

If the vectors $\boldsymbol{x}_i$, $i = 1, \ldots, k$, are the columns of an $n \times k$ matrix $\boldsymbol{X}$, then another way of writing (3.11) is

$$\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}, \tag{3.12}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $\boldsymbol{b}$ is a k–vector with typical element b_i. In order to see that (3.11) and (3.12) are equivalent, it is enough to check that the typical elements of the two left-hand sides are the same; see Exercise 3.5. The set of vectors $\boldsymbol{x}_i$, $i = 1, \ldots, k$, is linearly independent if it is not linearly dependent, that is, if there are no coefficients c_i such that (3.10) is true, or (equivalently) no coefficients b_i such that (3.11) is true, or (equivalently, once more) no vector $\boldsymbol{b}$ such that (3.12) is true.

It is easy to show that, if the columns of $\boldsymbol{X}$ are linearly dependent, the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is not invertible. As we have seen, if they are linearly dependent, there must exist a nonzero vector $\boldsymbol{b}$ such that $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}$. Premultiplying this equation, which is (3.12), by $\boldsymbol{X}^\top$ yields

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}. \tag{3.13}$$

Now suppose that the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible. If so, there exists a matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ such that $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X}) = \boldsymbol{I}$. Thus equation (3.13) implies that

$$\boldsymbol{b} = \boldsymbol{I}\boldsymbol{b} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}.$$

But this is a contradiction, since we have assumed that $\boldsymbol{b} \neq \boldsymbol{0}$. Therefore, we conclude that the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ cannot exist when the columns of $\boldsymbol{X}$ are linearly dependent. Thus a necessary condition for the existence of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is that the columns of $\boldsymbol{X}$ should be linearly independent. With a little more work, it can be shown that this condition is also sufficient, and so, if the regressors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are linearly independent, $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible.

If the k columns of $\boldsymbol{X}$ are not linearly independent, then they span a subspace of dimension less than k, say k', where k' is the largest number of columns of $\boldsymbol{X}$ that are linearly independent of each other. The number k' is called the **rank** of $\boldsymbol{X}$. Look again at Figure 3.8, and imagine that the angle θ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ tends to zero. If $\theta = 0$, then $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are parallel, and we can write $\boldsymbol{x}_1 = \alpha\boldsymbol{x}_2$, for some scalar α. But this means that $\boldsymbol{x}_1 - \alpha\boldsymbol{x}_2 = \boldsymbol{0}$, and so a relation of the form (3.11) holds between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which are therefore linearly dependent. In the figure, if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are parallel, then only one dimension is used, and there is no need for the second dimension in the plane of the page. Thus, in this case, $k = 2$ and $k' = 1$.

When the dimension of $\mathcal{S}(\boldsymbol{X})$ is $k' < k$, $\mathcal{S}(\boldsymbol{X})$ must be identical to $\mathcal{S}(\boldsymbol{X}')$, where $\boldsymbol{X}'$ is an $n \times k'$ matrix consisting of any k' linearly independent columns of $\boldsymbol{X}$. For example, consider the following $\boldsymbol{X}$ matrix, which is 5×3:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{3.14}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The columns of this matrix are not linearly independent, since

$$\boldsymbol{x}_1 = \frac{1}{4}\boldsymbol{x}_2 + \boldsymbol{x}_3.$$

However, any two of the columns are linearly independent, and so

$$\mathcal{S}(\boldsymbol{X}) = \mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_3) = \mathcal{S}(\boldsymbol{x}_2, \boldsymbol{x}_3);$$

see Exercise 3.8. For the remainder of this chapter, unless the contrary is explicitly assumed, we will assume that the columns of any regressor matrix $\boldsymbol{X}$ are linearly independent.

3.3 The Geometry of OLS Estimation

We studied the geometry of vector spaces in the preceding section because the numerical properties of OLS estimates are easily understood in terms of that geometry. The geometrical interpretation of OLS estimation of linear regression models is simple and intuitive. In many cases, it entirely does away with the need for algebraic proofs.

As we saw in Section 3.2, any point in a subspace $\mathcal{S}(\boldsymbol{X})$, where $\boldsymbol{X}$ is an $n \times k$ matrix, can be represented as a linear combination of the columns of $\boldsymbol{X}$. We can partition $\boldsymbol{X}$ in terms of its columns explicitly, as follows:

$$\boldsymbol{X} = [\, \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_k \,].$$

In order to compute the matrix product $\boldsymbol{X\beta}$ in terms of this partitioning, we need to partition the vector $\boldsymbol{\beta}$ by its rows. Since $\boldsymbol{\beta}$ has only one column, the elements of the partitioned vector are just the individual elements of $\boldsymbol{\beta}$. Thus we find that

$$\boldsymbol{X\beta} = [\, \boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_k \,] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \ldots + \boldsymbol{x}_k\beta_k = \sum_{i=1}^{k} \beta_i \boldsymbol{x}_i,$$

which is just a linear combination of the columns of $\boldsymbol{X}$. In fact, it is clear from the definition (3.09) that any linear combination of the columns of $\boldsymbol{X}$, and thus any element of the subspace $\mathcal{S}(\boldsymbol{X}) = \mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$, can be written as $\boldsymbol{X\beta}$ for some $\boldsymbol{\beta}$. The specific linear combination (3.09) is constructed by using $\boldsymbol{\beta} = [b_1 \vdots \ldots \vdots b_k]$. Thus every n–vector $\boldsymbol{X\beta}$ belongs to $\mathcal{S}(\boldsymbol{X})$, which is, in general, a k–dimensional subspace of E^n. In particular, the vector $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ constructed using the OLS estimator $\hat{\boldsymbol{\beta}}$ belongs to this subspace.

The estimator $\hat{\boldsymbol{\beta}}$ satisfies equations (2.47), and so we have

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}. \tag{3.15}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

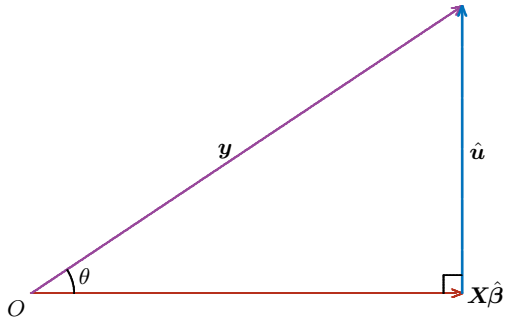


Figure 3.10 Residuals and fitted values

These equations have a simple geometrical interpretation. Note first that each element of the left-hand side of (3.15) is a scalar product. By the rule for selecting a single row of a matrix product (see Section 2.4), the i^{th} element is

$$\boldsymbol{x}_i^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \langle \boldsymbol{x}_i, \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \rangle, \tag{3.16}$$

since $\boldsymbol{x}_i$, the i^{th} column of $\boldsymbol{X}$, is the transpose of the i^{th} row of $\boldsymbol{X}^\top$. By (3.15), the scalar product in (3.16) is zero, and so the vector $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is orthogonal to all of the regressors, that is, all of the vectors $\boldsymbol{x}_i$ that represent the explanatory variables in the regression. For this reason, equations like (3.15) are often referred to as **orthogonality conditions**.

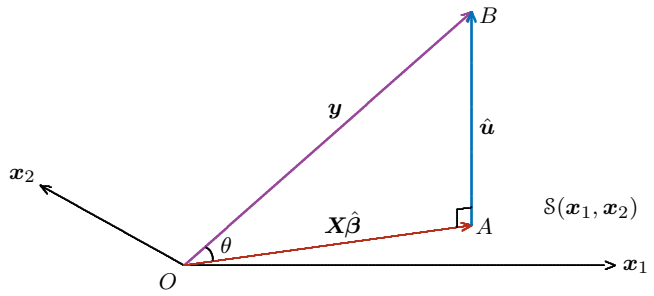
Recall from Section 2.5 that the vector $\boldsymbol{y} - \boldsymbol{X\beta}$, treated as a function of $\boldsymbol{\beta}$, is called the vector of residuals. This vector may be written as $\boldsymbol{u}(\boldsymbol{\beta})$. We are interested in $\boldsymbol{u}(\hat{\boldsymbol{\beta}})$, the vector of residuals evaluated at $\hat{\boldsymbol{\beta}}$, which is often called the vector of **least-squares residuals** and is usually written simply as $\hat{\boldsymbol{u}}$. We have just seen, in (3.16), that $\hat{\boldsymbol{u}}$ is orthogonal to all the regressors. This implies that $\hat{\boldsymbol{u}}$ is in fact orthogonal to *every* vector in $\mathcal{S}(\boldsymbol{X})$, the span of the regressors. To see this, remember that any element of $\mathcal{S}(\boldsymbol{X})$ can be written as $\boldsymbol{X\beta}$ for some $\boldsymbol{\beta}$, with the result that, by (3.15),

$$\langle \boldsymbol{X\beta}, \hat{\boldsymbol{u}} \rangle = (\boldsymbol{X\beta})^\top \hat{\boldsymbol{u}} = \boldsymbol{\beta}^\top \boldsymbol{X}^\top \hat{\boldsymbol{u}} = \boldsymbol{0}.$$

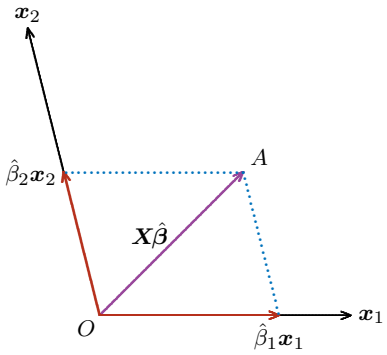
The vector $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is referred to as the vector of **fitted values**. Clearly, it lies in $\mathcal{S}(\boldsymbol{X})$, and, consequently, it must be orthogonal to $\hat{\boldsymbol{u}}$. Figure 3.10 is similar to Figure 3.9, but it shows the vector of least-squares residuals $\hat{\boldsymbol{u}}$ and the vector of fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{u}$ and $\boldsymbol{X\beta}$. The key feature of this figure, which is a consequence of the orthogonality conditions (3.15), is that the vector $\hat{\boldsymbol{u}}$ makes a right angle with the vector $\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Some things about the orthogonality conditions (3.15) are clearer if we add a third dimension to the picture. Accordingly, in panel (a) of Figure 3.11,

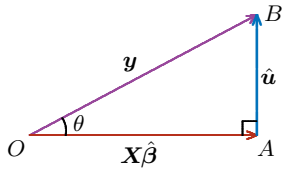
Copyright © 2021, Russell Davidson and James G. MacKinnon



(a) $\boldsymbol{y}$ projected on two regressors



(b) The span $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ of the regressors



(c) The vertical plane through $\boldsymbol{y}$

Figure 3.11 Linear regression in three dimensions

we consider the case of two regressors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which together span the horizontal plane labelled $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, seen in perspective from slightly above the plane. Although the perspective rendering of the figure does not make it clear, both the lengths of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and the angle between them are totally arbitrary, since they do not affect $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ at all. The vector $\boldsymbol{y}$ is intended to be viewed as rising up out of the plane spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

In the 3–dimensional setup, it is clear that, if $\hat{\boldsymbol{u}}$ is to be orthogonal to the horizontal plane, it must itself be vertical. Thus it is obtained by "dropping a perpendicular" from $\boldsymbol{y}$ to the horizontal plane. The least-squares interpretation of the estimator $\hat{\boldsymbol{\beta}}$ can now be seen to be a consequence of simple geometry. The shortest distance from $\boldsymbol{y}$ to the horizontal plane is obtained by descending vertically on to it, and the point in the horizontal plane vertically below $\boldsymbol{y}$, labeled A in the figure, is the closest point in the plane to $\boldsymbol{y}$. Thus $\|\hat{\boldsymbol{u}}\|$ minimizes $\|\boldsymbol{u}(\boldsymbol{\beta})\|$, the norm of $\boldsymbol{u}(\boldsymbol{\beta})$, with respect to $\boldsymbol{\beta}$. The squared norm,

Copyright © 2021, Russell Davidson and James G. MacKinnon

$\|\boldsymbol{u}(\boldsymbol{\beta})\|^2$, is just the sum of squared residuals, $\mathrm{SSR}(\boldsymbol{\beta})$; see (2.48). Since minimizing the norm of $\boldsymbol{u}(\boldsymbol{\beta})$ is the same thing as minimizing the squared norm, it follows that $\hat{\boldsymbol{\beta}}$ is the OLS estimator.

Panel (b) of the figure shows the horizontal plane $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ as a straightforward 2–dimensional picture, seen from directly above. The point A is the point directly underneath $\boldsymbol{y}$, and so, since $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}$ by definition, the vector represented by the line segment OA is the vector of fitted values, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Geometrically, it is much simpler to represent $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ than to represent just the vector $\hat{\boldsymbol{\beta}}$, because the latter lies in $\mathbb{R}^k$, a different space from the space E^n that contains the variables and all linear combinations of them. However, it is easy to see that the information in panel (b) does indeed determine $\hat{\boldsymbol{\beta}}$. Plainly, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ can be decomposed in just one way as a linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, as shown. The numerical value of $\hat{\beta}_1$ can be computed as the ratio of the length of the vector $\hat{\beta}_1\boldsymbol{x}_1$ to that of $\boldsymbol{x}_1$, and similarly for $\hat{\beta}_2$.

In panel (c) of Figure 3.11, we show the right-angled triangle that corresponds to dropping a perpendicular from $\boldsymbol{y}$, labelled in the same way as in panel (a). This triangle lies in the vertical plane that contains the vector $\boldsymbol{y}$. We can see that $\boldsymbol{y}$ is the hypotenuse of the triangle, the other two sides being $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$. Thus this panel corresponds to what we saw already in Figure 3.10. Since we have a right-angled triangle, we can apply Pythagoras' Theorem. It gives

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\hat{\boldsymbol{u}}\|^2. \tag{3.17}$$

If we write out the squared norms as scalar products, this becomes

$$\boldsymbol{y}^\top\boldsymbol{y} = \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{3.18}$$

In words, the **total sum of squares**, or **TSS**, is equal to the **explained sum of squares**, or **ESS**, plus the **sum of squared residuals**, or **SSR**. This is a fundamental property of OLS estimates, and it will prove to be very useful in many contexts. Intuitively, it lets us break down the total variation (TSS) of the dependent variable into the explained variation (ESS) and the unexplained variation (SSR), unexplained because the residuals represent the aspects of $\boldsymbol{y}$ about which we remain in ignorance.

Orthogonal Projections

When we estimate a linear regression model, we implicitly map the regressand $\boldsymbol{y}$ into a vector of fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and a vector of residuals $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Geometrically, these mappings are examples of orthogonal projections. A **projection** is a mapping that takes each point of E^n into a point in a subspace of E^n, while leaving all points in that subspace unchanged. Because of this, the subspace is called the **invariant subspace** of the projection. An **orthogonal projection** maps any point into the point of the subspace that is closest to it. If a point is already in the invariant subspace, it is mapped into itself.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The concept of an orthogonal projection formalizes the notion of "dropping a perpendicular" that we used in the last subsection when discussing least squares. Algebraically, an orthogonal projection on to a given subspace can be performed by premultiplying the vector to be projected by a suitable **projection matrix**. In the case of OLS, the two projection matrices that yield the vector of fitted values and the vector of residuals, respectively, are

$$\begin{aligned} \boldsymbol{P_X} &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \quad \text{and} \\ \boldsymbol{M_X} &= \mathbf{I} - \boldsymbol{P_X} = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \end{aligned} \tag{3.19}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. To see this, recall (3.02), the formula for the OLS estimates of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

From this, we see that

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{P_X}\boldsymbol{y}. \tag{3.20}$$

Therefore, the first projection matrix in (3.19), $\boldsymbol{P_X}$, projects on to $\mathcal{S}(\boldsymbol{X})$. For any n–vector $\boldsymbol{y}$, $\boldsymbol{P_X}\boldsymbol{y}$ always lies in $\mathcal{S}(\boldsymbol{X})$, because

$$\boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{X}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\big).$$

Since this takes the form $\boldsymbol{X}\boldsymbol{b}$ for $\boldsymbol{b} = \hat{\boldsymbol{\beta}}$, it is a linear combination of the columns of $\boldsymbol{X}$, and hence it belongs to $\mathcal{S}(\boldsymbol{X})$.

From (3.19), it is easy to show that $\boldsymbol{P_X}\boldsymbol{X} = \boldsymbol{X}$. Since any vector in $\mathcal{S}(\boldsymbol{X})$ can be written as $\boldsymbol{X}\boldsymbol{b}$ for some $\boldsymbol{b} \in \mathbb{R}^k$, we see that

$$\boldsymbol{P_X}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{b}. \tag{3.21}$$

We saw from (3.20) that the result of acting on any vector $\boldsymbol{y} \in E^n$ with $\boldsymbol{P_X}$ is a vector in $\mathcal{S}(\boldsymbol{X})$. Thus the invariant subspace of the projection $\boldsymbol{P_X}$ must be contained in $\mathcal{S}(\boldsymbol{X})$. But, by (3.21), *every* vector in $\mathcal{S}(\boldsymbol{X})$ is mapped into itself by $\boldsymbol{P_X}$. Therefore, the **image** of $\boldsymbol{P_X}$, which is a shorter name for its invariant subspace, is precisely $\mathcal{S}(\boldsymbol{X})$.

It is clear from (3.20) that, when $\boldsymbol{P_X}$ is applied to $\boldsymbol{y}$, it yields the vector of fitted values. Similarly, when $\boldsymbol{M_X}$, the second of the two projection matrices in (3.19), is applied to $\boldsymbol{y}$, it yields the vector of residuals:

$$\boldsymbol{M_X}\boldsymbol{y} = \big(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{u}}.$$

The image of $\boldsymbol{M_X}$ is $\mathcal{S}^\perp(\boldsymbol{X})$, the orthogonal complement of the image of $\boldsymbol{P_X}$. To see this, consider any vector $\boldsymbol{w} \in \mathcal{S}^\perp(\boldsymbol{X})$. It must satisfy the defining condition $\boldsymbol{X}^\top\boldsymbol{w} = \mathbf{0}$. From the definition (3.19) of $\boldsymbol{P_X}$, this implies that $\boldsymbol{P_X}\boldsymbol{w} = \mathbf{0}$,

Copyright © 2021, Russell Davidson and James G. MacKinnon

the zero vector. Since $\boldsymbol{M_X} = \mathbf{I} - \boldsymbol{P_X}$, we find that $\boldsymbol{M_X}\boldsymbol{w} = \boldsymbol{w}$. Thus $\mathcal{S}^\perp(\boldsymbol{X})$ must be contained in the image of $\boldsymbol{M_X}$. Next, consider any vector in the image of $\boldsymbol{M_X}$. It must take the form $\boldsymbol{M_X}\boldsymbol{y}$, where $\boldsymbol{y}$ is some vector in E^n. From this, it follows that $\boldsymbol{M_X}\boldsymbol{y}$ belongs to $\mathcal{S}^\perp(\boldsymbol{X})$. Observe that

$$(\boldsymbol{M_X}\boldsymbol{y})^\top\boldsymbol{X} = \boldsymbol{y}^\top\boldsymbol{M_X}\boldsymbol{X}, \tag{3.22}$$

an equality that relies on the symmetry of $\boldsymbol{M_X}$. Then, from (3.19), we have

$$\boldsymbol{M_X}\boldsymbol{X} = (\mathbf{I} - \boldsymbol{P_X})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{O}, \tag{3.23}$$

where $\mathbf{O}$ denotes a **zero matrix**, which in this case is $n \times k$. The result (3.22) says that any vector $\boldsymbol{M_X}\boldsymbol{y}$ in the image of $\boldsymbol{M_X}$ is orthogonal to $\boldsymbol{X}$, and thus belongs to $\mathcal{S}^\perp(\boldsymbol{X})$. We saw above that $\mathcal{S}^\perp(\boldsymbol{X})$ was contained in the image of $\boldsymbol{M_X}$, and so this image must coincide with $\mathcal{S}^\perp(\boldsymbol{X})$. For obvious reasons, the projection $\boldsymbol{M_X}$ is sometimes called the projection **off** $\mathcal{S}(\boldsymbol{X})$.

For any matrix to represent a projection, it must be **idempotent**. An idempotent matrix is one that, when multiplied by itself, yields itself again. Thus,

$$\boldsymbol{P_X}\boldsymbol{P_X} = \boldsymbol{P_X} \quad \text{and} \quad \boldsymbol{M_X}\boldsymbol{M_X} = \boldsymbol{M_X}.$$

These results are easily proved by a little algebra directly from (3.19), but the geometry of the situation makes them obvious. If we take any point, project it on to $\mathcal{S}(\boldsymbol{X})$, and then project it on to $\mathcal{S}(\boldsymbol{X})$ *again*, the second projection can have no effect at all, because the point is *already* in $\mathcal{S}(\boldsymbol{X})$, and so it is left unchanged. Since this implies that $\boldsymbol{P_X}\boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{P_X}\boldsymbol{y}$ for any vector $\boldsymbol{y}$, it must be the case that $\boldsymbol{P_X}\boldsymbol{P_X} = \boldsymbol{P_X}$, and similarly for $\boldsymbol{M_X}$.

Since, from (3.19),

$$\boldsymbol{P_X} + \boldsymbol{M_X} = \mathbf{I}, \tag{3.24}$$

any vector $\boldsymbol{y} \in E^n$ is equal to $\boldsymbol{P_X}\boldsymbol{y} + \boldsymbol{M_X}\boldsymbol{y}$. The pair of projections $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are said to be **complementary projections**, since the sum of $\boldsymbol{P_X}\boldsymbol{y}$ and $\boldsymbol{M_X}\boldsymbol{y}$ restores the original vector $\boldsymbol{y}$.

The fact that $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}^\perp(\boldsymbol{X})$ are orthogonal subspaces leads us to say that the two projection matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ define what is called an **orthogonal decomposition** of E^n, because the two vectors $\boldsymbol{M_X}\boldsymbol{y}$ and $\boldsymbol{P_X}\boldsymbol{y}$ lie in the two orthogonal subspaces. Algebraically, the orthogonality depends on the fact that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric matrices. To see this, we start from a further important property of $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$, which is that

$$\boldsymbol{P_X}\boldsymbol{M_X} = \mathbf{O}. \tag{3.25}$$

This equation is true for any complementary pair of projections satisfying (3.24), whether or not they are symmetric; see Exercise 3.9. We may say that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ **annihilate** each other. Now consider any vector $\boldsymbol{z} \in \mathcal{S}(\boldsymbol{X})$

Copyright © 2021, Russell Davidson and James G. MacKinnon

and any other vector $\boldsymbol{w} \in \mathcal{S}^\perp(\boldsymbol{X})$. We have $\boldsymbol{z} = \boldsymbol{P_X z}$ and $\boldsymbol{w} = \boldsymbol{M_X w}$. Thus the scalar product of the two vectors is

$$\langle \boldsymbol{P_X z}, \boldsymbol{M_X w} \rangle = \boldsymbol{z}^\top \boldsymbol{P_X^\top M_X w}.$$

Since $\boldsymbol{P_X}$ is symmetric, $\boldsymbol{P_X^\top} = \boldsymbol{P_X}$, and so the above scalar product is zero by (3.25). In general, however, if two complementary projection matrices are not symmetric, the spaces they project on to are not orthogonal; see Exercise 3.10.

The projection matrix $\boldsymbol{M_X}$ annihilates all points that lie in $\mathcal{S}(\boldsymbol{X})$, and $\boldsymbol{P_X}$ likewise annihilates all points that lie in $\mathcal{S}^\perp(\boldsymbol{X})$. These properties can be proved by straightforward algebra (see Exercise 3.12), but the geometry of the situation is very simple. Consider Figure 3.7. It is evident that, if we project any point in $\mathcal{S}^\perp(\boldsymbol{X})$ orthogonally on to $\mathcal{S}(\boldsymbol{X})$, we end up at the origin, as we do if we project any point in $\mathcal{S}(\boldsymbol{X})$ orthogonally on to $\mathcal{S}^\perp(\boldsymbol{X})$.

Provided that $\boldsymbol{X}$ has full rank, the subspace $\mathcal{S}(\boldsymbol{X})$ is k–dimensional, and so the first term in the decomposition $\boldsymbol{y} = \boldsymbol{P_X y} + \boldsymbol{M_X y}$ belongs to a k–dimensional space. Since $\boldsymbol{y}$ itself belongs to E^n, which has n dimensions, it follows that the complementary space $\mathcal{S}^\perp(\boldsymbol{X})$ must have $n - k$ dimensions. The number $n - k$ is called the **codimension** of $\boldsymbol{X}$ in E^n.

Geometrically, an orthogonal decomposition $\boldsymbol{y} = \boldsymbol{P_X y} + \boldsymbol{M_X y}$ can be represented by a right-angled triangle, with $\boldsymbol{y}$ as the hypotenuse and $\boldsymbol{P_X y}$ and $\boldsymbol{M_X y}$ as the other two sides. In terms of projections, equation (3.17), which is really just Pythagoras' Theorem, can be rewritten as

$$\|\boldsymbol{y}\|^2 = \|\boldsymbol{P_X y}\|^2 + \|\boldsymbol{M_X y}\|^2. \tag{3.26}$$

In Exercise 3.11, readers are asked to provide an algebraic proof of this equation. Since every term in (3.26) is nonnegative, we obtain the useful result that, for any orthogonal projection matrix $\boldsymbol{P_X}$ and any vector $\boldsymbol{y} \in E^n$,

$$\|\boldsymbol{P_X y}\| \leq \|\boldsymbol{y}\|. \tag{3.27}$$

In effect, this just says that the hypotenuse is longer than either of the other sides of a right-angled triangle.

In general, we will use $\boldsymbol{P}$ and $\boldsymbol{M}$ subscripted by matrix expressions to denote the matrices that, respectively, project on to and off the subspaces spanned by the columns of those matrix expressions. Thus $\boldsymbol{P_Z}$ would be the matrix that projects on to $\mathcal{S}(\boldsymbol{Z})$, $\boldsymbol{M_{X,W}}$ would be the matrix that projects off $\mathcal{S}(\boldsymbol{X}, \boldsymbol{W})$, or, equivalently, on to $\mathcal{S}^\perp(\boldsymbol{X}, \boldsymbol{W})$, and so on. It is frequently very convenient to express the quantities that arise in econometrics using these matrices, partly because the resulting expressions are relatively compact, and partly because the properties of projection matrices often make it easy to understand what those expressions mean. However, projection matrices are of little use for computation because they are of dimension $n \times n$. It is never efficient to calculate residuals or fitted values by explicitly using projection matrices, and it can be extremely inefficient if n is large.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Linear Transformations of Regressors

The span $\mathcal{S}(\boldsymbol{X})$ of the regressors of a linear regression can be defined in many equivalent ways. All that is needed is a set of k vectors that encompass all the k directions of the k–dimensional subspace. Consider what happens when we postmultiply $\boldsymbol{X}$ by any nonsingular $k \times k$ matrix $\boldsymbol{A}$. This is called a **nonsingular linear transformation**. Let $\boldsymbol{A}$ be partitioned by its columns, which may be denoted $\boldsymbol{a}_i$, $i = 1, \ldots, k$:

$$\boldsymbol{XA} = \boldsymbol{X}\,[\,\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_k\,] = [\,\boldsymbol{Xa}_1 \quad \boldsymbol{Xa}_2 \quad \cdots \quad \boldsymbol{Xa}_k\,].$$

Each block in the product takes the form $\boldsymbol{Xa}_i$, which is an n–vector that is a linear combination of the columns of $\boldsymbol{X}$. Thus any element of $\mathcal{S}(\boldsymbol{XA})$ must also be an element of $\mathcal{S}(\boldsymbol{X})$. But any element of $\mathcal{S}(\boldsymbol{X})$ is also an element of $\mathcal{S}(\boldsymbol{XA})$. To see this, note that any element of $\mathcal{S}(\boldsymbol{X})$ can be written as $\boldsymbol{X\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$. Since $\boldsymbol{A}$ is nonsingular, and thus invertible,

$$\boldsymbol{X\beta} = \boldsymbol{XAA^{-1}\beta} = (\boldsymbol{XA})(\boldsymbol{A^{-1}\beta}).$$

Because $\boldsymbol{A^{-1}\beta}$ is just a k–vector, this expression is a linear combination of the columns of $\boldsymbol{XA}$, that is, an element of $\mathcal{S}(\boldsymbol{XA})$. Since every element of $\mathcal{S}(\boldsymbol{XA})$ belongs to $\mathcal{S}(\boldsymbol{X})$, and every element of $\mathcal{S}(\boldsymbol{X})$ belongs to $\mathcal{S}(\boldsymbol{XA})$, these two subspaces must be identical.

Given the identity of $\mathcal{S}(\boldsymbol{X})$ and $\mathcal{S}(\boldsymbol{XA})$, it seems intuitively compelling to suppose that the orthogonal projections $\boldsymbol{P_X}$ and $\boldsymbol{P_{XA}}$ should be the same. This is in fact the case, as can be verified directly:

$$\begin{aligned}
\boldsymbol{P_{XA}} &= \boldsymbol{XA}(\boldsymbol{A^\top X^\top XA})^{-1}\boldsymbol{A^\top X^\top} \\
&= \boldsymbol{XAA^{-1}}(\boldsymbol{X^\top X})^{-1}(\boldsymbol{A^\top})^{-1}\boldsymbol{A^\top X^\top} \\
&= \boldsymbol{X}(\boldsymbol{X^\top X})^{-1}\boldsymbol{X^\top} = \boldsymbol{P_X}.
\end{aligned}$$

When expanding the inverse of the matrix $\boldsymbol{A^\top X^\top XA}$, we used the reversal rule for inverses; see Exercise 2.17.

We have already seen that the vectors of fitted values and residuals depend on $\boldsymbol{X}$ only through $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$. Therefore, they too must be invariant to any nonsingular linear transformation of the columns of $\boldsymbol{X}$. Thus if, in the regression $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, we replace $\boldsymbol{X}$ by $\boldsymbol{XA}$ for some nonsingular matrix $\boldsymbol{A}$, the residuals and fitted values do not change, even though $\hat{\boldsymbol{\beta}}$ changes. We will discuss an example of this important result shortly.

When the set of regressors contains a constant, it is necessary to express it as a vector, just like any other regressor. The coefficient of this **constant vector** is then the parameter that we usually call the constant term. The constant vector is just $\boldsymbol{\iota}$, the vector of which each element equals 1. Consider the n–vector $\beta_1\boldsymbol{\iota} + \beta_2\boldsymbol{x}$, where $\boldsymbol{x}$ is any nonconstant regressor, and β_1 and β_2 are

Copyright © 2021, Russell Davidson and James G. MacKinnon

scalar parameters. The t^{th} element of this vector is $\beta_1 + \beta_2 x_t$. Thus adding the vector $\beta_1 \boldsymbol{\iota}$ to $\beta_2 \boldsymbol{x}$ simply adds the scalar β_1 to each component of $\beta_2 \boldsymbol{x}$. For any regression which includes a constant term, then, the fact that we can perform arbitrary nonsingular transformations of the regressors without affecting residuals or fitted values implies that these vectors are unchanged if we add any constant amount to any one or more of the regressors.

Another implication of the invariance of residuals and fitted values under nonsingular transformations of the regressors is that these vectors are unchanged if we change the **units of measurement** of the regressors. Suppose, for instance, that the temperature is one of the explanatory variables in a regression with a constant term. A practical example in which the temperature could have good explanatory power is the modeling of electricity demand: More electrical power is consumed if the weather is very cold, or, in societies where air conditioners are common, very hot. In a few countries, notably the United States, temperatures are still measured in Fahrenheit degrees, while in most countries they are measured in Celsius (centigrade) degrees. It would be disturbing if our conclusions about the effect of temperature on electricity demand depended on whether we measured it using the Fahrenheit scale or the Celsius scale.

Let the n–vector of observations on the temperature variable be denoted as $\boldsymbol{T}$ in Celsius and as $\boldsymbol{F}$ in Fahrenheit, the constant vector being denoted, as usual, by $\boldsymbol{\iota}$. Then we have the relation

$$\boldsymbol{F} = 32\boldsymbol{\iota} + \frac{9}{5}\boldsymbol{T}.$$

If the constant is included in the transformation,

$$[\,\boldsymbol{\iota} \quad \boldsymbol{F}\,] = [\,\boldsymbol{\iota} \quad \boldsymbol{T}\,] \begin{bmatrix} 1 & 32 \\ 0 & 9/5 \end{bmatrix}. \tag{3.28}$$

Thus the constant and the two different temperature measures are related by a linear transformation that is easily seen to be nonsingular, since Fahrenheit degrees can be converted back into Celsius. This implies that the residuals and fitted values are unaffected by our choice of temperature scale.

Let us denote the constant term and the slope coefficient as β_1 and β_2 if we use the Celsius scale, and as α_1 and α_2 if we use the Fahrenheit scale. Then it is easy to see that these parameters are related by the equations

$$\beta_1 = \alpha_1 + 32\alpha_2 \quad \text{and} \quad \beta_2 = 9/5\,\alpha_2. \tag{3.29}$$

To see that this makes sense, suppose that the temperature is at freezing point, which is 0° Celsius and 32° Fahrenheit. Then the combined effect of the constant and the temperature on electricity demand is $\beta_1 + 0\beta_2 = \beta_1$ using the Celsius scale, and $\alpha_1 + 32\alpha_2$ using the Fahrenheit scale. These should be the same, and, according to (3.29), they are. Similarly, the effect of

Copyright © 2021, Russell Davidson and James G. MacKinnon

a 1-degree increase in the Celsius temperature is given by β_2. Now 1 Celsius degree equals $9/5$ Fahrenheit degrees, and the effect of a temperature increase of $9/5$ Fahrenheit degrees is given by $9/5\,\alpha_2$. We are assured by (3.29) that the two effects are the same.

3.4 The Frisch-Waugh-Lovell Theorem

In this section, we discuss an extremely useful property of least-squares estimates, which we will refer to as the **Frisch-Waugh-Lovell Theorem**, or **FWL Theorem** for short. It was introduced to econometricians by Frisch and Waugh (1933), and then reintroduced by Lovell (1963).

Deviations from the Mean

We begin by considering a particular nonsingular transformation of variables in a regression with a constant term. We saw at the end of the last section that residuals and fitted values are invariant under such transformations of the regressors. For simplicity, consider a model with a constant and just one explanatory variable:

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{x} + \boldsymbol{u}. \tag{3.30}$$

In general, $\boldsymbol{x}$ is not orthogonal to $\boldsymbol{\iota}$, but there is a very simple transformation which makes it so. This transformation replaces the observations in $\boldsymbol{x}$ by **deviations from the mean**. In order to perform the transformation, one first calculates the mean of the n observations of the vector $\boldsymbol{x}$,

$$\bar{x} \equiv \frac{1}{n}\sum_{t=1}^{n} x_t,$$

and then subtracts the constant $\bar{x}$ from each element of $\boldsymbol{x}$. This yields the vector of deviations from the mean, $\boldsymbol{z} \equiv \boldsymbol{x} - \bar{x}\boldsymbol{\iota}$. The vector $\boldsymbol{z}$ is easily seen to be orthogonal to $\boldsymbol{\iota}$, because

$$\boldsymbol{\iota}^\top \boldsymbol{z} = \boldsymbol{\iota}^\top(\boldsymbol{x} - \bar{x}\boldsymbol{\iota}) = n\bar{x} - \bar{x}\boldsymbol{\iota}^\top\boldsymbol{\iota} = n\bar{x} - n\bar{x} = 0.$$

The operation of expressing a variable in terms of the deviations from its mean is called **centering** the variable. In this case, the vector $\boldsymbol{z}$ is the **centered** version of the vector $\boldsymbol{x}$.

Since centering leads to a variable that is orthogonal to $\boldsymbol{\iota}$, it can be performed algebraically by the orthogonal projection matrix $\boldsymbol{M}_{\boldsymbol{\iota}}$. This can be verified by observing that

$$\boldsymbol{M}_{\boldsymbol{\iota}}\boldsymbol{x} = (\mathbf{I} - \boldsymbol{P}_{\boldsymbol{\iota}})\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^\top\boldsymbol{x} = \boldsymbol{x} - \bar{x}\boldsymbol{\iota} = \boldsymbol{z}, \tag{3.31}$$

as claimed. Here, we once again used the facts that $\boldsymbol{\iota}^\top\boldsymbol{\iota} = n$ and $\boldsymbol{\iota}^\top\boldsymbol{x} = n\bar{x}$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

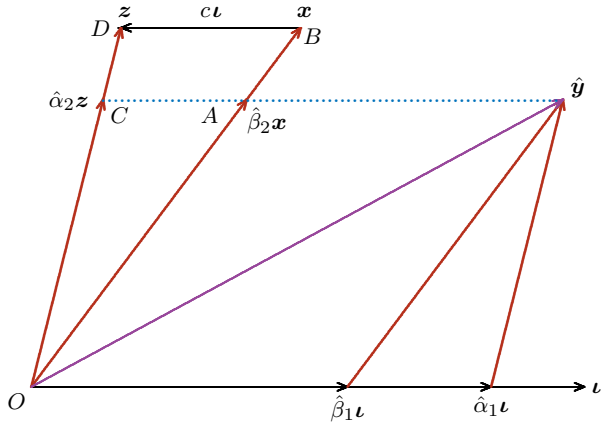


Figure 3.12 Adding a constant does not affect the slope coefficient

The idea behind the use of deviations from the mean is that it makes sense to separate the overall level of a dependent variable from its dependence on explanatory variables. Specifically, if we rewrite equation (3.30) in terms of $\boldsymbol{z}$, it becomes

$$\boldsymbol{y} = (\beta_1 + \beta_2\bar{x})\boldsymbol{\iota} + \beta_2\boldsymbol{z} + \boldsymbol{u} = \alpha_1\boldsymbol{\iota} + \alpha_2\boldsymbol{z} + \boldsymbol{u},$$

from which it is evident that

$$\alpha_1 = \beta_1 + \beta_2\bar{x}, \text{ and } \alpha_2 = \beta_2.$$

If, for some observation t, the value of x_t were exactly equal to the mean value, $\bar{x}$, then $z_t = 0$. Thus we find that $y_t = \alpha_1 + u_t$. We interpret this as saying that the expected value of y_t, when the explanatory variable takes on its average value, is the constant α_1.

The effect on y_t of a change of one unit in x_t is measured by the slope coefficient β_2. If we hold $\bar{x}$ at its value before x_t is changed, then the unit change in x_t induces a unit change in z_t. Thus a unit change in z_t, which is measured by the slope coefficient α_2, should have the same effect as a unit change in x_t. Accordingly, $\alpha_2 = \beta_2$, just as we found above.

The slope coefficients α_2 and β_2 would be the same with any constant in the place of $\bar{x}$. The reason for this can be seen geometrically, as illustrated in Figure 3.12. This figure, which is constructed in the same way as panel (b) of Figure 3.11, depicts the span of $\boldsymbol{\iota}$ and $\boldsymbol{x}$, with $\boldsymbol{\iota}$ in the horizontal direction. As before, the vector $\boldsymbol{y}$ is not shown, because a third dimension would be required; the vector would extend from the origin to a point off the plane of the page and directly above (or below) the point labelled $\hat{\boldsymbol{y}}$.

The figure shows the vector of fitted values $\hat{\boldsymbol{y}}$ as the vector sum $\hat{\beta}_1\boldsymbol{\iota} + \hat{\beta}_2\boldsymbol{x}$. The slope coefficient $\hat{\beta}_2$ is the ratio of the length of the vector $\hat{\beta}_2\boldsymbol{x}$ to that

Copyright © 2021, Russell Davidson and James G. MacKinnon

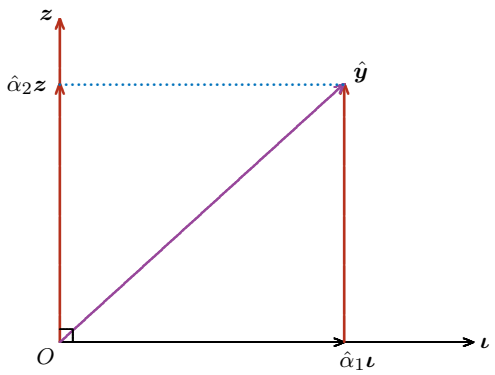


Figure 3.13 Orthogonal regressors may be omitted

of $\boldsymbol{x}$; geometrically, it is given by the ratio OA/OB. Then a new regressor $\boldsymbol{z}$ is defined by adding the constant value c, which is negative in the figure, to each component of $\boldsymbol{x}$, giving $\boldsymbol{z} = \boldsymbol{x} + c\boldsymbol{\iota}$. In terms of this new regressor, the vector $\hat{\boldsymbol{y}}$ is given by $\hat{\alpha}_1\boldsymbol{\iota} + \hat{\alpha}_2\boldsymbol{z}$, and $\hat{\alpha}_2$ is given by the ratio OC/OD. Since the ratios OA/OB and OC/OD are clearly the same, we see that $\hat{\alpha}_2 = \hat{\beta}_2$. A formal argument would use the fact that OAC and OBD are similar triangles.

When the constant c is chosen as $\bar{x}$, the vector $\boldsymbol{z}$ is said to be centered, and, as we saw above, it is orthogonal to $\boldsymbol{\iota}$. In this case, the estimate $\hat{\alpha}_2$ is the same whether it is obtained by regressing $\boldsymbol{y}$ on both $\boldsymbol{\iota}$ and $\boldsymbol{z}$, or just on $\boldsymbol{z}$ alone. This is illustrated in Figure 3.13, which shows what Figure 3.12 would look like when $\boldsymbol{z}$ is orthogonal to $\boldsymbol{\iota}$. Once again, the vector of fitted values $\hat{\boldsymbol{y}}$ is decomposed as $\hat{\alpha}_1\boldsymbol{\iota} + \hat{\alpha}_2\boldsymbol{z}$, with $\boldsymbol{z}$ now at right angles to $\boldsymbol{\iota}$.

Now suppose that $\boldsymbol{y}$ is regressed on $\boldsymbol{z}$ alone. This means that $\boldsymbol{y}$ is projected orthogonally on to $\mathcal{S}(\boldsymbol{z})$, which in the figure is the vertical line through $\boldsymbol{z}$. By definition,

$$\boldsymbol{y} = \hat{\alpha}_1\boldsymbol{\iota} + \hat{\alpha}_2\boldsymbol{z} + \hat{\boldsymbol{u}}, \tag{3.32}$$

where $\hat{\boldsymbol{u}}$ is orthogonal to both $\boldsymbol{\iota}$ and $\boldsymbol{z}$. But $\boldsymbol{\iota}$ is also orthogonal to $\boldsymbol{z}$, and so the only term on the right-hand side of (3.32) not to be annihilated by the projection on to $\mathcal{S}(\boldsymbol{z})$ is the middle term, which is left unchanged by it. Thus the fitted value vector from regressing $\boldsymbol{y}$ on $\boldsymbol{z}$ alone is just $\hat{\alpha}_2\boldsymbol{z}$, and so the OLS estimate is the same $\hat{\alpha}_2$ as given by the regression on both $\boldsymbol{\iota}$ and $\boldsymbol{z}$. Geometrically, we obtain this result because the projection of $\boldsymbol{y}$ on to $\mathcal{S}(\boldsymbol{z})$ is the same as the projection of $\hat{\boldsymbol{y}}$ on to $\mathcal{S}(\boldsymbol{z})$.

Incidentally, the fact that OLS residuals are orthogonal to all the regressors, including $\boldsymbol{\iota}$, leads to the important result that the residuals in any regression

Copyright © 2021, Russell Davidson and James G. MacKinnon

with a constant term sum to zero. In fact,

$$\boldsymbol{\iota}^\top \hat{\boldsymbol{u}} = \sum_{t=1}^{n} \hat{u}_t = 0;$$

recall equation (2.29). The residuals also sum to zero in any regression for which $\boldsymbol{\iota} \in \mathcal{S}(\boldsymbol{X})$, even if $\boldsymbol{\iota}$ does not explicitly appear in the list of regressors. This can happen if the regressors include certain sets of **dummy variables**, as we will see in Section 3.5.

Two Groups of Regressors

The results proved in the previous subsection are actually special cases of more general results that apply to any regression in which the regressors can logically be broken up into two groups. Such a regression can be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \tag{3.33}$$

where $\boldsymbol{X}_1$ is $n \times k_1$, $\boldsymbol{X}_2$ is $n \times k_2$, and $\boldsymbol{X}$ may be written as the partitioned matrix $[\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, with $k = k_1 + k_2$. In the case dealt with in the previous subsection, $\boldsymbol{X}_1$ is the constant vector $\boldsymbol{\iota}$ and $\boldsymbol{X}_2$ is either $\boldsymbol{x}$ or $\boldsymbol{z}$. Several other examples of partitioning $\boldsymbol{X}$ in this way will be considered in Section 3.5.

We begin by assuming that all the regressors in $\boldsymbol{X}_1$ are orthogonal to all the regressors in $\boldsymbol{X}_2$, so that $\boldsymbol{X}_2^\top \boldsymbol{X}_1 = \boldsymbol{O}$. Under this assumption, the vector of least-squares estimates $\hat{\boldsymbol{\beta}}_1$ from (3.33) is the same as the one obtained from the regression

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}_1, \tag{3.34}$$

and $\hat{\boldsymbol{\beta}}_2$ from (3.33) is likewise the same as the vector of estimates obtained from the regression $\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}_2$. In other words, when $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal, we can drop either set of regressors from (3.33) without affecting the coefficients of the other set.

The vector of fitted values from (3.33) is $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$, while that from (3.34) is $\boldsymbol{P}_1\boldsymbol{y}$, where we have used the abbreviated notation

$$\boldsymbol{P}_1 \equiv \boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top.$$

As we will show directly,

$$\boldsymbol{P}_1\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_1 = \boldsymbol{P}_1; \tag{3.35}$$

this is true whether or not $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal. Thus

$$\boldsymbol{P}_1\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{P}_1(\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2) = \boldsymbol{P}_1\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1. \tag{3.36}$$

The first equality above, which follows from (3.35), says that the projection of $\boldsymbol{y}$ on to $\mathcal{S}(\boldsymbol{X}_1)$ is the same as the projection of $\hat{\boldsymbol{y}} \equiv \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$ on to $\mathcal{S}(\boldsymbol{X}_1)$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The second equality follows from the definition of the fitted value vector from (3.33) as $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$; the third from the orthogonality of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, which implies that $\boldsymbol{P}_1\boldsymbol{X}_2 = \boldsymbol{O}$; and the last from the fact that $\boldsymbol{X}_1$ is invariant under the action of $\boldsymbol{P}_1$. Since $\boldsymbol{P}_1\boldsymbol{y}$ is equal to $\boldsymbol{X}_1$ postmultiplied by the OLS estimates from (3.34), the equality of the leftmost and rightmost expressions in (3.36) gives us the result that the same $\hat{\boldsymbol{\beta}}_1$ can be obtained either from (3.33) or from (3.34). The analogous result for $\hat{\boldsymbol{\beta}}_2$ is proved in just the same way.

We now drop the assumption that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal and prove (3.35), a very useful result that is true in general. In order to show that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_1 = \boldsymbol{P}_1$, we proceed as follows:

$$\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_1 = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top = \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top = \boldsymbol{P}_1.$$

The middle equality follows by noting that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}_1 = \boldsymbol{X}_1$, because all the columns of $\boldsymbol{X}_1$ are in $\mathcal{S}(\boldsymbol{X})$, and so are left unchanged by $\boldsymbol{P}_{\boldsymbol{X}}$. The other equality in (3.35), namely $\boldsymbol{P}_1\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_1$, is obtained directly by transposing $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_1 = \boldsymbol{P}_1$ and using the symmetry of $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{P}_1$. The two results in (3.35) tell us that the product of two orthogonal projections, where one projects on to a subspace of the image of the other, is the projection on to that subspace. See also Exercise 3.15, for the application of this result to the complementary projections $\boldsymbol{M}_{\boldsymbol{X}}$ and $\boldsymbol{M}_1$.

The general result corresponding to the one shown in Figure 3.12 can be stated as follows. If we transform the regressor matrix in (3.33) by adding $\boldsymbol{X}_1\boldsymbol{A}$ to $\boldsymbol{X}_2$, where $\boldsymbol{A}$ is a $k_1 \times k_2$ matrix, and leaving $\boldsymbol{X}_1$ as it is, we have the regression

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\alpha}_1 + (\boldsymbol{X}_2 + \boldsymbol{X}_1\boldsymbol{A})\boldsymbol{\alpha}_2 + \boldsymbol{u}. \tag{3.37}$$

Then $\hat{\boldsymbol{\alpha}}_2$ from (3.37) is the same as $\hat{\boldsymbol{\beta}}_2$ from (3.33). This can be seen immediately by expressing the right-hand side of (3.37) as a linear combination of the columns of $\boldsymbol{X}_1$ and of $\boldsymbol{X}_2$.

In the present general context, there is an operation analogous to that of centering. The result of centering a variable $\boldsymbol{x}$ is a variable $\boldsymbol{z}$ that is orthogonal to $\boldsymbol{\iota}$, the constant. We can create from $\boldsymbol{X}_2$ a set of variables orthogonal to $\boldsymbol{X}_1$ by acting on $\boldsymbol{X}_2$ with the orthogonal projection $\boldsymbol{M}_1 \equiv \boldsymbol{I} - \boldsymbol{P}_1$, so as to obtain $\boldsymbol{M}_1\boldsymbol{X}_2$. This allows us to run the regression

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\alpha}_2 + \boldsymbol{u} \\ &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \big(\boldsymbol{X}_2 - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2\big)\boldsymbol{\alpha}_2 + \boldsymbol{u}.\end{aligned}$$

The first line above is a regression model with two groups of regressors, $\boldsymbol{X}_1$ and $\boldsymbol{M}_1\boldsymbol{X}_2$, which are mutually orthogonal. Therefore, $\hat{\boldsymbol{\alpha}}_2$ is unchanged if we omit $\boldsymbol{X}_1$. The second line makes it clear that this regression is a special case of (3.37), which implies that $\hat{\boldsymbol{\alpha}}_2$ is equal to $\hat{\boldsymbol{\beta}}_2$ from (3.33). Consequently, we see that the two regressions

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u} \quad \text{and} \tag{3.38}$$

$$\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{v} \tag{3.39}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

yield the same estimates of $\boldsymbol{\beta}_2$.

Although regressions (3.33) and (3.39) give the same estimates of $\boldsymbol{\beta}_2$, they do not give the same residuals, as we have indicated by writing $\boldsymbol{u}$ for one regression and $\boldsymbol{v}$ for the other. We can see why the residuals are not the same by looking again at Figure 3.13, in which the constant $\boldsymbol{\iota}$ plays the role of $\boldsymbol{X}_1$, and the centered variable $\boldsymbol{z}$ plays the role of $\boldsymbol{M}_1\boldsymbol{X}_2$. The point corresponding to $\boldsymbol{y}$ can be thought of as lying somewhere on a line through the point $\hat{\boldsymbol{y}}$ and sticking perpendicularly out from the page. The residual vector from regressing $\boldsymbol{y}$ on both $\boldsymbol{\iota}$ and $\boldsymbol{z}$ is thus represented by the line segment from $\hat{\boldsymbol{y}}$, in the page, to $\boldsymbol{y}$, vertically above the page. However, if $\boldsymbol{y}$ is regressed on $\boldsymbol{z}$ alone, the residual vector is the sum of this line segment and the segment from $\hat{\alpha}_2\boldsymbol{z}$ and $\hat{\boldsymbol{y}}$, that is, the top side of the rectangle in the figure. If we want the same residuals in regression (3.33) and a regression like (3.39), we need to purge the dependent variable of the second segment, which can be seen from the figure to be equal to $\hat{\alpha}_1\boldsymbol{\iota}$.

This suggests replacing $\boldsymbol{y}$ by what we get by projecting $\boldsymbol{y}$ off $\boldsymbol{\iota}$. This projection would be the line segment perpendicular to the page, translated in the horizontal direction so that it intersected the page at the point $\hat{\alpha}_2\boldsymbol{z}$ rather than $\hat{\boldsymbol{y}}$. In the general context, the analogous operation replaces $\boldsymbol{y}$ by $\boldsymbol{M}_1\boldsymbol{y}$, the projection off $\boldsymbol{X}_1$ rather than off $\boldsymbol{\iota}$. When we perform this projection, (3.39) is replaced by the regression

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals}, \tag{3.40}$$

which yields the same vector of OLS estimates $\hat{\boldsymbol{\beta}}_2$ as regression (3.33), and also the same vector of residuals. This regression is sometimes called the **FWL regression**. We used the notation "+ residuals" instead of "+ $\boldsymbol{u}$" in (3.40) because, in general, the difference between $\boldsymbol{M}_1\boldsymbol{y}$ and $\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2$ is not the same thing as the vector $\boldsymbol{u}$ in (3.33). If $\boldsymbol{u}$ is interpreted as a vector of disturbances, then (3.40) would not be true if "residuals" were replaced by $\boldsymbol{u}$.

We can now formally state the FWL Theorem. Although the conclusions of the theorem have been established gradually in this section, we also provide a short formal proof.

Theorem 3.1. (Frisch-Waugh-Lovell Theorem)

1. The OLS estimates of $\boldsymbol{\beta}_2$ from regressions (3.33) and (3.40) are numerically identical.
2. The OLS residuals from regressions (3.33) and (3.40) are numerically identical.

Proof: By the standard formula (2.45), the estimate of $\boldsymbol{\beta}_2$ from (3.40) is

$$(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y}. \tag{3.41}$$

Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ denote the two vectors of OLS estimates from (3.33). Then

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}. \tag{3.42}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Premultiplying the leftmost and rightmost expressions in (3.42) by $\boldsymbol{X}_2^\top\boldsymbol{M}_1$, we obtain

$$\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2. \tag{3.43}$$

The first term on the right-hand side of (3.42) has dropped out because $\boldsymbol{M}_1$ annihilates $\boldsymbol{X}_1$. To see that the last term also drops out, observe that

$$\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X}_2 = \boldsymbol{O}. \tag{3.44}$$

The first equality follows from (3.35) (see also Exercise 3.15), and the second from (3.23), which shows that $\boldsymbol{M}_{\boldsymbol{X}}$ annihilates all the columns of $\boldsymbol{X}$, in particular those of $\boldsymbol{X}_2$. Premultiplying $\boldsymbol{y}$ by the transpose of (3.44) shows that $\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{0}$. We can now solve (3.43) for $\hat{\boldsymbol{\beta}}_2$ to obtain

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y},$$

which is expression (3.41). This proves the first part of the theorem.

If we had premultiplied (3.42) by $\boldsymbol{M}_1$ instead of by $\boldsymbol{X}_2^\top\boldsymbol{M}_1$, we would have obtained

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}, \tag{3.45}$$

where the last term is unchanged from (3.42) because $\boldsymbol{M}_1\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{M}_{\boldsymbol{X}}$. The regressand in (3.45) is the regressand from regression (3.40). Because $\hat{\boldsymbol{\beta}}_2$ is the estimate of $\boldsymbol{\beta}_2$ from (3.40), by the first part of the theorem, the first term on the right-hand side of (3.45) is the vector of fitted values from that regression. Thus the second term must be the vector of residuals from regression (3.40). But $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ is also the vector of residuals from regression (3.33), and this therefore proves the second part of the theorem. ■

3.5 Applications of the FWL Theorem

A regression like (3.33), in which the regressors are broken up into two groups, can arise in many situations. In this section, we will study three of these, namely, seasonal dummy variables, time trends, and fixed effects. In all cases, the FWL Theorem allows us to obtain explicit expressions based on (3.41) for subsets of the parameter estimates of a linear regression.

Seasonal Dummy Variables

For a variety of reasons, it is sometimes desirable for the explanatory variables of a regression model to include variables that can take on only two possible values, which are usually 0 and 1. Such variables are called **indicator variables**, because they indicate a subset of the observations, namely, those for which the value of the variable is 1. Indicator variables are a special case of **dummy variables**, which can take on more than two possible values.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Seasonal variation provides a good reason to employ dummy variables. It is common for economic data that are indexed by time to take the form of **quarterly data**, where each year in the sample period is represented by four observations, one for each quarter, or season, of the year. Many economic activities are strongly affected by the season, for obvious reasons like Christmas shopping, or summer holidays, or the difficulty of doing outdoor work during very cold weather. This seasonal variation, or **seasonality**, in economic activity is likely to be reflected in the economic **time series** that are used in regression models. The term "time series" is used to refer to any variable the observations of which are indexed by time. Of course, time-series data are sometimes annual, in which case there is no seasonal variation to worry about, and sometimes monthly, in which case there are twelve "seasons" instead of four. For simplicity, we consider only the case of quarterly data.

Since there are four seasons, there may be four **seasonal dummy variables**, each taking the value 1 for just one of the four seasons. Let us denote these variables as $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, $\boldsymbol{s}_3$, and $\boldsymbol{s}_4$. If we consider a sample the first observation of which corresponds to the first quarter of some year, these variables look like

$$\boldsymbol{s}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}, \quad \boldsymbol{s}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}, \quad \boldsymbol{s}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}, \quad \boldsymbol{s}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix}. \tag{3.46}$$

An important property of these variables is that, since every observation must correspond to some season, the sum of the seasonal dummies must indicate every season. This means that this sum is a vector every component of which equals 1. Algebraically,

$$\boldsymbol{s}_1 + \boldsymbol{s}_2 + \boldsymbol{s}_3 + \boldsymbol{s}_4 = \boldsymbol{\iota}, \tag{3.47}$$

as is clear from (3.46). Since $\boldsymbol{\iota}$ represents the constant in a regression, (3.47) means that the five-variable set consisting of all four seasonal dummies plus the constant is linearly dependent. Consequently, one of the five variables must be dropped if all the regressors are to be linearly independent.

Just which one of the five variables is dropped makes no difference to the fitted values and residuals of a regression, because it is easy to check that

$$\mathcal{S}(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4) = \mathcal{S}(\boldsymbol{\iota}, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4) = \mathcal{S}(\boldsymbol{\iota}, \boldsymbol{s}_1, \boldsymbol{s}_3, \boldsymbol{s}_4),$$

and so on. However the parameter estimates associated with the set of four variables that we choose to keep have different interpretations depending on

Copyright © 2021, Russell Davidson and James G. MacKinnon

that choice. Suppose first that we drop the constant and run the regression

$$\boldsymbol{y} = \alpha_1 \boldsymbol{s}_1 + \alpha_2 \boldsymbol{s}_2 + \alpha_3 \boldsymbol{s}_3 + \alpha_4 \boldsymbol{s}_4 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{3.48}$$

where the matrix $\boldsymbol{X}$ contains other explanatory variables. If observation t corresponds to the first season, the values of $\boldsymbol{s}_2$, $\boldsymbol{s}_3$, and $\boldsymbol{s}_4$ are all 0, and that of $\boldsymbol{s}_1$ is 1. Thus, if we write out the t^{th} observation of (3.48), we get

$$y_t = \alpha_1 + \boldsymbol{X}_t\boldsymbol{\beta} + u_t.$$

For all t belonging to the first season, the constant term in the regression is evidently α_1. If we repeat this exercise for t in each of the other seasons, we see at once that α_i is the constant for season i. Thus the introduction of the seasonal dummies gives us a different constant for every season.

An alternative is to retain the constant and drop $\boldsymbol{s}_1$. This yields

$$\boldsymbol{y} = \alpha_0 \boldsymbol{\iota} + \gamma_2 \boldsymbol{s}_2 + \gamma_3 \boldsymbol{s}_3 + \gamma_4 \boldsymbol{s}_4 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}.$$

It is clear that, in this specification, the overall constant α_0 is really the constant for season 1. For an observation belonging to season 2, the constant is $\alpha_0 + \gamma_2$, for an observation belonging to season 3, it is $\alpha_0 + \gamma_3$, and so on. The easiest way to interpret this is to think of season 1 as the reference season. The coefficients γ_i, $i = 2, 3, 4$, measure the difference between α_0, the constant for the reference season, and the constant for season i. Since we could have dropped any of the seasonal dummies, the choice of reference season is, of course, entirely arbitrary.

Another alternative is to retain the constant and use the three dummy variables defined by

$$\boldsymbol{s}'_1 = \boldsymbol{s}_1 - \boldsymbol{s}_4, \quad \boldsymbol{s}'_2 = \boldsymbol{s}_2 - \boldsymbol{s}_4, \quad \boldsymbol{s}'_3 = \boldsymbol{s}_3 - \boldsymbol{s}_4. \tag{3.49}$$

These new dummy variables are not actually indicator variables, because their components for season 4 are equal to -1, but they have the advantage that, for each complete year, the sum of their components for that year is 0. Thus, for any sample whose size is a multiple of 4, each of the $\boldsymbol{s}'_i$, $i = 1, 2, 3$, is orthogonal to the constant. We can write the regression as

$$\boldsymbol{y} = \delta_0 \boldsymbol{\iota} + \delta_1 \boldsymbol{s}'_1 + \delta_2 \boldsymbol{s}'_2 + \delta_3 \boldsymbol{s}'_3 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{3.50}$$

For t in season i, $i = 1, 2, 3$, the constant term is $\delta_0 + \delta_i$. For t belonging to season 4, it is $\delta_0 - \delta_1 - \delta_2 - \delta_3$. Thus the average of the constants for all four seasons is just δ_0, the coefficient of the constant, $\boldsymbol{\iota}$. Accordingly, the δ_i, $i = 1, 2, 3$, measure the difference between the average constant δ_0 and the constant specific to season i. Season 4 is more complicated, because of the arithmetic needed to ensure that the average does indeed work out to δ_0.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Let $\boldsymbol{S}$ denote whatever $n \times 4$ matrix we choose to use in order to span the constant and the four seasonal variables $\boldsymbol{s}_i$. Then any of the regressions we have considered so far can be written as

$$\boldsymbol{y} = \boldsymbol{S\delta} + \boldsymbol{X\beta} + \boldsymbol{u}. \tag{3.51}$$

This regression has two groups of regressors, as required for the application of the FWL Theorem. That theorem implies that the estimates $\hat{\boldsymbol{\beta}}$ and the residuals $\hat{\boldsymbol{u}}$ can also be obtained by running the FWL regression

$$\boldsymbol{M_S y} = \boldsymbol{M_S X\beta} + \text{residuals}, \tag{3.52}$$

where, as the notation suggests, $\boldsymbol{M_S} \equiv \mathbf{I} - \boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{S})^{-1}\boldsymbol{S}^\top$.

The effect of the projection $\boldsymbol{M_S}$ on $\boldsymbol{y}$ and on the explanatory variables in the matrix $\boldsymbol{X}$ can be considered as a form of **seasonal adjustment**. By making $\boldsymbol{M_S y}$ orthogonal to all the seasonal variables, we are, in effect, purging it of its seasonal variation. Consequently, $\boldsymbol{M_S y}$ can be called a **seasonally adjusted**, or **deseasonalized**, version of $\boldsymbol{y}$, and similarly for the explanatory variables. In practice, such seasonally adjusted variables can be conveniently obtained as the residuals from regressing $\boldsymbol{y}$ and each of the columns of $\boldsymbol{X}$ on the variables in $\boldsymbol{S}$. The FWL Theorem tells us that we get the same results in terms of estimates of $\boldsymbol{\beta}$ and residuals whether we run (3.51), in which the variables are unadjusted and seasonality is explicitly accounted for, or run (3.52), in which all the variables are seasonally adjusted by regression. This was, in fact, the subject of the famous paper by Lovell (1963).

The equivalence of (3.51) and (3.52) is sometimes used to claim that, in estimating a regression model with time-series data, it does not matter whether one uses "raw" data, along with seasonal dummies, or seasonally adjusted data. Such a conclusion is completely unwarranted. Official seasonal adjustment procedures are almost never based on regression; using official seasonally adjusted data is therefore *not* equivalent to using residuals from regression on a set of seasonal variables. Moreover, if (3.51) is not a sensible model (and it would not be if, for example, the seasonal pattern were more complicated than that given by $\boldsymbol{S\delta}$), then (3.52) is not a sensible specification either. Seasonality is actually an important practical problem in applied work with time-series data. For more detailed treatments, see Hylleberg (1986, 1992) and Ghysels and Osborn (2001).

The deseasonalization performed by the projection $\boldsymbol{M_S}$ makes all variables orthogonal to the constant as well as to the seasonal dummies. Thus the effect of $\boldsymbol{M_S}$ is not only to deseasonalize, but also to center, the variables on which it acts. Sometimes this is undesirable; if so, we may use the three variables $\boldsymbol{s}'_i$ given in (3.49). Since they are themselves orthogonal to the constant, no centering takes place if only these three variables are used for seasonal adjustment. An explicit constant should normally be included in any regression that uses variables seasonally adjusted in this way.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Time Trends

Another sort of constructed, or artificial, variable that is often encountered in models of time-series data is a **time trend**. The simplest sort of time trend is the **linear time trend**, represented by the vector $\boldsymbol{T}$, with typical element $T_t \equiv t$. Thus $\boldsymbol{T} = [1 \vdots 2 \vdots 3 \vdots 4 \vdots \dots]$. Imagine that we have a regression with a constant and a linear time trend:

$$\boldsymbol{y} = \gamma_1\boldsymbol{\iota} + \gamma_2\boldsymbol{T} + \boldsymbol{X\beta} + \boldsymbol{u}.$$

For observation t, y_t is equal to $\gamma_1 + \gamma_2 t + \boldsymbol{X}_t\boldsymbol{\beta} + u_t$. Thus the overall level of y_t increases or decreases steadily as t increases. Instead of just a constant, we now have the affine function of time, $\gamma_1 + \gamma_2 t$. An increasing time trend might be appropriate, for instance, in a model of a production function where technical progress is taking place. An explicit model of technical progress might well be difficult to construct, in which case a linear time trend could serve as a simple way to take account of the phenomenon.

It is often desirable to make the time trend orthogonal to the constant by centering it, that is, operating on it with $\boldsymbol{M}_{\boldsymbol{\iota}}$. If we do this with a sample with an odd number of elements, the result is a variable that looks like

$$[\,\cdots \vdots -3 \vdots -2 \vdots -1 \vdots 0 \vdots 1 \vdots 2 \vdots 3 \vdots \cdots\,].$$

If the sample size is even, the variable is made up of the half integers $\pm 1/2$, $\pm 3/2$, $\pm 5/2, \dots$. In both cases, the coefficient of $\boldsymbol{\iota}$ is the average value of the linear function of time over the whole sample.

Sometimes it is appropriate to use constructed variables that are more complicated than a linear time trend. A simple case would be a quadratic time trend, with typical element t^2. Typically, if a quadratic time trend were included, a linear time trend would be as well. In fact, any deterministic function of the time index t can be used, including the trigonometric functions $\sin t$ and $\cos t$, which could be used to account for oscillatory behavior. With such variables, it is again usually preferable to make them orthogonal to the constant by centering them.

The FWL Theorem applies just as well with time trends of various sorts as it does with seasonal dummy variables. It is possible to project all the other variables in a regression model off the time trend variables, thereby obtaining **detrended** variables. The parameter estimates and residuals are the same as if the trend variables were explicitly included in the regression. This was in fact the type of situation dealt with by Frisch and Waugh (1933).

Fixed Effects

When the observations in a sample fall naturally into a number of distinct groups, often based on location or time period, it can be convenient to use

Copyright © 2021, Russell Davidson and James G. MacKinnon

two (or perhaps more) subscripts to identify each observation instead of the single "t" subscript that we have been using so far. For example, we might let x_{gi} denote the i^{th} observation on a variable x that belongs to group g. Then, if there are G groups and the g^{th} group has n_g observations, the regression model (3.01) could be written as

$$y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\beta} + u_{gi}, \quad g = 1, \dots, G, \quad i = 1, \dots, n_g, \tag{3.53}$$

where y_{gi} and u_{gi} are scalars, and $\boldsymbol{X}_{gi}$ is a row vector of length k. The total number of observations n is $\sum_{g=1}^{G} n_g$. There is always a mapping from the values of g and i to the value of t, which will depend on just how the observations are ordered. The simplest way to order them is first by g and then by i. This implies that

$$t = \sum_{j=1}^{g-1} n_j + i, \quad i = 1, \dots, n_g. \tag{3.54}$$

In the remainder of this section, we will assume that this is indeed how the observations are ordered, because doing so simplifies a number of things.

There are two situations in which it may be useful to rewrite the regression (3.01) in the form of (3.53). The first is when the properties of the disturbances depend on them having two subscripts. The second is when some of the regressors are dummy variables that explicitly depend on the g and i subscripts. Of course, these two situations are not mutually exclusive. The former case will be discussed in Section 6.4, and the latter will be discussed here.

In many cases, it is plausible that the constant term should differ across each of the G groups. This is exactly what is implicitly assumed when seasonal dummy variables are used; the groups being the seasons. With different constant terms for each group, the model (3.53) is called the **fixed-effects model**. It can be written as

$$y_{gi} = \boldsymbol{X}_{gi}\boldsymbol{\beta} + \eta_g + u_{gi}, \quad g = 1, \dots, G, \quad i = 1, \dots, n_g, \tag{3.55}$$

where the η_g are scalars that are called **fixed effects**. The η_g have to be estimated.

In a way entirely analogous to (3.51), the fixed-effects regression (3.55) can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\eta} + \boldsymbol{u}, \tag{3.56}$$

where the n–vectors $\boldsymbol{y}$ and $\boldsymbol{u}$ have typical elements y_{gi} and u_{gi}, respectively, the $n \times k$ matrix $\boldsymbol{X}$ has typical row $\boldsymbol{X}_{gi}$, and the $n \times G$ matrix $\boldsymbol{D}$ contains G dummy variables. Each column of $\boldsymbol{D}$ corresponds to one of the fixed effects.

Copyright © 2021, Russell Davidson and James G. MacKinnon

For example, the third column has 1 in the n_3 positions for which $g = 3$ and 0 everywhere else. As the notation suggests, the G–vector of coefficients $\boldsymbol{\eta}$ has typical element η_g.

As in the case of seasonal dummies, the constant vector $\boldsymbol{\iota}$ is a linear combination of the columns of $\boldsymbol{D}$. Consequently, in order to ensure that the matrix of regressors $[\boldsymbol{X} \ \boldsymbol{D}]$ has full rank, the matrix $\boldsymbol{X}$ must not contain either a constant or any group of variables that collectively add up to a constant vector.

The fixed-effects regression model (3.56) can, of course, be estimated using any routine for OLS estimation. However, when both n and G are large, OLS estimation can be computationally demanding. In such a case, the FWL Theorem can be used to make computing the OLS estimator $\hat{\boldsymbol{\beta}}$ very much faster. Let $\boldsymbol{M}_{\boldsymbol{D}}$ denote the projection matrix $\mathbf{I} - \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top$. Then, by the FWL Theorem, we see that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{y}. \tag{3.57}$$

Expression (3.57) is valuable for computation because calculating the vector of residuals $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{x}$ for any vector $\boldsymbol{x}$ is extremely easy. The matrix $\boldsymbol{D}^\top\boldsymbol{D}$ is a $G \times G$ diagonal matrix with typical diagonal element n_g, and the vector $\boldsymbol{D}^\top\boldsymbol{x}$ is a G–vector with $\sum_{i=1}^{n_1} x_{1i}$ in position 1, $\sum_{i=1}^{n_1} x_{2i}$ in position 2, and so on. Therefore,

$$(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_G \end{bmatrix}, \quad \text{and} \quad \boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{x} = \boldsymbol{x} - \begin{bmatrix} \boldsymbol{\iota}_{n_1}\bar{x}_1 \\ \boldsymbol{\iota}_{n_2}\bar{x}_2 \\ \vdots \\ \boldsymbol{\iota}_{n_G}\bar{x}_G \end{bmatrix}, \tag{3.58}$$

where $\bar{x}_g$ denotes the sample mean of the elements of $\boldsymbol{x}$ that correspond to group g, and $\boldsymbol{\iota}_{n_g}$ denotes an n_g–vector of 1s, for $g = 1, \dots, G$. Thus each element of the vector $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{x}$ is simply the deviation of x_{gi} from its group mean $\bar{x}_g$.

Even when both n and G are extremely large, it is inexpensive to compute $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{y}$ and $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{X}$, because the former, and every column of the latter, is just a vector of deviations from group means. Computing $\hat{\boldsymbol{\beta}}$ then simply requires regressing $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{y}$ on $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{X}$. This FWL regression has k regressors. In contrast, computing $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$ jointly would require least-squares estimation of regression (3.56), which has $k + G$ regressors.

Since the FWL regression does not directly calculate $\hat{\boldsymbol{\eta}}$, we need to perform a few additional computations if the estimated fixed effects are of interest. Replacing $\boldsymbol{\beta}$ and $\boldsymbol{u}$ in regression (3.56) by their estimates from the FWL regression and rearranging yields the equation

$$\boldsymbol{D}\hat{\boldsymbol{\eta}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{u}}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Premultiplying both sides of this equation by $\boldsymbol{D}^\top$, we obtain

$$\boldsymbol{D}^\top\boldsymbol{D}\hat{\boldsymbol{\eta}} = \boldsymbol{D}^\top\boldsymbol{y} - \boldsymbol{D}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{D}^\top\hat{\boldsymbol{u}} = \boldsymbol{D}^\top\boldsymbol{y} - \boldsymbol{D}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{3.59}$$

The second equality here holds because the residual vector $\hat{\boldsymbol{u}}$ is orthogonal to each of the regressors in $\boldsymbol{D}$. This fact implies that the residuals sum to zero over each of the G groups. Solving equations (3.59) yields the result that

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{3.60}$$

This is just the vector of OLS estimates from a regression of $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ on $\boldsymbol{D}$. However, we do not actually have to run this regression in order to compute $\hat{\boldsymbol{\eta}}$ using expression (3.60). By the same arguments that led to the first equality in (3.58), we see that

$$\hat{\boldsymbol{\eta}} = \begin{bmatrix} \bar{y}_1 - \bar{\boldsymbol{X}}_1\hat{\boldsymbol{\beta}} \\ \bar{y}_2 - \bar{\boldsymbol{X}}_2\hat{\boldsymbol{\beta}} \\ \vdots \\ \bar{y}_G - \bar{\boldsymbol{X}}_G\hat{\boldsymbol{\beta}} \end{bmatrix}.$$

Thus, for all G, the estimated fixed effect $\hat{\eta}_g$ is simply the sample mean of $y_{gi} - \boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}}$ over the observations that belong to group g.

3.6 Influential Observations and Leverage

One important feature of OLS estimation, which we have not stressed up to this point, is that each element of the vector of parameter estimates $\hat{\boldsymbol{\beta}}$ is simply a weighted average of the elements of the vector $\boldsymbol{y}$. To see this, define $\boldsymbol{c}_i$ as the i^{th} row of the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, and observe from equation (3.02) that $\hat{\beta}_i = \boldsymbol{c}_i\boldsymbol{y}$. This fact will prove to be of great importance when we discuss the statistical properties of least-squares estimation in the next chapter.

Because each element of $\hat{\boldsymbol{\beta}}$ is a weighted average, some observations may affect the value of $\hat{\boldsymbol{\beta}}$ much more than others do. Consider Figure 3.14. This figure is an example of a **scatter diagram**, a long-established way of graphing the relation between two variables. Each point in the figure has Cartesian coordinates (x_t, y_t), where x_t and y_t are typical elements of a vector $\boldsymbol{x}$ and a vector $\boldsymbol{y}$, respectively. In the figure, there are 99 small dots and two larger ones. Suppose that we run the regression

$$\boldsymbol{y} = \beta_1\boldsymbol{\iota} + \beta_2\boldsymbol{x} + \boldsymbol{u}$$

using only the 99 observations represented by small dots. The fitted values from this regression all lie on the so-called **regression line**, which is the straight line with equation

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

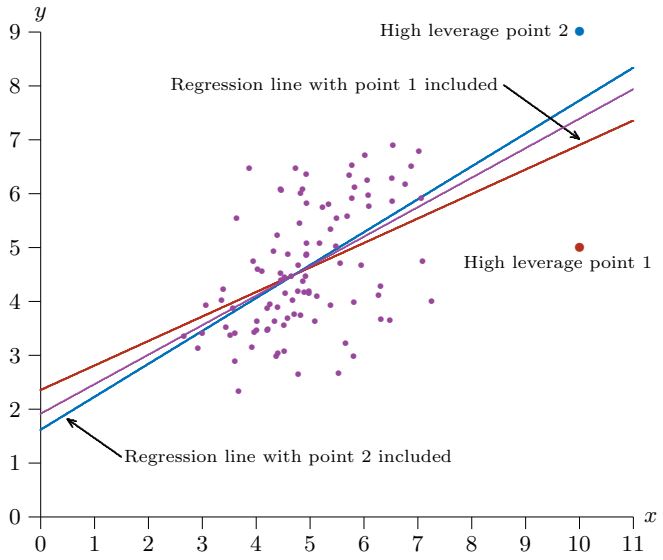


Figure 3.14 An observation with high leverage

The slope of this line is just $\hat{\beta}_2$, which is why β_2 is sometimes called the **slope coefficient**; see Section 2.1. Similarly, because $\hat{\beta}_1$ is the intercept that the regression line makes with the y axis, the constant term β_1 is sometimes called the **intercept**. The regression line is entirely determined by the estimated coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$.

Figure 3.14 actually shows three regression lines. Two of these are obtained by adding just one more observation to the sample. When the observation labeled "High leverage point 1" is added, the line becomes flatter. When the observation labeled "High leverage point 2" is added, the line becomes steeper. Both these points are well to the right of the other observations. They therefore exert a good deal of **leverage** on the regression line, pulling it towards themselves.

The two high-leverage points in Figure 3.14 have the same x coordinate. We could imagine locating the high-leverage point anywhere between the two of them. If the point were quite close to the original regression line, then the line would not move much when the high-leverage point was added to the sample. The further away from the line it is, the more the line will move when the point is added. Thus the **influence** of the high-leverage point becomes greater as its y coordinate becomes further away from the point on the original regression line associated with it x coordinate.

The extra points in Figure 3.14 have high leverage because their x coordinate is much larger than that of any other point in the sample. If the x coor-

Copyright © 2021, Russell Davidson and James G. MacKinnon

dinate were smaller, moving them closer to the main cloud of points, then they would have much smaller influence. Thus it is the x coordinate that potentially gives an observation high leverage, but it is the y coordinate that determines whether the high leverage is actually exploited, resulting in substantial influence on the regression line. In a moment, we will generalize these conclusions to regressions with any number of regressors.

If one or a few observations in a regression are highly **influential**, in the sense that deleting them from the sample would change some elements of $\hat{\boldsymbol{\beta}}$ substantially, the prudent econometrician will normally want to scrutinize the data carefully. It may be that these **influential observations** are erroneous, or at least untypical of the rest of the sample. Because a single erroneous observation can have an enormous effect on $\hat{\boldsymbol{\beta}}$, it is important to ensure that any influential observations are not in error. Even if the data are all correct, the interpretation of the regression results may change if it is known that a few observations are primarily responsible for those results, especially if those observations differ systematically in some way from the rest of the data.

Leverage

The effect of a single observation on $\hat{\boldsymbol{\beta}}$ can be seen by comparing $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}^{(t)}$, the estimate of $\boldsymbol{\beta}$ that would be obtained if the t^{th} observation were omitted from the sample. To see the effect of omitting the t^{th} observation, we can "remove" it by using a dummy variable. The appropriate dummy variable is $\boldsymbol{e}_t$, an n–vector which has t^{th} element 1 and all other elements 0. The vector $\boldsymbol{e}_t$ is called a **unit basis vector**, unit because its norm is 1, and basis because the set of all the $\boldsymbol{e}_t$, for $t = 1, \dots, n$, span, or constitute a **basis** for, the full space E^n; see Exercise 3.22. Considered as an indicator variable, $\boldsymbol{e}_t$ indexes the singleton subsample that contains only observation t.

Including $\boldsymbol{e}_t$ as a regressor leads to a regression of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \alpha \boldsymbol{e}_t + \boldsymbol{u}, \tag{3.61}$$

and, by the FWL Theorem, this gives the same parameter estimates and residuals as the FWL regression

$$\boldsymbol{M}_t \boldsymbol{y} = \boldsymbol{M}_t \boldsymbol{X}\boldsymbol{\beta} + \text{residuals}, \tag{3.62}$$

where $\boldsymbol{M}_t \equiv \boldsymbol{M}_{\boldsymbol{e}_t} = \mathbf{I} - \boldsymbol{e}_t(\boldsymbol{e}_t^\top \boldsymbol{e}_t)^{-1}\boldsymbol{e}_t^\top$ is the orthogonal projection off the vector $\boldsymbol{e}_t$. It is easy to see that $\boldsymbol{M}_t \boldsymbol{y}$ is just $\boldsymbol{y}$ with its t^{th} component replaced by 0. Since $\boldsymbol{e}_t^\top \boldsymbol{e}_t = 1$, and because $\boldsymbol{e}_t^\top \boldsymbol{y}$ is just the t^{th} component of $\boldsymbol{y}$,

$$\boldsymbol{M}_t \boldsymbol{y} = \boldsymbol{y} - \boldsymbol{e}_t \boldsymbol{e}_t^\top \boldsymbol{y} = \boldsymbol{y} - y_t \boldsymbol{e}_t.$$

Thus y_t is subtracted from $\boldsymbol{y}$ for the t^{th} observation only. Similarly, $\boldsymbol{M}_t \boldsymbol{X}$ is just $\boldsymbol{X}$ with its t^{th} row replaced by 0s. Running regression (3.62) gives the

Copyright © 2021, Russell Davidson and James G. MacKinnon

same parameter estimates as those that would be obtained if observation t were deleted from the sample. Since the vector $\hat{\boldsymbol{\beta}}$ is defined exclusively in terms of scalar products of the variables, replacing the t^{th} elements of these variables by 0 is tantamount to leaving observation t out when computing those scalar products.

Let us denote by $\boldsymbol{P}_{\boldsymbol{Z}}$ and $\boldsymbol{M}_{\boldsymbol{Z}}$, respectively, the orthogonal projections on to and off $\mathcal{S}(\boldsymbol{X}, \boldsymbol{e}_t)$. The fitted values and residuals from regression (3.61) are then given by

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y} + \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\boldsymbol{e}_t + \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}. \tag{3.63}$$

Now premultiply (3.63) by $\boldsymbol{P}_{\boldsymbol{X}}$ to obtain

$$\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)} + \hat{\alpha}\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_t, \tag{3.64}$$

where we have used the fact that $\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{P}_{\boldsymbol{X}} = \mathbf{O}$, because $\boldsymbol{M}_{\boldsymbol{Z}}$ annihilates both $\boldsymbol{X}$ and $\boldsymbol{e}_t$. But $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and so (3.64) gives

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}}) = -\hat{\alpha}\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_t. \tag{3.65}$$

We can compute the difference between $\hat{\boldsymbol{\beta}}^{(t)}$ and $\hat{\boldsymbol{\beta}}$ using this equation if we can compute the value of $\hat{\alpha}$.

In order to calculate $\hat{\alpha}$, we once again use the FWL Theorem, which tells us that the estimate of α from (3.61) is the same as the estimate from the FWL regression

$$\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y} = \hat{\alpha}\, \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{e}_t + \text{residuals}.$$

Therefore, using (3.02) and the idempotency of $\boldsymbol{M}_{\boldsymbol{X}}$,

$$\hat{\alpha} = \frac{\boldsymbol{e}_t^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}}{\boldsymbol{e}_t^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{e}_t}. \tag{3.66}$$

Now $\boldsymbol{e}_t^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$ is the t^{th} element of $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$, the vector of residuals from the regression including all observations. We may denote this element as $\hat{u}_t$. In like manner, $\boldsymbol{e}_t^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{e}_t$, which is just a scalar, is the t^{th} diagonal element of $\boldsymbol{M}_{\boldsymbol{X}}$. Substituting these into (3.66), we obtain

$$\hat{\alpha} = \frac{\hat{u}_t}{1 - h_t}, \tag{3.67}$$

where h_t denotes the t^{th} diagonal element of $\boldsymbol{P}_{\boldsymbol{X}}$, which is equal to 1 minus the t^{th} diagonal element of $\boldsymbol{M}_{\boldsymbol{X}}$. The rather odd notation h_t comes from the fact that $\boldsymbol{P}_{\boldsymbol{X}}$ is sometimes referred to as the **hat matrix**, because the vector of fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y}$ is sometimes written as $\hat{\boldsymbol{y}}$, and $\boldsymbol{P}_{\boldsymbol{X}}$ is therefore said to "put a hat on" $\boldsymbol{y}$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Finally, if we premultiply (3.65) by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and use (3.67), we find that

$$\hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}} = -\hat{\alpha}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t = \frac{-1}{1-h_t}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_t^\top\hat{u}_t. \tag{3.68}$$

The second equality uses the facts that $\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}^\top$ and that the final factor of $\boldsymbol{e}_t$ selects the t^{th} column of $\boldsymbol{X}^\top$, which is the transpose of the t^{th} row, $\boldsymbol{X}_t$. Expression (3.68) makes it clear that, when either $\hat{u}_t$ is large or h_t is large, or both, the effect of the t^{th} observation on at least some elements of $\hat{\boldsymbol{\beta}}$ is likely to be substantial. Such an observation is said to be influential.

From the rightmost expression in (3.68), it is evident that the influence of an observation depends on both $\hat{u}_t$ and h_t. It is greater if the observation has a large residual, which, as we saw in Figure 3.14, is related to its y coordinate. On the other hand, h_t is related to the x coordinate of a point, which, as we also saw in the figure, determines the leverage, or potential influence, of the corresponding observation. We say that observations for which h_t is large have **high leverage** or are **leverage points**. A leverage point is not necessarily influential, but it has the potential to be influential.

The Diagonal Elements of the Hat Matrix

Since the leverage of the t^{th} observation depends on h_t, the t^{th} diagonal element of the hat matrix, it is worth studying the properties of these diagonal elements in a little more detail. We can express h_t as

$$h_t = \boldsymbol{e}_t^\top\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t = \|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t\|^2. \tag{3.69}$$

Since the rightmost expression here is a square, $h_t \geq 0$. Moreover, because $\|\boldsymbol{e}_t\| = 1$, we obtain by applying the inequality (3.27) to the vector $\boldsymbol{e}_t$ that $h_t = \|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t\|^2 \leq 1$. Thus

$$0 \leq h_t \leq 1. \tag{3.70}$$

The geometrical reason for these bounds on the value of h_t can be found in Exercise 3.30.

The lower bound in (3.70) can be strengthened when there is a constant term. In that case, none of the h_t can be less than $1/n$. This follows from equation (3.69), because if $\boldsymbol{X}$ consisted only of a constant vector $\boldsymbol{\iota}$, $\boldsymbol{e}_t^\top\boldsymbol{P}_{\boldsymbol{\iota}}\boldsymbol{e}_t$ would equal $1/n$. If other regressors are present, then we have

$$h_t = \|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t\|^2 \geq \|\boldsymbol{P}_{\boldsymbol{\iota}}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{e}_t\|^2 = \|\boldsymbol{P}_{\boldsymbol{\iota}}\boldsymbol{e}_t\|^2 = 1/n.$$

Here we have used the fact that $\boldsymbol{P}_{\boldsymbol{\iota}}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{\iota}}$ since $\boldsymbol{\iota}$ is in $\mathcal{S}(\boldsymbol{X})$ by assumption, and, for the inequality, we have used (3.27). Although h_t cannot be 0 in normal circumstances, there is a special case in which it equals 1. If one column of $\boldsymbol{X}$ is the dummy variable $\boldsymbol{e}_t$, then $h_t = \boldsymbol{e}_t^\top\boldsymbol{P}_X\boldsymbol{e}_t = \boldsymbol{e}_t^\top\boldsymbol{e}_t = 1$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

In a regression with n observations and k regressors, the average of the h_t is equal to k/n. In order to demonstrate this, we need to use some properties of the **trace** of a square matrix. If $\boldsymbol{A}$ is an $n \times n$ matrix, its trace, denoted $\text{Tr}(\boldsymbol{A})$, is the sum of the elements on its principal diagonal. Thus

$$\text{Tr}(\boldsymbol{A}) \equiv \sum_{i=1}^{n} A_{ii}.$$

The trace of a product of two not necessarily square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is unaffected by the order in which the two matrices are multiplied together. If the dimensions of $\boldsymbol{A}$ are $n \times m$, then, in order for the product $\boldsymbol{AB}$ to be square, those of $\boldsymbol{B}$ must be $m \times n$. This implies further that the product $\boldsymbol{BA}$ exists and is $m \times m$. We have

$$\text{Tr}(\boldsymbol{AB}) = \sum_{i=1}^{n}(\boldsymbol{AB})_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}B_{ji} = \sum_{j=1}^{m}(\boldsymbol{BA})_{jj} = \text{Tr}(\boldsymbol{BA}). \tag{3.71}$$

The result (3.71) can be extended. If we consider a (square) product of several matrices, the trace is invariant under what is called a **cyclic permutation** of the factors. Thus, as can be seen by successive applications of (3.71),

$$\text{Tr}(\boldsymbol{ABC}) = \text{Tr}(\boldsymbol{CAB}) = \text{Tr}(\boldsymbol{BCA}). \tag{3.72}$$

We now return to the h_t. Their sum is

$$\begin{aligned}\sum_{t=1}^{n} h_t = \text{Tr}(\boldsymbol{P}_{\boldsymbol{X}}) &= \text{Tr}\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big) \\ &= \text{Tr}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\big) = \text{Tr}(\mathbf{I}_k) = k.\end{aligned} \tag{3.73}$$

The first equality in the second line makes use of (3.72). Then, because we are multiplying a $k \times k$ matrix by its inverse, we get a $k \times k$ identity matrix, the trace of which is obviously just k. It follows from (3.73) that the average of the h_t equals k/n. When, for a given regressor matrix $\boldsymbol{X}$, the diagonal elements of $\boldsymbol{P}_{\boldsymbol{X}}$ are all close to their average value, no observation has very much leverage. Such an $\boldsymbol{X}$ matrix is sometimes said to have a **balanced design**. On the other hand, if some of the h_t are much larger than k/n, and others consequently smaller, the $\boldsymbol{X}$ matrix is said to have an **unbalanced design**.

The h_t tend to be larger for values of the regressors that are farther away from their average over the sample. As an example, Figure 3.15 plots them as a function of x_t for a particular sample of 100 observations for the model

$$y_t = \beta_1 + \beta_2 x_t + u_t.$$

The elements x_t of the regressor are perfectly well behaved, being drawings from the standard normal distribution. Although the average value of the h_t

Copyright © 2021, Russell Davidson and James G. MacKinnon

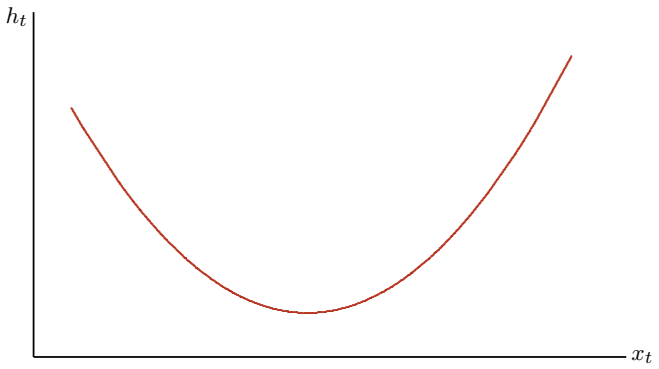


Figure 3.15 A graph of h_t as a function of x_t

is $2/100 = 0.02$, h_t varies from 0.0100 for values of x_t near the sample mean to 0.0695 for the largest value of x_t, which is about 2.4 standard deviations above the sample mean. Thus, even in this very typical case, some observations have a great deal more leverage than others. Those observations with the greatest amount of leverage are those for which x_t is farthest from the sample mean, in accordance with the intuition of Figure 3.14.

3.7 Final Remarks

In this chapter, we have discussed the numerical properties of OLS estimation of linear regression models from a geometrical point of view. This perspective often provides a much simpler way to understand such models than does a purely algebraic approach. For example, the fact that certain matrices are idempotent becomes quite clear as soon as one understands the notion of an orthogonal projection. Most of the results discussed in this chapter are thoroughly fundamental, and many of them will be used again and again throughout the book. In particular, the FWL Theorem will turn out to be extremely useful in many contexts.

The use of geometry as an aid to the understanding of linear regression has a long history; see Herr (1980). One valuable reference on linear models that takes the geometric approach is Seber (1980). A good expository paper that is reasonably accessible is Bryant (1984), and a detailed treatment is provided by Ruud (2000).

It is strongly recommended that readers attempt the exercises which follow this chapter before starting Chapter 4, in which we turn our attention to the statistical properties of OLS estimation. Many of the results of this chapter will be useful in establishing these properties, and the exercises are designed to enhance understanding of these results.

Copyright © 2021, Russell Davidson and James G. MacKinnon

3.8 Exercises

3.1 Consider two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in E^2. Let $\boldsymbol{x} = [x_1 \vdots x_2]$ and $\boldsymbol{y} = [y_1 \vdots y_2]$. Show trigonometrically that $\boldsymbol{x}^\top\boldsymbol{y} \equiv x_1y_1 + x_2y_2$ is equal to $\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|\cos\theta$, where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.

3.2 A vector in E^n can be **normalized** by multiplying it by the reciprocal of its norm. Show that, for any $\boldsymbol{x} \in E^n$ with $\boldsymbol{x} \neq \boldsymbol{0}$, the norm of $\boldsymbol{x}/\|\boldsymbol{x}\|$ is 1.

Now consider two vectors $\boldsymbol{x}, \boldsymbol{y} \in E^n$. Compute the norm of the sum and of the difference of $\boldsymbol{x}$ normalized and $\boldsymbol{y}$ normalized, that is, of

$$\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} + \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \quad \text{and} \quad \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}.$$

By using the fact that the norm of any nonzero vector is positive, prove the Cauchy-Schwartz inequality (3.08):

$$|\boldsymbol{x}^\top\boldsymbol{y}| \leq \|\boldsymbol{x}\|\,\|\boldsymbol{y}\|. \tag{3.08}$$

Show that this inequality becomes an equality when $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel. **Hint:** Show first that $\boldsymbol{x}$ and $\boldsymbol{y}$ are parallel if and only if $\boldsymbol{x}/\|\boldsymbol{x}\| = \pm\,\boldsymbol{y}/\|\boldsymbol{y}\|$.

3.3 The **triangle inequality** states that, for $\boldsymbol{x}, \boldsymbol{y} \in E^n$,

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|. \tag{3.74}$$

Draw a 2–dimensional picture to illustrate this result. Prove the result algebraically by computing the squares of both sides of the above inequality, and then using (3.08). In what circumstances does (3.74) hold with equality?

3.4 Suppose that $\boldsymbol{x} = [1.0 \vdots 1.5 \vdots 1.2 \vdots 0.7]$ and $\boldsymbol{y} = [3.2 \vdots 4.4 \vdots 2.5 \vdots 2.0]$. What are $\|\boldsymbol{x}\|$, $\|\boldsymbol{y}\|$, and $\boldsymbol{x}^\top\boldsymbol{y}$? Use these quantities to calculate θ, the angle θ between $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\cos\theta$.

3.5 Show explicitly that the left-hand sides of (3.11) and (3.12) are the same. This can be done either by comparing typical elements or by using the results in Section 3.3 on partitioned matrices.

3.6 Prove that, if the k columns of $\boldsymbol{X}$ are linearly independent, each vector $\boldsymbol{z}$ in $\mathcal{S}(\boldsymbol{X})$ can be expressed as $\boldsymbol{X}\boldsymbol{b}$ for one and only one k–vector $\boldsymbol{b}$. **Hint:** Suppose that there are two different vectors, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, such that $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{b}_i$, $i = 1, 2$, and show that this implies that the columns of $\boldsymbol{X}$ are linearly dependent.

3.7 Consider the vectors $\boldsymbol{x}_1 = [1 \vdots 2 \vdots 4]$, $\boldsymbol{x}_2 = [2 \vdots 3 \vdots 5]$, and $\boldsymbol{x}_3 = [3 \vdots 6 \vdots 12]$. What is the dimension of the subspace that these vectors span?

3.8 Consider the example of the three vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ defined in (3.14). Show that any vector $\boldsymbol{z} \equiv b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2$ in $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ also belongs to $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_3)$ and $\mathcal{S}(\boldsymbol{x}_2, \boldsymbol{x}_3)$. Give explicit formulas for $\boldsymbol{z}$ as a linear combination of $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$, and of $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$.

3.9 Prove algebraically that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{O}$. This is equation (3.25). Use only the requirement (3.24) that $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}$ are complementary projections, and the idempotency of $\boldsymbol{P}_{\boldsymbol{X}}$.

3.10 Let $\boldsymbol{X}$ and $\boldsymbol{W}$ be two $n \times k$ matrices such that $\mathcal{S}(\boldsymbol{X}) \neq \mathcal{S}(\boldsymbol{W})$. Show that the $n \times n$ matrix $\boldsymbol{P} \equiv \boldsymbol{X}(\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{W}^\top$ is idempotent but not symmetric.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Characterize the spaces that $\boldsymbol{P}$ and $\mathbf{I} - \boldsymbol{P}$ project on to, and show that they are not orthogonal. Projections like $\boldsymbol{P}$ are called **oblique projections**.

3.11 Prove algebraically that equation (3.26), which is really Pythagoras' Theorem for linear regression, holds. Use the facts that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric, idempotent, and orthogonal to each other.

3.12 Show algebraically that, if $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are complementary orthogonal projections, then $\boldsymbol{M_X}$ annihilates all vectors in $\mathcal{S}(\boldsymbol{X})$, and $\boldsymbol{P_X}$ annihilates all vectors in $\mathcal{S}^{\perp}(\boldsymbol{X})$.

3.13 Consider the two regressions

$$\boldsymbol{y} = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \boldsymbol{u}, \text{ and}$$
$$\boldsymbol{y} = \alpha_1 \boldsymbol{z}_1 + \alpha_2 \boldsymbol{z}_2 + \alpha_3 \boldsymbol{z}_3 + \boldsymbol{u},$$

where $\boldsymbol{z}_1 = \boldsymbol{x}_1 - 2\boldsymbol{x}_2$, $\boldsymbol{z}_2 = \boldsymbol{x}_2 + 4\boldsymbol{x}_3$, and $\boldsymbol{z}_3 = 2\boldsymbol{x}_1 - 3\boldsymbol{x}_2 + 5\boldsymbol{x}_3$. Let $\boldsymbol{X} = [\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \boldsymbol{x}_3]$ and $\boldsymbol{Z} = [\boldsymbol{z}_1 \;\; \boldsymbol{z}_2 \;\; \boldsymbol{z}_3]$. Show that the columns of $\boldsymbol{Z}$ can be expressed as linear combinations of the columns of $\boldsymbol{X}$, that is, that $\boldsymbol{Z} = \boldsymbol{XA}$, for some 3×3 matrix $\boldsymbol{A}$. Find the elements of this matrix $\boldsymbol{A}$.

Show that the matrix $\boldsymbol{A}$ is invertible, by showing that the columns of $\boldsymbol{X}$ are linear combinations of the columns of $\boldsymbol{Z}$. Give the elements of $\boldsymbol{A}^{-1}$. Show that the two regressions give the same fitted values and residuals.

Precisely how is the OLS estimate $\hat{\beta}_1$ related to the OLS estimates $\hat{\alpha}_i$, for $i = 1, \ldots, 3$? Precisely how is $\hat{\alpha}_1$ related to the $\hat{\beta}_i$, for $i = 1, \ldots, 3$?

3.14 Let $\boldsymbol{X}$ be an $n \times k$ matrix of full rank. Consider the $n \times k$ matrix $\boldsymbol{XA}$, where $\boldsymbol{A}$ is a *singular* $k \times k$ matrix. Show that the columns of $\boldsymbol{XA}$ are linearly dependent, and that $\mathcal{S}(\boldsymbol{XA}) \subset \mathcal{S}(\boldsymbol{X})$.

3.15 Use the result (3.35) to show that $\boldsymbol{M_X M}_1 = \boldsymbol{M}_1 \boldsymbol{M_X} = \boldsymbol{M_X}$, where $\boldsymbol{X}$ is partitioned as $[\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$.

3.16 Consider the following linear regression:

$$\boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u},$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}_1$ is $n \times k_1$, and $\boldsymbol{X}_2$ is $n \times k_2$. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the OLS parameter estimates from running this regression.

Now consider the following regressions, all to be estimated by OLS:

(a) $\boldsymbol{y} = \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(b) $\boldsymbol{P}_1 \boldsymbol{y} = \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(c) $\boldsymbol{P}_1 \boldsymbol{y} = \boldsymbol{P}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(d) $\boldsymbol{P_X y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(e) $\boldsymbol{P_X y} = \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(f) $\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(g) $\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(h) $\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(i) $\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{M}_1 \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$;
(j) $\boldsymbol{P_X y} = \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Here $\boldsymbol{P}_1$ projects orthogonally on to the span of $\boldsymbol{X}_1$, and $\boldsymbol{M}_1 = \mathbf{I} - \boldsymbol{P}_1$. For which of the above regressions are the estimates of $\boldsymbol{\beta}_2$ the same as for the original regression? Why? For which are the residuals the same? Why?

3.17 Consider the linear regression

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u},$$

where $\boldsymbol{\iota}$ is an n–vector of 1s, and $\boldsymbol{X}_2$ is an $n \times (k-1)$ matrix of observations on the remaining regressors. Show, using the FWL Theorem, that the OLS estimators of β_1 and $\boldsymbol{\beta}_2$ can be written as

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} n & \boldsymbol{\iota}^\top \boldsymbol{X}_2 \\ \mathbf{0} & \boldsymbol{X}_2^\top \boldsymbol{M}_\iota \boldsymbol{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\iota}^\top \boldsymbol{y} \\ \boldsymbol{X}_2^\top \boldsymbol{M}_\iota \boldsymbol{y} \end{bmatrix},$$

where, as usual, $\boldsymbol{M}_\iota$ is the matrix that takes deviations from the sample mean.

3.18 Using equations (3.35), show that $\boldsymbol{P_X} - \boldsymbol{P}_1$ is an orthogonal projection matrix. That is, show that $\boldsymbol{P_X} - \boldsymbol{P}_1$ is symmetric and idempotent.

⋆**3.19** Show that $\boldsymbol{P_X} - \boldsymbol{P}_1 = \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2}$, where $\boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2}$ is the projection on to the span of $\boldsymbol{M}_1 \boldsymbol{X}_2$. This can be done most easily by showing that any vector in $\mathcal{S}(\boldsymbol{M}_1 \boldsymbol{X}_2)$ is invariant under the action of $\boldsymbol{P_X} - \boldsymbol{P}_1$, and that any vector orthogonal to this span is annihilated by $\boldsymbol{P_X} - \boldsymbol{P}_1$.

3.20 Let $\boldsymbol{\iota}$ be a vector of 1s, and let $\boldsymbol{X}$ be an $n \times 3$ matrix, with full rank, of which the first column is $\boldsymbol{\iota}$. What can you say about the matrix $\boldsymbol{M}_\iota \boldsymbol{X}$? What can you say about the matrix $\boldsymbol{P}_\iota \boldsymbol{X}$? What is $\boldsymbol{M}_\iota \boldsymbol{M_X}$ equal to? What is $\boldsymbol{P}_\iota \boldsymbol{M_X}$ equal to?

3.21 Express the four seasonal variables, $\boldsymbol{s}_i$, $i = 1, 2, 3, 4$, defined in (3.46), as functions of the constant $\boldsymbol{\iota}$ and the three variables $\boldsymbol{s}'_i$, $i = 1, 2, 3$, defined in (3.49).

3.22 Show that the full n–dimensional space E^n is the span of the set of **unit basis vectors** $\boldsymbol{e}_t$, $t = 1, \ldots, n$, where all the components of $\boldsymbol{e}_t$ are zero except for the t^{th}, which is equal to 1.

3.23 The file **earnings-data.txt** contains data on the weekly earnings of 46,302 women who lived in California between 1992 and 2015. Regress the log of earnings on age, age squared, the five education dummies, and as many of the 24 year dummies as you can. How many regressors does your model contain? What happens if you add a constant term?

Regress each of the log of earnings, age, age squared, and the five education dummies on the year dummies that you included in your model. Then regress the residuals from the first of these regressions on the residuals from the other seven regressions. Do the estimates look familiar? Explain.

3.24 Verify numerically that regressing the log earnings variable on the full set of education dummy variables, without a constant term, and taking the fitted values is equivalent to replacing each observation by the group mean of log earnings for women with that level of education.

How is the regression you just ran related to regressing the log earnings variable on a constant term and the last four education dummies, `ed2` through `ed5`? How do you interpret the coefficients on the dummy variables in this regression?

Copyright © 2021, Russell Davidson and James G. MacKinnon

3.25 The file **tbrate-data.txt** contains data for three quarterly time series for the United States: r_t, the interest rate on 90-day treasury bills, π_t, the rate of inflation, and dy_t, the quarterly percentage change in seasonally adjusted real GDP at annual rates. For the period 1955:1 to 2014:4, run the regression

$$\Delta r_t = \beta_1 + \beta_2 dy_t + \beta_3 dy_{t-1} + \beta_4 \pi_t + \beta_5 r_{t-1} + u_t, \tag{3.75}$$

where Δ is the **first-difference operator**, defined so that $\Delta x_t = x_t - x_{t-1}$. Plot the residuals and fitted values against time. Then regress the residuals on the fitted values and on a constant. What do you learn from this second regression? Now regress the fitted values on the residuals and on a constant. What do you learn from this third regression?

3.26 For the same sample period, regress Δr_t on a constant, dy_t, dy_{t-1}, and r_{t-1}. Save the residuals from this regression, and call them $\hat{e}_t$. Then regress π_t on a constant, dy_t, dy_{t-1}, and r_{t-1}. Save the residuals from this regression, and call them $\hat{v}_t$. Now regress $\hat{e}_t$ on $\hat{v}_t$. How are the estimated coefficient and the residuals from this last regression related to anything that you obtained when you estimated regression (3.75)?

3.27 Calculate the diagonal elements of the hat matrix for regression (3.75) and use them to calculate a measure of leverage. Plot this measure against time. On the basis of this plot, which observations seem to have unusually high leverage?

3.28 Show explicitly that the t^{th} residual from running regression (3.61) is 0.

3.29 Calculate a vector of "omit 1" residuals $\hat{\boldsymbol{u}}^{(\cdot)}$ for regression (3.75). The t^{th} element of $\hat{\boldsymbol{u}}^{(\cdot)}$ is the residual for the t^{th} observation calculated from a regression that uses data for every observation except the t^{th}. Try to avoid running 240 regressions in order to do this! Regress $\hat{\boldsymbol{u}}^{(\cdot)}$ on the ordinary residuals $\hat{\boldsymbol{u}}$. Is the estimated coefficient roughly the size you expected it to be? Would it be larger or smaller if you were to omit some of the high-leverage observations?

3.30 Show that the leverage measure h_t is the square of the cosine of the angle between the unit basis vector $\boldsymbol{e}_t$ and its projection on to the span $\mathcal{S}(\boldsymbol{X})$ of the regressors.

3.31 Suppose the matrix $\boldsymbol{X}$ is 150×5 and has full rank. Let $\boldsymbol{P}_{\boldsymbol{X}}$ be the matrix that projects on to $\mathcal{S}(\boldsymbol{X})$ and let $\boldsymbol{M}_{\boldsymbol{X}} = \mathbf{I} - \boldsymbol{P}_{\boldsymbol{X}}$. What is $\text{Tr}(\boldsymbol{P}_{\boldsymbol{X}})$? What is $\text{Tr}(\boldsymbol{M}_{\boldsymbol{X}})$? What would these be if $\boldsymbol{X}$ did not have full rank but instead had rank 3?

3.32 Generate a figure like Figure 3.15 for yourself. Begin by drawing 100 observations of a regressor x_t from the $\text{N}(0, 1)$ distribution. Then compute and save the h_t for a regression of any regressand on a constant and x_t. Plot the points (x_t, h_t), and you should obtain a graph similar to the one in Figure 3.15.

Now add one more observation, x_{101}. Start with $x_{101} = \bar{x}$, the average value of the x_t, and then increase x_{101} progressively until $x_{101} = \bar{x} + 20$. For each value of x_{101}, compute the leverage measure h_{101}. How does h_{101} change as x_{101} gets larger? Why is this in accord with the result that $h_t = 1$ if the regressors include the dummy variable $\boldsymbol{e}_t$?

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 4

The Statistical Properties of Ordinary Least Squares

4.1 Introduction

In the previous chapter, we studied the numerical properties of ordinary least squares estimation, properties that hold no matter how the data may have been generated. In this chapter, we turn our attention to the **statistical** properties of OLS, ones that depend on how the data were actually generated. These properties can never be shown to hold numerically for any actual data set, but they can be proved to hold if we are willing to make certain assumptions. Most of the properties that we will focus on concern the first two moments of the least-squares estimator.

In Section 2.5, we introduced the concept of a **data-generating process**, or **DGP**. For any data set that we are trying to analyze, the DGP is simply the mechanism that actually generated the data. Most real DGPs for economic data are probably very complicated, and economists do not pretend to understand every detail of them. However, for the purpose of studying the statistical properties of estimators, it is almost always necessary to assume that the DGP is quite simple. For instance, when we are studying the (multiple) linear regression model

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{4.01}$$

we may wish to assume that the data were actually generated by the DGP

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta}_0 + u_t, \quad u_t \sim \text{NID}(0, \sigma_0^2). \tag{4.02}$$

The symbol "$\sim$" in (4.01) and (4.02) means "is distributed as." We introduced the abbreviation IID, which means "independently and identically distributed," in Section 2.3. In the model (4.01), the notation $\text{IID}(0, \sigma^2)$ means that the u_t are statistically independent and all have the same distribution, with expectation 0 and variance σ^2. Similarly, in the DGP (4.02), the notation $\text{NID}(0, \sigma_0^2)$ means that the u_t are *normally*, independently, and identically distributed, with expectation 0 and variance σ_0^2. In both cases, it is implicitly being assumed that the distribution of u_t is in no way dependent on $\boldsymbol{X}_t$.

The differences between the regression model (4.01) and the DGP (4.02) may seem subtle, but they are important. A key feature of a DGP is that it constitutes a **complete specification**, where that expression means, as in Section 2.3, that enough information is provided for the DGP to be simulated on a computer. For that reason, in (4.02) we must provide specific values for the parameters $\boldsymbol{\beta}$ and σ^2 (the zero subscripts on these parameters are intended to remind us of this), and we must specify from what distribution the disturbances are to be drawn (here, the normal distribution).

A **model** is defined as a set of data-generating processes. Since a model is a set, we will sometimes use the notation $\mathbb{M}$ to denote it. In the case of the **linear regression model** (4.01), this set consists of all DGPs of the form (4.01) in which the coefficient vector $\boldsymbol{\beta}$ takes some value in $\mathbb{R}^k$, the variance σ^2 is some positive real number, and the distribution of u_t varies over all possible distributions that have expectation 0 and variance σ^2. Although the DGP (4.02) evidently belongs to this set, it is considerably more restrictive.

A subset of the set of DGPs of the form (4.02) defines what is called the **classical normal linear model**, where the name indicates that the disturbances are normally distributed. The subset results from the imposition of the restriction of **exogeneity** on the regressors in the matrix $\boldsymbol{X}$; see the definition of this concept in the next section. The model (4.01) is larger than the classical normal linear model, not only because it does not require exogeneity of the regressors, but also because, although (4.01) specifies the first two moments of the disturbances, and requires them to be mutually independent, it says no more about them, and in particular it does not require them to be normal. All of the results we prove in this chapter, and many of those in the next, apply to the linear regression model (4.01), with no normality assumption. However, in order to obtain some of the results in the next two chapters, it will be necessary to limit attention to the classical normal linear model.

For most of this chapter, we assume that whatever model we are studying, the linear regression model or the classical normal linear model, is **correctly specified**. By this, we mean that the DGP that actually generated our data belongs to the model under study. A model is **misspecified** if that is not the case. It is crucially important, when studying the properties of an estimation procedure, to distinguish between properties which hold only when the model is correctly specified, and properties, like those treated in the previous chapter, which hold no matter what the DGP. We can talk about statistical properties only if we specify the DGP.

In the remainder of this chapter, we study a number of the most important statistical properties of ordinary least-squares estimation, by which we mean least-squares estimation of linear regression models. In Section 4.2, we discuss the concept of bias and prove that $\hat{\boldsymbol{\beta}}$, the OLS estimator of $\boldsymbol{\beta}$, is unbiased under certain conditions. Then, in Section 4.3, we discuss the concepts of asymptotic constructions and consistency, and prove that $\hat{\boldsymbol{\beta}}$ is consistent under considerably weaker conditions. In Section 4.4, we turn our attention to

Copyright © 2021, Russell Davidson and James G. MacKinnon

the covariance matrix of $\hat{\boldsymbol{\beta}}$. In Section 4.5, we discuss what determines the precision of OLS estimates and introduce the concept of collinearity. This leads to a discussion of the efficiency of least-squares estimation in Section 4.6, in which we prove the famous Gauss-Markov Theorem. In Section 4.7, we discuss the estimation of σ^2 and the relationship between disturbances and least-squares residuals. Up to this point, we will assume that the DGP belongs to the model being estimated. In Section 4.8, we relax this assumption and consider the consequences of estimating a model that is misspecified in certain ways. Finally, in Section 4.9, we discuss ways of measuring how well a regression fits, in particular, the measure called R^2.

4.2 Bias and Unbiasedness

One desirable statistical property of any estimator is for it to be **unbiased**. Suppose that $\hat{\theta}$ is an estimator of some parameter θ, the true value of which is θ_0 for some given DGP. The **estimation error** is the difference $\hat{\theta} - \theta_0$ between the estimator and the true value. Since the estimator is a random variable, so is the estimation error. The expectation of the estimation error is what we call the **bias**. It is defined to be $\mathrm{E}(\hat{\theta}) - \theta_0$, and it is defined for the given DGP.

Suppose now that the DGP belongs to a model $\mathbb{M}$. We can in principle compute the bias of the estimator $\hat{\theta}$ for every DGP $\mu \in \mathbb{M}$. If we denote the true value of θ for μ as θ_μ, then the bias for that DGP is $\mathrm{E}_\mu(\hat{\theta}) - \theta_\mu$, where we use the notation E_μ for the expectation when all random variables are generated by μ. If the bias of an estimator is zero for every $\mu \in \mathbb{M}$, then the estimator is said to be unbiased. Otherwise, it is said to be **biased**. Intuitively, if we were to use an unbiased estimator to calculate estimates for a very large number of samples, then the average value of those estimates would tend to the quantity being estimated. We would always prefer an unbiased estimator to a biased one if their other statistical properties were the same.

As we first saw in Section 2.4, the linear regression model (4.01) can also be written, using matrix notation, as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \tag{4.03}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are n–vectors, $\boldsymbol{X}$ is an $n \times k$ matrix, and $\boldsymbol{\beta}$ is a k–vector. In (4.03), the notation $\mathrm{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I})$ is the matrix version of that in (4.01): it is just another way of saying that each element of the vector $\boldsymbol{u}$ is independently and identically distributed with expectation 0 and variance σ^2. This notation is convenient to use when the model is written in matrix notation. Its meaning should become clearer in Section 4.4.

Recall from Section 2.5 that the OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{4.04}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Thus we have, for a DGP μ with true parameter vector $\boldsymbol{\beta}_\mu$,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}_\mu + \boldsymbol{u}) \\ &= \boldsymbol{\beta}_\mu + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}.\end{aligned} \tag{4.05}$$

The expectation of the second line here is

$$\mathrm{E}_\mu(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_\mu + \mathrm{E}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}\big). \tag{4.06}$$

Thus $\hat{\boldsymbol{\beta}}$ is unbiased if and only if the second term on the right-hand side of equation (4.06) is equal to a zero vector. What is not entirely obvious is just what assumptions are needed to ensure that this condition holds. We will discuss these assumptions in the next subsection.

Although it is desirable for an estimator to be unbiased, few estimators in econometrics actually have this property. A closely related concept is that of an **unbiased estimating equation**. Many estimators, including many biased ones, are based on unbiased estimating equations. As we will see in later chapters, such estimators often have good theoretical properties. Quite generally, the left-hand side of an estimating equation is a function of data and parameters, called the **estimating function**. The estimator of the model parameters is found by solving the estimating equation or equations, that is, by setting the estimating functions to zero. An estimating equation is unbiased precisely when the corresponding estimating function is a zero function; recall the definition in Section 2.5.

In general, an estimating equation takes the form

$$g(\boldsymbol{y}, \boldsymbol{\theta}) = 0, \tag{4.07}$$

where $\boldsymbol{\theta}$ is a vector of model parameters, $\boldsymbol{y}$ represents the data, and $g(\boldsymbol{y}, \boldsymbol{\theta})$ is the estimating function. We say that (4.07) is unbiased for a model $\mathbb{M}$ if, for every DGP $\mu \in \mathbb{M}$,

$$\mathrm{E}_\mu\, g(\boldsymbol{y}, \boldsymbol{\theta}_\mu) = \mathbf{0}.$$

As above, the notation E_μ here denotes the expectation when the data $\boldsymbol{y}$ are generated by the DGP μ, and $\boldsymbol{\theta}_\mu$ denotes the true value of $\boldsymbol{\theta}$ for that DGP. If $\boldsymbol{\theta}$ is a k-vector, we need k estimating equations in order to define it. The bias of the estimator $\hat{\boldsymbol{\theta}}$ defined implicitly by a set of estimating equations is defined as $\mathrm{E}_\mu(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\mu)$, and the estimator $\hat{\boldsymbol{\theta}}$ is unbiased if this bias is zero for all $\mu \in \mathbb{M}$.

The OLS estimator of the model (4.03) is defined by the estimating equations (2.47) that we first saw in Section 2.5:

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}. \tag{4.08}$$

Suppose that the DGP is given by (4.03) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, so that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$. Then the OLS estimating functions, the left-hand side of equations (4.08),

Copyright © 2021, Russell Davidson and James G. MacKinnon

become the components of the k-vector $\boldsymbol{X}^\top(\boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \boldsymbol{u})$. To see whether these estimating equations are unbiased, we evaluate this quantity at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. The result is just $\boldsymbol{X}^\top\boldsymbol{u}$. Therefore, the estimating equations (4.08) are unbiased whenever

$$\mathrm{E}(\boldsymbol{X}^\top\boldsymbol{u}) = \mathbf{0}. \tag{4.09}$$

Assumptions About Disturbances and Regressors

In certain cases, it may be reasonable to treat the matrix $\boldsymbol{X}$ as **nonstochastic**, or **fixed**. For example, this would certainly be a reasonable assumption to make if the data pertained to an experiment, and the experimenter had chosen the values of all the variables that enter into $\boldsymbol{X}$ before $\boldsymbol{y}$ was determined. In this case, we have $\mathrm{E}(\boldsymbol{X}^\top\boldsymbol{u}) = \boldsymbol{X}^\top\mathrm{E}(\boldsymbol{u})$, since we may move the nonstochastic factor $\boldsymbol{X}^\top$ outside the expectation operator. But then, by (4.09), the OLS estimating equations are unbiased, because part of the model specification is that $\mathrm{E}(\boldsymbol{u}) = \mathbf{0}$,

Similarly, in this case, the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is not random, and the second term in (4.06) becomes

$$\mathrm{E}\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}\big) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathrm{E}(\boldsymbol{u}). \tag{4.10}$$

Therefore, the OLS estimator itself is unbiased.

Unfortunately, the assumption that $\boldsymbol{X}$ is fixed, convenient though it may be for showing unbiasedness, is frequently not a reasonable assumption to make in applied econometric work. More commonly, at least some of the columns of $\boldsymbol{X}$ correspond to variables that are no less random than $\boldsymbol{y}$ itself, and it would often stretch credibility to treat them as fixed. Luckily, we can still show that $\hat{\boldsymbol{\beta}}$ and its estimating equations are unbiased in some quite reasonable circumstances without making such a strong assumption.

A weaker assumption is that the explanatory variables which form the columns of $\boldsymbol{X}$ are **exogenous**. The concept of **exogeneity** was introduced in Section 2.3. When applied to the matrix $\boldsymbol{X}$, it implies that any randomness in the DGP that generated $\boldsymbol{X}$ is independent of the disturbances $\boldsymbol{u}$ in the DGP for $\boldsymbol{y}$. This independence in turn implies that

$$\mathrm{E}(\boldsymbol{u} \,|\, \boldsymbol{X}) = \mathbf{0}. \tag{4.11}$$

In words, this says that the expectation of the entire vector $\boldsymbol{u}$, that is, of every one of the u_t, is zero conditional on the entire matrix $\boldsymbol{X}$. See Section 2.2 for a discussion of conditional expectations. Although condition (4.11) is weaker than the condition of independence of $\boldsymbol{X}$ and $\boldsymbol{u}$, it is convenient to refer to (4.11) as an **exogeneity** assumption.

Given the exogeneity assumption (4.11), it is easy to show that both the estimator $\hat{\boldsymbol{\beta}}$ and the estimating equations (4.08) are unbiased. Because the

Copyright © 2021, Russell Davidson and James G. MacKinnon

expectation of $\boldsymbol{X}^\top$ conditional on $\boldsymbol{X}$ is just itself, and the expectation of $\boldsymbol{u}$ conditional on $\boldsymbol{X}$ is assumed to be $\mathbf{0}$, it is clear that $\mathrm{E}(\boldsymbol{X}^\top \boldsymbol{u}) = \mathbf{0}$. Similarly,

$$\mathrm{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} \,|\, \boldsymbol{X}\big) = \mathbf{0}; \tag{4.12}$$

see equation (2.17). Then, applying the Law of Iterated Expectations, we see that the unconditional expectation of the left-hand side of (4.12) must be equal to the expectation of the right-hand side, which is just $\mathbf{0}$.

Assumption (4.11) is perfectly reasonable in the context of some types of data. In particular, suppose that a sample consists of **cross-section data**, in which each observation might correspond to an individual firm, household, person, or city. For many cross-section data sets, there may be no reason to believe that u_t is in any way related to the values of the regressors for any of the observations. On the other hand, suppose that a sample consists of **time-series data**, in which each observation might correspond to a year, quarter, month, or day, as would be the case, for instance, if we wished to estimate a consumption function, as in Chapter 2. Even if we are willing to assume that u_t is in no way related to current and past values of the regressors, it must be related to future values if current values of the dependent variable affect future values of some of the regressors. Thus, in the context of time-series data, the exogeneity assumption (4.11) is a very strong one that we may often not feel comfortable in making.

The assumption that we made in Section 2.3 about the disturbances and the explanatory variables, namely, that

$$\mathrm{E}(u_t \,|\, \boldsymbol{X}_t) = 0, \tag{4.13}$$

is substantially weaker than assumption (4.11), because (4.11) rules out the possibility that the expectation of u_t depends on the values of the regressors for any observation, while (4.13) merely rules out the possibility that it depends on their values for the current observation. For reasons that will become apparent in the next subsection, we refer to (4.13) as a **predeterminedness** condition. Equivalently, we say that the regressors are **predetermined** with respect to the disturbances. Yet another way of expressing this is to call the disturbances **innovations**.

The OLS Estimator Can Be Biased

We have just seen that the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased if we make assumption (4.11) that the explanatory variables $\boldsymbol{X}$ are exogenous, but we remarked that this assumption can sometimes be uncomfortably strong. If we are not prepared to go beyond the predeterminedness assumption (4.13), which it is rarely sensible to do if we are using time-series data, then it turns out that $\hat{\boldsymbol{\beta}}$ is, in general, biased. It is easy to see this in the context of a simple model involving time-series data.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Many regression models for time-series data include one or more **lagged variables** among the regressors. The first lag of a time-series variable that takes on the value z_t at time t has value z_{t-1}. Similarly, the second lag of z_t has value z_{t-2}, and the p^{th} lag has value z_{t-p}. In some models, lags of the dependent variable itself are used as regressors. Indeed, in some cases, the only regressors, except perhaps for a constant term and time trend or dummy variables, are **lagged dependent variables**. Such models are said to be **autoregressive**, because the conditional expectation of the dependent variable depends on lagged values of the dependent variable itself.

A simple example of an autoregressive model is

$$y_t = \beta_1 + \beta_2 y_{t-1} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2). \tag{4.14}$$

This model is said to be **first-order autoregressive**, or **AR(1)**, because only one lag of y_t appears on the right-hand side. The model (4.14) can also be written in vector notation as

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{y}_1 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{4.15}$$

where, as usual, $\boldsymbol{\iota}$ is a vector of 1s, the vector $\boldsymbol{y}$ has typical element y_t, and the vector $\boldsymbol{y}_1$ has typical element y_{t-1}.

It is perfectly reasonable to assume that the predeterminedness condition (4.13) holds for the model (4.14), because this condition amounts to saying that $\mathrm{E}(u_t) = 0$ for every possible value of y_{t-1}. The lagged dependent variable y_{t-1} is then said to be predetermined with respect to the disturbance u_t. Not only is y_{t-1} realized before u_t, but its realized value has no impact on the expectation of u_t. However, it is clear that the exogeneity assumption (4.11), which would here require that $\mathrm{E}(\boldsymbol{u} \,|\, \boldsymbol{y}_1) = \mathbf{0}$, cannot possibly hold, because y_{t-1} depends on u_{t-1}, u_{t-2}, and so on. Assumption (4.11) evidently fails to hold for any model in which the regression function includes a lagged dependent variable.

To see the consequences of assumption (4.11) not holding, we use (4.15) and the FWL Theorem to write out $\hat{\beta}_2$ explicitly as

$$\hat{\beta}_2 = (\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y}_1)^{-1} \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y}.$$

Here $\boldsymbol{M}_{\boldsymbol{\iota}}$ denotes the projection matrix $\mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top$, which centers any vector it multiplies; recall (3.31). If we replace $\boldsymbol{y}$ by $\beta_{10}\boldsymbol{\iota} + \beta_{20}\boldsymbol{y}_1 + \boldsymbol{u}$, where β_{10} and β_{20} are specific values of the parameters, and use the fact that $\boldsymbol{M}_{\boldsymbol{\iota}}$ annihilates the constant vector, we find that

$$\begin{aligned} \hat{\beta}_2 &= (\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y}_1)^{-1} \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{\iota}} (\boldsymbol{y}_1 \beta_{20} + \boldsymbol{u}) \\ &= \beta_{20} + (\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y}_1)^{-1} \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{u}. \end{aligned} \tag{4.16}$$

This is evidently just a special case of (4.05).

Copyright © 2021, Russell Davidson and James G. MacKinnon

It is clear that $\hat{\beta}_2$ is unbiased if and only if the second term in the second line of (4.16) has expectation zero. But this term does *not* have expectation zero. Because $\boldsymbol{y}_1$ is stochastic, we cannot simply move the expectations operator, as we did in (4.10), and then take the unconditional expectation of $\boldsymbol{u}$. Because $\mathrm{E}(\boldsymbol{u} \mid \boldsymbol{y}_1) \neq \mathbf{0}$, we also cannot take expectations conditional on $\boldsymbol{y}_1$, in the way that we took expectations conditional on $\boldsymbol{X}$ in (4.12), and then rely on the Law of Iterated Expectations. In fact, as readers are asked to demonstrate in Exercise 4.1, the estimator $\hat{\beta}_2$ is biased.

It seems reasonable that, if $\hat{\beta}_2$ is biased, so must be $\hat{\beta}_1$. The equivalent of the second line of (4.16) is

$$\hat{\beta}_1 = \beta_{10} + (\boldsymbol{\iota}^\top \boldsymbol{M}_{\boldsymbol{y}_1} \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top \boldsymbol{M}_{\boldsymbol{y}_1} \boldsymbol{u}, \tag{4.17}$$

where the notation should be self-explanatory. Once again, because the vector $\boldsymbol{y}_1$ depends on $\boldsymbol{u}$, we cannot employ the methods that we used in (4.10) or (4.12) to prove that the second term on the right-hand side of (4.17) has expectation zero. In fact, that is not so, and $\hat{\beta}_1$ is consequently biased, as readers are also asked to demonstrate in Exercise 4.1.

The problems we have just encountered when dealing with the autoregressive model (4.14) evidently affect every regression model with random regressors for which the exogeneity assumption (4.11) does not hold. Thus, for all such models, the least-squares estimator of the parameters of the regression function is biased. Assumption (4.11) cannot possibly hold when the regressor matrix $\boldsymbol{X}$ contains lagged dependent variables, and it probably fails to hold for most other models that involve time-series data.

In contrast to the OLS estimator, the OLS estimating equations (4.08) are unbiased whenever the regressors are predetermined. By assumption (4.13),

$$\begin{aligned} \mathrm{E}(\boldsymbol{X}^\top \boldsymbol{u}) &= \mathrm{E}\Big(\sum_{t=1}^n \boldsymbol{X}_t^\top u_t\Big) = \sum_{t=1}^n \mathrm{E}(\boldsymbol{X}_t^\top u_t) \\ &= \sum_{t=1}^n \mathrm{E}\big(\mathrm{E}(\boldsymbol{X}_t^\top u_t \mid \boldsymbol{X}_t)\big) = \sum_{t=1}^n \mathrm{E}\big(\boldsymbol{X}_t^\top \mathrm{E}(u_t \mid \boldsymbol{X}_t)\big) = \mathbf{0}. \end{aligned}$$

Intuitively, the conditions for $\mathrm{E}(\boldsymbol{X}^\top \boldsymbol{u})$ to equal $\mathbf{0}$ are much weaker than the conditions for $\mathrm{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\big)$ to equal $\mathbf{0}$, because we are not inverting a matrix that depends on $\boldsymbol{u}$ or multiplying $\boldsymbol{X}^\top \boldsymbol{u}$ by a random matrix. Thus the bias in the estimator $\hat{\boldsymbol{\beta}}$ arises from the fact that we have to solve the estimating equations (4.08), even though they are themselves unbiased.

Copyright © 2021, Russell Davidson and James G. MacKinnon

4.3 Asymptotic Theory and Consistency

It is sometimes possible to prove that an estimator always has some property, such as unbiasedness, as we did in the previous section for the OLS estimator in certain circumstances. Theoretical results like that one, which hold for every sample, are said to be **exact**. However, it is often difficult or impossible to obtain exact results. We must then be content with approximate results. In econometrics, this is usually accomplished by use of **asymptotic theory**, where the sample size is assumed to tend to infinity in some way. The asymptotic approximation is given by what happens in the limit as the sample size becomes infinite.

A sample of infinite size cannot exist in the real world. In order to calculate the limit, we make use of an **asymptotic construction**, a purely mathematical rule or rules that specify the properties of a sample of arbitrarily large size. For instance, we can imagine simulating data and letting the sample size n become as large as we want. In the case of a model with cross-section data, this is very easy. We can pretend that the original sample is taken from a population of infinite size, and we can imagine drawing more and more observations from that population. In the case of a pure time-series model like (4.14), we can easily generate samples of any size we want, just by letting the simulations run on for long enough. Thus, in both of these cases, we can reasonably think of letting n tend to infinity.

Even in the case of a model with fixed regressors, there are ways to let n tend to infinity. Suppose that the original $\boldsymbol{X}$ matrix is of dimension $m \times k$. Then we can create $\boldsymbol{X}$ matrices of dimensions $2m \times k$, $3m \times k$, $4m \times k$, and so on, simply by stacking as many copies of the original $\boldsymbol{X}$ matrix as we like. By simulating vectors of disturbances of the appropriate dimension, we can then generate n–vectors $\boldsymbol{y}$ for any n that is an integer multiple of m.

In general, an asymptotic construction is meant to reflect the properties of the finite data set being analyzed. The two examples above illustrate this clearly. But in some cases, more than one construction may suggest itself as appropriate. For example, consider the model

$$y_t = \beta_1 + \beta_2 \frac{1}{t} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2). \tag{4.18}$$

Since both regressors here are nonstochastic, the least-squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased, whatever the sample size. Suppose that we have an actual data set with n observations, potentially correctly modeled by (4.18). The values of the second regressor thus range from 1 to $1/n$. If we continue to apply the formula (4.18) for samples of sizes greater than n, the value of the regressor gets smaller and smaller, and ultimately tends to zero. However, if we adopt the construction of the previous paragraph, the regressor remains bounded below by $1/m$. We will return to this example shortly.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Stochastic Limits

In Section 2.5, we described an estimator as a rule for obtaining estimates from any set of data. The rule itself is quite deterministic. Data that are generated by the sort of DGP we have looked at so far are realizations of random vectors or matrices, and so the estimates computed from them are also realizations of random variables. We can therefore think of an estimator as being a random variable whose realizations are estimates. It is a deterministic function of random variables.

For any estimator, an asymptotic construction tells us how to generate a *sequence* of estimators, one for each sample size. If we wish to talk about the limit of this sequence, we must be able to define the limit of a sequence of random variables. Unfortunately, there is more than one form of **stochastic convergence**. Fortunately, for the purposes of this book, we need to consider only two of these. The first is convergence in probability, whereby we can find the **probability limit**, or **plim** for short, of a sequence of random variables. In general, a probability limit is itself a random variable.

Let $\{Y_n\}$ denote a sequence of scalar random variables Y_n, $n = 1, \dots, \infty$. For any such sequence, there must be a *joint* distribution defined for any finite set of elements of the sequence. If the sequence converges in probability, then we may write

$$\operatorname*{plim}_{n\to\infty} Y_n = Y_\infty, \tag{4.19}$$

where Y_∞ is the plim of the sequence. For equation (4.19) to be true, what we need is that, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\big(|Y_n - Y_\infty| > \varepsilon\big) = 0. \tag{4.20}$$

Since a probability is a real number between 0 and 1, the limit in (4.20) is the ordinary limit of a sequence of (deterministic) real numbers. Convergence in probability means that, for any chosen tolerance ε, however small, we can find N large enough so that, for all $n > N$, the probability is smaller than that tolerance of finding the absolute value of the difference between Y_n and the limiting random variable Y_∞ to be greater than the tolerance.

If instead of a sequence of scalar random variables, we consider a sequence of random vectors, or matrices, denoted by $\{\boldsymbol{Y}_n\}$, then $\operatorname{plim}_{n\to\infty} \boldsymbol{Y}_n = \boldsymbol{Y}_\infty$ means that

$$\lim_{n\to\infty} \Pr\big(\|\boldsymbol{Y}_n - \boldsymbol{Y}_\infty\| > \varepsilon\big) = 0.$$

Here $\|\cdot\|$ denotes the Euclidean norm of a vector (see Section 3.2), which simplifies to the absolute value when its argument is a scalar.

The second form of stochastic convergence that we will need is quite different from convergence in probability. It is called **convergence in distribution**, or **convergence in law**, or sometimes **weak convergence**. For a probability limit to exist, there must exist a joint distribution for Y_∞ and any finite set of

Copyright © 2021, Russell Davidson and James G. MacKinnon

the Y_n. However, if the Y_n were all mutually independent, equation (4.20) could never be true, unless the Y_n were actually nonrandom for all n greater than some threshold. In that case, convergence in probability would reduce to the ordinary convergence of a deterministic sequence. Suppose that the Y_n are not only independent but also IID. Then the sequence $\{Y_n\}$ does converge in distribution. However, it cannot possibly converge to a random variable. That should be obvious as soon as we ask ourselves what such a limiting random variable might be.

For convergence in distribution, it is not the random variables themselves that converge, but instead the sequence of CDFs of the random variables. If the random variables are IID, then they all have the same CDF, and so the limiting CDF is simply the CDF of each element of the sequence. In general, a sequence $\{Y_n\}$ of scalar random variables converges in distribution to the distribution characterized by the CDF F if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

where F_n is the CDF of Y_n, for all real x at which F is continuous. We write the relation as

$$Y_n \xrightarrow{d} F. \tag{4.21}$$

It is not necessary for any finite set of the elements of the sequence to have a well-defined joint distribution. The elements of such a finite set may be independent or may be linked by an arbitrary pattern of dependence, without this having any effect on the convergence in distribution (or otherwise) of the sequence, since only the *marginal* CDFs of the individual elements are required to converge.

It is sometimes convenient to write $Y_n \xrightarrow{d} Y_\infty$ as an alternative to (4.21). Here there is a random variable on the right-hand side of the relation, even though Y_∞ may be quite independent of the elements of the sequence. What is required is that the CDF of Y_∞ is the limit of the CDFs F_n as $n \to \infty$.

It is a little trickier to define convergence in distribution for a sequence of vector-valued random variables, although the idea is straightforward. Let the joint CDF of element n of the sequence $\{\boldsymbol{Y}_n\}$ be denoted by $F_n(\boldsymbol{y})$, for $n = 1, \dots, \infty$. The sequence converges in distribution if the sequence $\{F_n(\boldsymbol{y})\}$ converges to a joint CDF $F_\infty(\boldsymbol{y})$, avoiding points at which F_∞ is not continuous. Here, we make no effort to make this final condition precise.

As we discussed above, the convergence in distribution of a sequence in no way implies convergence in probability. In contrast, it can be shown that, if a sequence $\{Y_n\}$ converges in probability, it also converges in distribution. However, there is a special case in which the two concepts coincide, namely, when the limiting distribution is **degenerate**. A distribution is said to be degenerate when all the probability mass of the distribution is concentrated on one single point, so that the plim is nonstochastic. This special case arises frequently in econometrics.

Copyright © 2021, Russell Davidson and James G. MacKinnon

A simple example of a **nonstochastic plim** is the limit of the proportion of heads in a series of independent tosses of an unbiased coin. Suppose that Z_t is a random variable equal to 1 if the coin comes up heads, and equal to 0 if it comes up tails. After n tosses, the proportion of heads is just

$$Y_n \equiv \frac{1}{n}\sum_{t=1}^{n} Z_t.$$

If the coin really is unbiased, $\mathrm{E}(Y_n) = 1/2$. Thus it should come as no surprise to learn that $\operatorname{plim} Y_n = 1/2$. Proving this requires a certain amount of effort, however, and we will therefore not attempt a proof here. For a detailed discussion and proof, see Davidson and MacKinnon (1993, Section 4.2).

The coin-tossing example is really a special case of an extremely powerful result in probability theory, which is called a **law of large numbers**, or **LLN**. Suppose that $\bar{Y}_n$ is the sample mean of Y_t, $t = 1, \ldots, n$, a sequence of random variables, each with expectation μ_Y. Then, provided the Y_t are independent (or at least, not too dependent), a law of large numbers would state that

$$\operatorname*{plim}_{n\to\infty} \bar{Y}_n = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} Y_t = \mu_Y. \tag{4.22}$$

In words, $\bar{Y}_n$ has a nonstochastic plim which is equal to the common expectation of each of the Y_t.

It is not hard to see intuitively why (4.22) is true under certain conditions. Suppose, for example, that the Y_t are IID, with variance σ^2. Then we see at once that

$$\mathrm{E}(\bar{Y}_n) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(Y_t) = \frac{1}{n}\sum_{t=1}^{n} \mu_Y = \mu_Y, \quad \text{and}$$

$$\operatorname{Var}(\bar{Y}_n) = \left(\frac{1}{n}\right)^2 \sum_{t=1}^{n} \sigma^2 = \frac{1}{n}\sigma^2.$$

Thus $\bar{Y}_n$ has expectation μ_Y and a variance which tends to zero as $n \to \infty$. In the limit, we expect that, on account of the shrinking variance, $\bar{Y}_n$ will become a nonstochastic quantity equal to its expectation μ_Y. The law of large numbers assures us that this is indeed the case.

Another useful way to think about laws of large numbers is to note that, as $n \to \infty$, we are collecting more and more information about the expectation of the Y_t, with each individual observation providing a smaller and smaller fraction of that information. Thus, eventually, the random components of the individual Y_t cancel out, and the sample mean $\bar{Y}_n$ converges to the population mean μ_Y. For this to happen, we need to make some assumption in order to prevent any one of the Y_t from having too much impact on $\bar{Y}_n$. The assumption that they are IID is sufficient for this. Alternatively, if they are not IID, we

Copyright © 2021, Russell Davidson and James G. MacKinnon

could assume that the variance of each Y_t is greater than some finite nonzero lower bound, but smaller than some finite upper bound. We also need to assume that there is not too much dependence among the Y_t in order to ensure that the random components of the individual Y_t really do cancel out.

There are actually many laws of large numbers, which differ principally in the conditions that they impose on the random variables which are being averaged. In almost all cases, the result (4.22) is replaced by

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} Y_t = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(Y_t),$$

where the different elements of the sequence of the Y_t may have different expectations. We will not attempt to prove any of these LLNs. Section 4.5 of Davidson and MacKinnon (1993) provides a simple proof of a relatively elementary law of large numbers. More advanced LLNs are discussed in Section 4.7 of that book, and, in more detail, in Davidson (1994).

Probability limits have some very convenient properties. For example, suppose that $\{Y_n\}$, $n = 1, \ldots, \infty$, is a sequence of random variables which has a nonstochastic plim Y_∞ as $n \to \infty$, and $\eta(Y_n)$ is a smooth function of Y_n. Then $\operatorname{plim} \eta(Y_n) = \eta(Y_\infty)$. Another useful property is that, if we have two sequences $\{Y_n\}$ and $\{Z_n\}$ that converge in probability, then $\operatorname{plim} Y_n Z_n = \operatorname{plim} Y_n \operatorname{plim} Z_n$. These features of plims are emphatically not shared by expectations. When $\eta(\cdot)$ is a nonlinear function, $\mathrm{E}\big(\eta(Y)\big) \neq \eta\big(\mathrm{E}(Y)\big)$, and $\mathrm{E}(YZ) \neq \mathrm{E}(Y)\,\mathrm{E}(Z)$ unless Y and Z are independent. Thus, it is often very easy to calculate plims in circumstances where it would be difficult or impossible to calculate expectations.

However, working with plims can be a little bit tricky. The problem is that many of the stochastic quantities we encounter in econometrics do not have probability limits unless we divide them by n or, perhaps, by some power of n. For example, consider the matrix $\boldsymbol{X}^\top\boldsymbol{X}$, which appears in the formula (4.04) for $\hat{\boldsymbol{\beta}}$. Each element of this matrix is a scalar product of two of the columns of $\boldsymbol{X}$, that is, two n–vectors. Thus it is a sum of n numbers. As $n \to \infty$, we would expect that, for any sensible asymptotic construction, such a sum would tend to infinity as well. Therefore, the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ would not generally have a plim. However, it is not at all unreasonable to assume that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}, \tag{4.23}$$

where $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ is a finite nonstochastic matrix, because each element of the matrix on the left-hand side of equation (4.23) is now an average of n numbers:

$$\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)_{ij} = \frac{1}{n}\sum_{t=1}^{n} x_{ti}\,x_{tj}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

In effect, when we write (4.23), we are implicitly making some assumption sufficient for a LLN to hold for the sequences generated by the squares of the regressors and their cross-products. Thus there should not be too much dependence between $x_{ti}x_{tj}$ and $x_{si}x_{sj}$ for $s \neq t$, and the variances of these quantities should not differ too much as t and s vary.

For a more detailed treatment of stochastic convergence, see van der Vaart (1998), especially Chapter 2. Advanced treatments of asymptotic theory in econometrics include Davidson (1994), Gallant (1997), and White (2000).

Same-Order Notation

At this point, it is convenient to introduce the concept of the **same-order relation** and its associated notation. Almost all of the quantities that we encounter in econometrics depend on the sample size. In many cases, when we are using asymptotic theory, the only thing about these quantities that concerns us is the rate at which they change as the sample size changes. The same-order relation provides a very convenient way to deal with such cases.

To begin with, let us suppose that $f(n)$ is a real-valued function of the positive integer n, and r is a rational number. Then we say that $f(n)$ is of the same order as n^r if there exists a constant K, independent of n, and a positive integer N such that

$$\left|\frac{f(n)}{n^r}\right| < K \text{ for all } n > N. \tag{4.24}$$

When $f(n)$ is of the same order as n^r, we can write $f(n) = O(n^r)$. Of course, equation (4.24) does not express an equality in the usual sense. But, as we will see in a moment, this "big O" notation is often very convenient.

The definition we have just given is appropriate only if $f(n)$ is a deterministic function. However, in most econometric applications, some or all of the quantities with which we are concerned are stochastic rather than deterministic. To deal with such quantities, we need to make use of the **stochastic same-order relation**. Let $\{a_n\}$ be a sequence of random variables indexed by the positive integer n. Then we say that a_n is of order n^r in probability if, for all $\varepsilon > 0$, there exist a constant K and a positive integer N such that

$$\Pr\left(\left|\frac{a_n}{n^r}\right| > K\right) < \varepsilon \text{ for all } n > N. \tag{4.25}$$

When a_n is of order n^r in probability, we can write $a_n = O_p(n^r)$. In most cases, it is obvious that a quantity is stochastic, and there is no harm in writing $O(n^r)$ when we really mean $O_p(n^r)$. The properties of the same-order relations are the same in the deterministic and stochastic cases.

The same-order relations are useful because we can manipulate them as if they were simply powers of n. Suppose, for example, that we are dealing with

Copyright © 2021, Russell Davidson and James G. MacKinnon

two functions, $f(n)$ and $g(n)$, which are $O(n^r)$ and $O(n^q)$, respectively. Then

$$\begin{aligned} f(n)g(n) &= O(n^r)O(n^q) = O(n^{r+q}), \text{ and} \\ f(n) + g(n) &= O(n^r) + O(n^q) = O(n^{\max(r,q)}). \end{aligned} \tag{4.26}$$

In the first line here, we see that the order of the product of the two functions is just n to the power $r + q$. In the second line, we see that the order of the sum of the functions is just n to the power that is the maximum of r and q. Both these properties are often very useful in asymptotic analysis.

In equation (4.23), we made the assumption that $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ has a probability limit of $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$, which is a finite, positive definite, deterministic matrix. From the definition (4.20) of a probability limit, it follows that each element of the matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ is $O_p(1)$. Moreover, the definition (4.25) lets us write $\boldsymbol{X}^\top\boldsymbol{X} = O_p(n)$. Similarly, it is very reasonable to assume that $\boldsymbol{X}^\top\boldsymbol{y} = O_p(n)$. In that case, using the first line of equations (4.26), we have

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = O_p(n^{-1})O_p(n) = O_p(1). \tag{4.27}$$

It is customary to write $O(1)$ instead of $O(n^0)$, as we did in the last expression here, but the same-order relation is still about n. Equation (4.27) says that $\hat{\boldsymbol{\beta}}$ does not systematically get larger or smaller as $n \to \infty$, and this is of course a desirable property of any sensible asymptotic construction.

Consistency

Even when the OLS estimator is biased, it may turn out to be **consistent**. Since consistency is an asymptotic property, whether or not it holds depends on the asymptotic construction. Given a model $\mathbb{M}$ and an estimator $\hat{\boldsymbol{\beta}}$ of its parameters, and given a suitable asymptotic construction that allows $\hat{\boldsymbol{\beta}}$ to be defined for arbitrary sample size n, the estimator is consistent if, for every DGP $\mu \in \mathbb{M}$,

$$\operatorname*{plim}_{n\to\infty,\mu} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_\mu,$$

where as before the notation means that, for samples generated by the DGP μ, the plim of the sequence of estimators $\hat{\boldsymbol{\beta}}$ is nonstochastic and equal to the value of $\boldsymbol{\beta}$ associated with μ.

We now show that, under a suitable asymptotic construction, the OLS estimator $\hat{\boldsymbol{\beta}}$ is consistent. When the DGP μ is a special case of the regression model (4.03) that is being estimated, we saw in (4.05) that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_\mu + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}. \tag{4.28}$$

To demonstrate that $\hat{\boldsymbol{\beta}}$ is consistent, we need to show that the second term on the right-hand side here has a plim of zero. This term is the product of two

Copyright © 2021, Russell Davidson and James G. MacKinnon

matrix expressions, $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ and $\boldsymbol{X}^\top\boldsymbol{u}$. Neither $\boldsymbol{X}^\top\boldsymbol{X}$ nor $\boldsymbol{X}^\top\boldsymbol{u}$ has a probability limit. However, we can divide both of these expressions by n without changing the value of this term, since $n \cdot n^{-1} = 1$. By doing so, we convert them into quantities that, for many reasonable asymptotic constructions, have nonstochastic plims. The plim of the second term in (4.28) becomes

$$\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1} \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{u} = \big(\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}\big)^{-1} \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{u} = \boldsymbol{0}. \tag{4.29}$$

In writing the first equality here, we have assumed, first, that equation (4.23) holds, and, second, that $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ is nonsingular. As we discuss below, the second of these assumptions is nontrivial. It can fail even when the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsingular for any finite n.

To obtain the second equality in (4.29), we start with assumption (4.13), which can reasonably be made even when there are lagged dependent variables among the regressors. This assumption tells us that $\mathrm{E}(\boldsymbol{X}_t^\top u_t \mid \boldsymbol{X}_t) = \boldsymbol{0}$, and the Law of Iterated Expectations then tells us that $\mathrm{E}(\boldsymbol{X}_t^\top u_t) = \boldsymbol{0}$. Thus, assuming that we can apply a law of large numbers,

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{u} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}_t^\top u_t = \boldsymbol{0}.$$

Equations (4.28) and (4.29) together give us the result that $\hat{\boldsymbol{\beta}}$ is consistent.

We have just seen that the OLS estimator $\hat{\boldsymbol{\beta}}$ is consistent under considerably weaker assumptions about the relationship between the disturbances and the regressors than were needed to prove that it is unbiased; compare (4.13) and (4.11). This may wrongly suggest that consistency is a weaker condition than unbiasedness. Actually, it is neither weaker nor stronger. Consistency and unbiasedness are simply different concepts. Consistency is an asymptotic property, while unbiasedness is a property that may hold in samples of any size. Sometimes, least-squares estimators may be biased but consistent, for example, in models where $\boldsymbol{X}$ includes lagged dependent variables. In other circumstances, however, these estimators may be unbiased but not consistent.

As an example, consider again the model (4.18), repeated here for ease of reference:

$$y_t = \beta_1 + \beta_2 \frac{1}{t} + u_t, \quad u_t \sim \mathrm{IID}(0, \sigma^2). \tag{4.18}$$

The least-squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are evidently unbiased. However, $\hat{\beta}_2$ is not consistent if we use the first of the asymptotic constructions suggested above, where the formula (4.18) is used unchanged for arbitrarily large samples. The problem is that, as $n \to \infty$, each observation provides less and less information about β_2. This happens because the regressor $1/t$ tends to zero, and hence varies less and less across observations as t becomes larger. As a consequence, the matrix $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ can be shown to be singular; see Exercise 4.5. Therefore, equation (4.29) does not hold, and the second term on the right-hand side of equation (4.28) does not have a probability limit of zero.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The above example illustrates the importance of the choice of the asymptotic construction. Even though $\hat{\beta}_2$ is inconsistent under one asymptotic construction, it is consistent under the alternative asymptotic construction in which we create arbitrarily large samples by stacking copies of the original $\boldsymbol{X}$ matrix. The model (4.18) is in any case rather a curious one, since $\hat{\beta}_1$ is consistent even though $\hat{\beta}_2$ is not, as readers are asked to show in Exercise 4.6. The estimator $\hat{\beta}_1$ is consistent because, as the sample size n gets larger, we obtain an amount of information about β_1 that is roughly proportional to n. In contrast, $\hat{\beta}_2$ is not consistent because each successive observation gives us less and less information about β_2.

We make use of asymptotic constructions in order to obtain approximations to the statistical properties of estimators (or test statistics) for finite sample sizes. A good construction provides a good approximation. A useful guideline for the choice of a good construction is that, as the sample size increases, the data generated for observations past those of the actual sample should have properties, both stochastic and nonstochastic, as similar as possible to those of the observations in the real sample. They should be "**more of the same**." This guideline suggests that the second of the constructions suggested for model (4.18) is preferable to the first.

An estimator that is not consistent is said to be **inconsistent**. There are two types of inconsistency, which are actually quite different. If an unbiased estimator, like $\hat{\beta}_2$ in the previous example, is inconsistent, it is so because it does not tend to any nonstochastic probability limit. In contrast, many inconsistent estimators do tend to nonstochastic probability limits, but they tend to the wrong ones.

To illustrate the various types of inconsistency, and the relationship between bias and inconsistency, imagine that we are trying to estimate the population mean, μ, from a sample of data $y_t, t = 1, \ldots, n$. A sensible estimator would be the sample mean, $\bar{y}$. Under reasonable assumptions about the way the y_t are generated, $\bar{y}$ is unbiased and consistent. Three not very sensible estimators are the following:

$$\hat{\mu}_1 = \frac{1}{n+1}\sum_{t=1}^{n} y_t,$$

$$\hat{\mu}_2 = \frac{1.01}{n}\sum_{t=1}^{n} y_t, \text{ and}$$

$$\hat{\mu}_3 = 0.01\, y_1 + \frac{0.99}{n-1}\sum_{t=2}^{n} y_t.$$

The first of these estimators, $\hat{\mu}_1$, is biased but consistent. It is evidently equal to $n/(n+1)$ times $\bar{y}$. Thus its expectation is $\big(n/(n+1)\big)\mu$, which tends to μ as $n \to \infty$, and it is consistent whenever $\bar{y}$ is. The second estimator, $\hat{\mu}_2$, is clearly biased and inconsistent. Its expectation is 1.01μ, since it is equal to $1.01\,\bar{y}$, and it actually tends to a plim of 1.01μ as $n \to \infty$. The third estimator, $\hat{\mu}_3$,

Copyright © 2021, Russell Davidson and James G. MacKinnon

is perhaps the most interesting. It is clearly unbiased, since it is a weighted average of two estimators, y_1 and the average of y_2 through y_n, each of which is unbiased. The second of these two estimators is also consistent. However, $\hat{\mu}_3$ itself is not consistent, because it does not converge to a nonstochastic plim. Instead, it converges to the random quantity $0.99\mu + 0.01y_1$.

4.4 Covariance Matrices and Precision Matrices

Although it is valuable to know that the least-squares estimator $\hat{\boldsymbol{\beta}}$ can be either unbiased or, under weaker conditions, consistent, this information by itself is not very useful. If we are to interpret any given set of OLS parameter estimates, we need to know, at least approximately, how the vector $\hat{\boldsymbol{\beta}}$ is actually distributed.

For purposes of inference, the most important feature of the distribution of any vector of parameter estimates is the matrix of its central second moments. This matrix is the analog, for vector random variables, of the variance of a scalar random variable. If $\boldsymbol{b}$ is any random vector, its matrix of central second moments may be denoted by $\mathrm{Var}(\boldsymbol{b})$, using the same notation that we would use for a variance in the scalar case. Usage, perhaps somewhat illogically, dictates that this matrix should be called the **covariance matrix**, although the terms **variance matrix** and **variance-covariance matrix** are also sometimes used. Whatever it is called, the covariance matrix is an extremely important concept which comes up over and over again in econometrics.

The covariance matrix $\mathrm{Var}(\boldsymbol{b})$ of a random k–vector $\boldsymbol{b}$, with typical element b_i, organizes all the central second moments of the b_i into a $k \times k$ symmetric matrix. The i^{th} diagonal element of $\mathrm{Var}(\boldsymbol{b})$ is $\mathrm{Var}(b_i)$, the variance of b_i. The ij^{th} off-diagonal element of $\mathrm{Var}(\boldsymbol{b})$ is $\mathrm{Cov}(b_i, b_j)$, the **covariance** of b_i and b_j. The concept of covariance was introduced in Exercise 2.10. In terms of the random variables b_i and b_j, the definition is

$$\mathrm{Cov}(b_i, b_j) \equiv \mathrm{E}\Big(\big(b_i - \mathrm{E}(b_i)\big)\big(b_j - \mathrm{E}(b_j)\big)\Big). \tag{4.30}$$

Many of the properties of covariance matrices follow immediately from (4.30). For example, it is easy to see that, if $i = j$, $\mathrm{Cov}(b_i, b_j) = \mathrm{Var}(b_i)$. Moreover, since from (4.30) it is obvious that $\mathrm{Cov}(b_i, b_j) = \mathrm{Cov}(b_j, b_i)$, $\mathrm{Var}(\boldsymbol{b})$ must be a symmetric matrix. The full covariance matrix $\mathrm{Var}(\boldsymbol{b})$ can be expressed readily using matrix notation. It is just

$$\mathrm{Var}(\boldsymbol{b}) \equiv \mathrm{E}\Big(\big(\boldsymbol{b} - \mathrm{E}(\boldsymbol{b})\big)\big(\boldsymbol{b} - \mathrm{E}(\boldsymbol{b})\big)^\top\Big), \tag{4.31}$$

as is obvious from (4.30). An important special case of equation (4.31) arises when $\mathrm{E}(\boldsymbol{b}) = \boldsymbol{0}$. In this case, $\mathrm{Var}(\boldsymbol{b}) = \mathrm{E}(\boldsymbol{b}\boldsymbol{b}^\top)$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The special case in which $\mathrm{Var}(\boldsymbol{b})$ is diagonal, so that all the covariances are zero, is of particular interest. If b_i and b_j are statistically independent, then $\mathrm{Cov}(b_i, b_j) = 0$; see Exercise 2.13. The converse is not true, however. It is perfectly possible for two random variables that are not statistically independent to have covariance 0; for an extreme example of this, see Exercise 2.14.

The **correlation** between b_i and b_j is

$$\rho(b_i, b_j) \equiv \frac{\mathrm{Cov}(b_i, b_j)}{\big(\mathrm{Var}(b_i)\,\mathrm{Var}(b_j)\big)^{1/2}}. \tag{4.32}$$

It is often useful to think in terms of correlations rather than covariances, because, according to the result of Exercise 4.10, the former always lie between -1 and 1. We can arrange the correlations between all the elements of $\boldsymbol{b}$ into a symmetric matrix called the **correlation matrix**. It is clear from (4.32) that all the elements on the principal diagonal of this matrix must be 1. This demonstrates that the correlation of any random variable with itself equals 1.

In addition to being symmetric, $\mathrm{Var}(\boldsymbol{b})$ must be a **positive semidefinite** matrix; see Exercise 4.9. In most cases, covariance matrices and correlation matrices are **positive definite** rather than positive semidefinite, and their properties depend crucially on this fact.

Positive Definite Matrices

A $k \times k$ symmetric matrix $\boldsymbol{A}$ is said to be positive definite if, for all nonzero k–vectors $\boldsymbol{x}$, the matrix product $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$, which is just a scalar, is positive. The quantity $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ is called a **quadratic form**. A quadratic form always involves a k–vector, in this case $\boldsymbol{x}$, and a $k \times k$ matrix, in this case $\boldsymbol{A}$. By the rules of matrix multiplication,

$$\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{k}\sum_{j=1}^{k} x_i x_j A_{ij}. \tag{4.33}$$

If this quadratic form can take on zero values but not negative values, the matrix $\boldsymbol{A}$ is said to be positive semidefinite.

Any matrix of the form $\boldsymbol{B}^\top\boldsymbol{B}$ is positive semidefinite. To see this, observe that $\boldsymbol{B}^\top\boldsymbol{B}$ is symmetric and that, for any nonzero $\boldsymbol{x}$,

$$\boldsymbol{x}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{x} = (\boldsymbol{B}\boldsymbol{x})^\top(\boldsymbol{B}\boldsymbol{x}) = \|\boldsymbol{B}\boldsymbol{x}\|^2 \geq 0. \tag{4.34}$$

This result can hold with equality only if $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$. But, in that case, since $\boldsymbol{x} \neq \boldsymbol{0}$, the columns of $\boldsymbol{B}$ are linearly dependent. We express this circumstance by saying that $\boldsymbol{B}$ does not have **full column rank**. Note that $\boldsymbol{B}$ can have full rank but not full column rank if $\boldsymbol{B}$ has fewer rows than columns, in which case the maximum possible rank equals the number of rows. However, a matrix

Copyright © 2021, Russell Davidson and James G. MacKinnon

with full column rank necessarily also has full rank. When $\boldsymbol{B}$ does have full column rank, it follows from (4.34) that $\boldsymbol{B}^\top\boldsymbol{B}$ is positive definite. Similarly, if $\boldsymbol{A}$ is positive definite, then any matrix of the form $\boldsymbol{B}^\top\boldsymbol{A}\boldsymbol{B}$ is positive definite if $\boldsymbol{B}$ has full column rank and positive semidefinite otherwise.

It is easy to see that the diagonal elements of a positive definite matrix must all be positive. Suppose that this is not the case and that, say, A_{22} is negative. Then, if we chose $\boldsymbol{x}$ to be the vector $\boldsymbol{e}_2$, that is, a vector with 1 as its second element and all other elements equal to 0 (see Section 3.6), we could make $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} < 0$. From (4.33), the quadratic form would just be $\boldsymbol{e}_2^\top\boldsymbol{A}\boldsymbol{e}_2 = A_{22} < 0$. For a positive semidefinite matrix, the diagonal elements may be 0. Unlike the diagonal elements, the off-diagonal elements of $\boldsymbol{A}$ may be of either sign.

A particularly simple example of a positive definite matrix is the identity matrix, $\mathbf{I}$. Because all the off-diagonal elements are zero, (4.33) tells us that a quadratic form in $\mathbf{I}$ is

$$\boldsymbol{x}^\top\mathbf{I}\boldsymbol{x} = \sum_{i=1}^{k} x_i^2,$$

which is certainly positive for all nonzero vectors $\boldsymbol{x}$. The identity matrix was used in (4.03) in a notation that may not have been clear at the time. There we specified that $\boldsymbol{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I})$. This is just a compact way of saying that the vector of disturbances $\boldsymbol{u}$ is assumed to have expectation $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$.

A positive definite matrix cannot be singular, because, if $\boldsymbol{A}$ is singular, there must exist a nonzero $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \mathbf{0}$. But then $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = 0$ as well, which means that $\boldsymbol{A}$ is not positive definite. Thus the inverse of a positive definite matrix always exists. It too is a positive definite matrix, as readers are asked to show in Exercise 4.13.

There is a sort of converse of the result that any matrix of the form $\boldsymbol{B}^\top\boldsymbol{B}$, where $\boldsymbol{B}$ has full column rank, is positive definite. It is that, if the $k \times k$ matrix $\boldsymbol{A}$ is symmetric and positive definite, then there always exists a full-rank $k \times k$ matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}^\top\boldsymbol{B}$. For any given matrix $\boldsymbol{A}$, the corresponding matrix $\boldsymbol{B}$ is not unique. In particular, $\boldsymbol{B}$ can be chosen to be symmetric, but it can also be chosen to be upper or lower triangular. Details of a simple algorithm (Crout's algorithm) for finding a triangular matrix $\boldsymbol{B}$ can be found in Press, Teukolsky, Vetterling, and Flannery (2007).

For a scalar parameter, the accuracy of an estimator is often taken to be proportional to the inverse of its variance, and this is sometimes called the **precision** of the estimator. For a parameter vector, the **precision matrix** is defined as the inverse of the covariance matrix of the estimator. It exists and is positive definite if and only if the covariance matrix is positive definite.

The OLS Covariance Matrix

The notation we used in the specification (4.03) of the linear regression model can now be explained. It serves both to indicate that the disturbances have

Copyright © 2021, Russell Davidson and James G. MacKinnon

expectation zero, and also to specify the covariance matrix of the disturbances, usually called the **disturbance covariance matrix** or **error covariance matrix**.

If the disturbances are IID, they all have the same variance σ^2, and the covariance of any pair of them is zero. Thus the covariance matrix of the vector $\boldsymbol{u}$ is $\sigma^2\mathbf{I}$, and we have

$$\text{Var}(\boldsymbol{u}) = \text{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2\mathbf{I}. \tag{4.35}$$

A matrix like this one, which is proportional to an identity matrix, is called a **scalar matrix**. Notice that (4.35) does not require the disturbances to be independent, or even that they all have the same distribution. All that is required is that they all have the same variance and that the covariance of each pair of disturbances is zero. This weaker condition is often expressed by saying that the disturbances are **white noise**. Although this expression is a mixed metaphor, it has been in use for a long time, because it provides an easy shorthand way to refer to this condition

When equation (4.35) does not hold, we denote the $n \times n$ disturbance covariance matrix by $\boldsymbol{\Omega}$. When the diagonal elements of the matrix $\boldsymbol{\Omega}$ differ, the disturbances are said to be **heteroskedastic**, or to display **heteroskedasticity**. In the opposite case, when all the disturbances have the same variance, they are said to be **homoskedastic**, or to display **homoskedasticity**.

When $\boldsymbol{\Omega}$ has nonzero off-diagonal elements, the disturbances are said to be **autocorrelated**, or, for time series, **serially correlated**. Autocorrelation may also arise outside a time-series context. For instance, if the observations of a sample characterize different locations in space, they may well display **spatial correlation**. Of course, autocorrelated disturbances may or may not also be heteroskedastic, and heteroskedastic disturbances may or may not also be autocorrelated.

If we assume that the matrix $\boldsymbol{X}$ is exogenous, we can calculate the covariance matrix of $\hat{\boldsymbol{\beta}}$. From (4.05), we know that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}$. By equation (4.31), under the assumption that $\hat{\boldsymbol{\beta}}$ is unbiased, $\text{Var}(\hat{\boldsymbol{\beta}})$ is the expectation of the $k \times k$ matrix

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{4.36}$$

Taking this expectation, conditional on $\boldsymbol{X}$, gives

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\text{E}(\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned} \tag{4.37}$$

This form of covariance matrix is often called a **sandwich covariance matrix**, for the obvious reason that the matrix $\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$ is sandwiched between the two instances of the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. The diagonal elements of $\text{Var}(\hat{\boldsymbol{\beta}})$ are often particularly interesting, since the square root of the k^{th} diagonal element is the standard deviation of $\hat{\beta}_k$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

If $\boldsymbol{\Omega} = \sigma_0^2\mathbf{I}$, so that there is neither heteroskedasticity nor autocorrelation, then equations (4.37) simplify greatly. They become

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\sigma_0^2\mathbf{I}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\ &= \sigma_0^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma_0^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned} \tag{4.38}$$

This is the standard result for the covariance matrix of $\hat{\boldsymbol{\beta}}$ under the conventional, but often too strong, assumptions that $\hat{\boldsymbol{\beta}}$ is unbiased and that the covariance matrix of the disturbances is a scalar matrix.

4.5 Precision of the Least-Squares Estimates

We now investigate what determines the precision of the least-squares estimates $\hat{\boldsymbol{\beta}}$ when the disturbances are homoskedastic and serially uncorrelated, so that equation (4.38) holds. Recall that the precision matrix is the inverse of the covariance matrix. Thus, in this case, the precision matrix is

$$\frac{1}{\sigma_0^2}\boldsymbol{X}^\top\boldsymbol{X}. \tag{4.39}$$

There are really only three things that matter. The first of these is σ_0^2, the true variance of the disturbances. Not surprisingly, the precision matrix (4.39) is inversely proportional to σ_0^2. Thus the more random variation there is in the disturbances, the less precise are the parameter estimates.

The second thing that affects the precision of $\hat{\boldsymbol{\beta}}$ is the sample size n. It is illuminating to rewrite expression (4.39) as

$$\frac{n}{\sigma_0^2}\Big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\Big). \tag{4.40}$$

If we make the assumption (4.23), then $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ is $O_p(1)$. Thus it does not vary systematically with n. In that case, the precision matrix (4.40) must be $O_p(n)$. Thus, if we were to double the sample size, we would expect the precision of $\hat{\boldsymbol{\beta}}$ to be roughly doubled and the standard deviations of the individual $\hat{\beta}_i$ to be divided by $\sqrt{2}$.

As an example, suppose that we are estimating a regression model with just a constant term. We can write the model as $\boldsymbol{y} = \beta_1\boldsymbol{\iota} + \boldsymbol{u}$, where $\boldsymbol{\iota}$ is an n–vector of ones. Plugging in $\boldsymbol{\iota}$ for $\boldsymbol{X}$ in (4.04) and (4.38), we find that

$$\begin{aligned}\hat{\beta}_1 = (\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^\top\boldsymbol{y} = \frac{1}{n}\sum_{t=1}^{n} y_t, \text{ and} \\ \mathrm{Var}(\hat{\beta}_1) = \sigma_0^2(\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1} = \frac{1}{n}\sigma_0^2.\end{aligned}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Thus, in this particularly simple case, the precision of the least-squares estimator is exactly proportional to n, since the variance is proportional to $1/n$.

The third thing that affects the precision of $\hat{\boldsymbol{\beta}}$ is the matrix $\boldsymbol{X}$. Suppose that we are interested in a particular coefficient which, without loss of generality, we may call β_1. Then, if $\boldsymbol{\beta}_2$ denotes the $(k-1)$–vector of the remaining coefficients, we can rewrite the regression model (4.03) as

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \tag{4.41}$$

where $\boldsymbol{X}$ has been partitioned into $\boldsymbol{x}_1$ and $\boldsymbol{X}_2$ to conform with the partition of $\boldsymbol{\beta}$. By the FWL Theorem, regression (4.41) yields the same estimate of β_1 as the FWL regression

$$\boldsymbol{M}_2\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{x}_1\beta_1 + \text{residuals},$$

where, as in Section 3.4, $\boldsymbol{M}_2 \equiv \mathbf{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top$. This estimate is $\hat{\beta}_1 = \boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{y}/\boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{x}_1$, and, by calculations similar to the ones that culminated in equations (4.38), its variance is $\sigma_0^2(\boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{x}_1)^{-1}$. Thus we see that the precision of $\hat{\beta}_1$ is

$$\boldsymbol{x}_1^\top\boldsymbol{M}_2\boldsymbol{x}_1/\sigma_0^2, \tag{4.42}$$

which is equal to the squared length of the vector $\boldsymbol{M}_2\boldsymbol{x}_1$, divided by the variance of the disturbances.

The intuition behind equation (4.42) is simple. How much information the sample gives us about β_1 is proportional to the squared Euclidean length of the vector $\boldsymbol{M}_2\boldsymbol{x}_1$, which is the numerator of (4.42). When $\|\boldsymbol{M}_2\boldsymbol{x}_1\|$ is big, either because n is large or because at least some elements of $\boldsymbol{M}_2\boldsymbol{x}_1$ are large, $\hat{\beta}_1$ is relatively precise. When $\|\boldsymbol{M}_2\boldsymbol{x}_1\|$ is small, either because n is small or because all the elements of $\boldsymbol{M}_2\boldsymbol{x}_1$ are small, $\hat{\beta}_1$ is relatively imprecise.

The squared Euclidean length of the vector $\boldsymbol{M}_2\boldsymbol{x}_1$ is just the sum of squared residuals from the regression

$$\boldsymbol{x}_1 = \boldsymbol{X}_2\boldsymbol{c} + \text{residuals}. \tag{4.43}$$

Thus the precision of $\hat{\beta}_1$, expression (4.42), is proportional to the sum of squared residuals from regression (4.43). When $\boldsymbol{x}_1$ is well explained by the other columns of $\boldsymbol{X}$, this SSR is small, and the variance of $\hat{\beta}_1$ is consequently large. When $\boldsymbol{x}_1$ is not well explained by the other columns of $\boldsymbol{X}$, this SSR is large, and the variance of $\hat{\beta}_1$ is consequently small.

As the above discussion makes clear, the precision with which β_1 is estimated depends on $\boldsymbol{X}_2$ just as much as it depends on $\boldsymbol{x}_1$. Sometimes, if we just regress $\boldsymbol{y}$ on a constant and $\boldsymbol{x}_1$, we may obtain what seems to be a very precise estimate of β_1, but if we then include some additional regressors, the estimate becomes much less precise. The reason for this is that the additional regressors do a much better job of explaining $\boldsymbol{x}_1$ in regression (4.43) than does

Copyright © 2021, Russell Davidson and James G. MacKinnon

a constant alone. As a consequence, the length of $\boldsymbol{M}_2\boldsymbol{x}_1$ is much less than the length of $\boldsymbol{M}_\iota\boldsymbol{x}_1$.

This type of situation is sometimes referred to as **collinearity**, and the regressor $\boldsymbol{x}_1$ is said to be **collinear** with some of the other regressors. This terminology is not very satisfactory, since, if a regressor were collinear with other regressors in the usual mathematical sense of the term, the regressors would be linearly dependent. It would be better to speak of **approximate collinearity**, although econometricians seldom bother with this nicety. Collinearity can cause difficulties for applied econometric work, but these difficulties are essentially the same as the ones caused by having a sample size that is too small. In either case, the data simply do not contain enough information to allow us to obtain precise estimates of all the coefficients.

What we have called collinearity is often called **multicollinearity** by econometricians. This is in fact an abuse of the term multicollinearity as it was introduced by Frisch (1934); see Hendry (1995, Section 7.8) for details.

The covariance matrix of $\hat{\boldsymbol{\beta}}$ summarizes all that we know about its second moments. In practice, of course, we rarely know this matrix, but we usually can estimate it. For the model (4.01), this merely involves getting an estimate of σ_0^2. However, under less restrictive assumptions than those of model (4.01), other ways have to be found to estimate the covariance matrix. Some of these will be discussed in Sections 5.4 and 5.5. Using an appropriate estimate of the covariance matrix, we can then, under appropriate assumptions, make exact or approximate inferences about the parameter vector $\boldsymbol{\beta}$. Just how we can do this will be discussed at length in Chapters 5, 6, and 7.

Linear Functions of Parameter Estimates

The covariance matrix of $\hat{\boldsymbol{\beta}}$ can be used to calculate the variance of any linear (strictly speaking, affine) function of $\hat{\boldsymbol{\beta}}$. Suppose that we are interested in the variance of $\hat{\gamma}$, where $\gamma = \boldsymbol{w}^\top\boldsymbol{\beta}$, $\hat{\gamma} = \boldsymbol{w}^\top\hat{\boldsymbol{\beta}}$, and $\boldsymbol{w}$ is a k–vector of known coefficients. By choosing $\boldsymbol{w}$ appropriately, we can make γ equal to any one of the β_i, or to the sum of the β_i, or to any linear combination of the β_i in which we might be interested. For instance, if we set $\boldsymbol{w}$ equal to a unit basis vector, with $\boldsymbol{w} = \boldsymbol{e}_i$, then the scalar product $\boldsymbol{w}^\top\boldsymbol{\beta} = \boldsymbol{e}_i^\top\boldsymbol{\beta} = \beta_i$. If $\boldsymbol{w} = \boldsymbol{\iota}$, the vector of ones, $\boldsymbol{w}^\top\boldsymbol{\beta} = \sum_i \beta_i$. Again, if $\gamma = 3\beta_1 - \beta_4$, $\boldsymbol{w}$ would be a vector with 3 as the first element, -1 as the fourth element, and 0 for all the other elements.

By (4.31), we see that

$$\begin{aligned}\operatorname{Var}(\hat{\gamma}) &= \operatorname{Var}(\boldsymbol{w}^\top\hat{\boldsymbol{\beta}}) = \operatorname{E}\big(\boldsymbol{w}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\boldsymbol{w}\big)\\ &= \boldsymbol{w}^\top\operatorname{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big)\boldsymbol{w}\\ &= \boldsymbol{w}^\top\operatorname{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{w}. \qquad (4.44)\end{aligned}$$

Notice that, in general, the variance of $\hat{\gamma}$ depends on every element of the covariance matrix of $\hat{\boldsymbol{\beta}}$; this is made explicit in expression (4.85), which readers

Copyright © 2021, Russell Davidson and James G. MacKinnon

are asked to derive in Exercise 4.16. Of course, if some elements of the vector $\boldsymbol{w}$ are equal to 0, $\operatorname{Var}(\hat{\gamma})$ does not depend on the corresponding rows and columns of the covariance matrix.

It may be illuminating to consider the special case used as an example above, in which $\gamma = 3\beta_1 - \beta_4$. In this case, the result (4.44) implies that

$$\begin{aligned}\operatorname{Var}(\hat{\gamma}) &= w_1^2\operatorname{Var}(\hat{\beta}_1) + w_4^2\operatorname{Var}(\hat{\beta}_4) + 2w_1w_4\operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_4)\\ &= 9\operatorname{Var}(\hat{\beta}_1) + \operatorname{Var}(\hat{\beta}_4) - 6\operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_4).\end{aligned}$$

Notice that the variance of $\hat{\gamma}$ depends on the covariance of $\hat{\beta}_1$ and $\hat{\beta}_4$ as well as on their variances. If this covariance is large and positive, $\operatorname{Var}(\hat{\gamma})$ may be small, even when $\operatorname{Var}(\hat{\beta}_1)$ and $\operatorname{Var}(\hat{\beta}_4)$ are both large.

The Variance of Forecast Errors

The result (4.44) can be used to obtain the variance of the error associated with any prediction based on a linear regression. In the time-series context, predictions of y_t are generally made before y_t is actually observed, and these predictions are usually called **forecasts**. We will use the nouns "forecast" and "prediction" interchangeably.

Suppose we have computed a vector of OLS estimates $\hat{\boldsymbol{\beta}}$ and wish to use them to predict y_s, for s not in $1, \dots, n$, using an observed vector of regressors $\boldsymbol{X}_s$. Then the forecast of y_s is simply $\boldsymbol{X}_s\hat{\boldsymbol{\beta}}$. For simplicity, let us assume that the variance of $\hat{\boldsymbol{\beta}}$ is given by (4.38), and that $\hat{\boldsymbol{\beta}}$ is unbiased, which implies that the prediction itself is unbiased. Therefore, the **prediction error** (or **forecast error**) has expectation zero, and its variance is

$$\begin{aligned}\operatorname{E}(y_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}})^2 &= \operatorname{E}(\boldsymbol{X}_s\boldsymbol{\beta}_0 + u_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}})^2\\ &= \operatorname{E}(u_s^2) + \operatorname{E}(\boldsymbol{X}_s\boldsymbol{\beta}_0 - \boldsymbol{X}_s\hat{\boldsymbol{\beta}})^2\\ &= \sigma_0^2 + \operatorname{Var}(\boldsymbol{X}_s\hat{\boldsymbol{\beta}}). \qquad (4.45)\end{aligned}$$

The first equality here depends on the assumption that the regression model is correctly specified, the second depends on the assumption that the disturbances are serially uncorrelated, which ensures that $\operatorname{E}(u_s\boldsymbol{X}_s\hat{\boldsymbol{\beta}}) = 0$, and the third uses the fact that $\hat{\boldsymbol{\beta}}$ is assumed to be unbiased.

Using the result (4.44), and recalling that $\boldsymbol{X}_s$ is a row vector, we see that expression (4.45) is equal to

$$\sigma_0^2 + \boldsymbol{X}_s\operatorname{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_s^\top = \sigma_0^2 + \sigma_0^2\boldsymbol{X}_s(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_s^\top. \qquad (4.46)$$

Thus we find that the variance of the forecast error is the sum of two terms. The first term is simply the variance of the disturbance u_s. If we knew the true value of $\boldsymbol{\beta}$, this would be the variance of the forecast error. The second

Copyright © 2021, Russell Davidson and James G. MacKinnon

term, which makes the forecast error larger than σ_0^2, arises because we are using the estimate $\hat{\boldsymbol{\beta}}$ instead of the true parameter vector $\boldsymbol{\beta}_0$. It expresses **parameter uncertainty**, and it can be thought of as the penalty we pay for our ignorance about $\boldsymbol{\beta}$. Of course, the result (4.46) can easily be generalized to the case in which we are forecasting a vector of values of the dependent variable; see Exercise 4.24.

In practice, the expected squared forecast error is almost always larger than expression (4.45) would suggest. That expression is based on the assumption that the correct model is actually $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}$, but we rarely know the correct model. Suppose the y_t were instead generated by the DGP

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u}, \tag{4.47}$$

where some but not all of the columns of $\boldsymbol{Z}$ may belong to $\mathcal{S}(\boldsymbol{X})$. If we nevertheless estimate the incorrect model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, and treat all the regressors as fixed, we find that

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{\gamma}_0 + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u} \\ &\equiv \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}. \end{aligned} \tag{4.48}$$

Equations (4.48) implicitly define the **pseudo-true** parameter vector $\boldsymbol{\beta}_0$ as $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{\gamma}_0$. More generally, when the regressors are stochastic, $\boldsymbol{\beta}_0$ may be defined as the plim of $\hat{\boldsymbol{\beta}}$ under the DGP (4.47) and an appropriate asymptotic construction.

We can now compute the expected squared forecast error when we incorrectly base our forecast on the false model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$. It is

$$\begin{aligned} \mathrm{E}(y_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}})^2 &= \mathrm{E}(\boldsymbol{Z}_s\boldsymbol{\gamma}_0 + u_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}})^2 \\ &= \mathrm{E}(u_s^2) + \mathrm{E}(\boldsymbol{Z}_s\boldsymbol{\gamma}_0 - \boldsymbol{X}_s\hat{\boldsymbol{\beta}})^2 \\ &= \mathrm{E}(u_s^2) + \mathrm{E}(\boldsymbol{Z}_s\boldsymbol{\gamma}_0 - \boldsymbol{X}_s\boldsymbol{\beta}_0 + \boldsymbol{X}_s\boldsymbol{\beta}_0 - \boldsymbol{X}_s\hat{\boldsymbol{\beta}})^2 \\ &= \sigma_0^2 + \mathrm{E}(\boldsymbol{Z}_s\boldsymbol{\gamma}_0 - \boldsymbol{X}_s\boldsymbol{\beta}_0)^\top(\boldsymbol{Z}_s\boldsymbol{\gamma}_0 - \boldsymbol{X}_s\boldsymbol{\beta}_0) + \mathrm{Var}(\boldsymbol{X}_s\hat{\boldsymbol{\beta}}). \end{aligned} \tag{4.49}$$

In the third line, we add and subtract the predictions that would be obtained if we knew the pseudo-true vector $\boldsymbol{\beta}_0$. The fourth line then follows from a little algebra and the exogeneity assumption that $\mathrm{E}(\boldsymbol{u}|\boldsymbol{X}, \boldsymbol{Z}) = \boldsymbol{0}$.

It is evident from the last line of equations (4.49) that the expected squared forecast error will be larger when the DGP is (4.47) than when it is a special case of the model we are using. The first and last terms in the last line of equations (4.49) are the same as the two terms in expression (4.45). The middle term, which can be thought of as a squared bias, arises because using a false model causes additional forecast errors. If $\boldsymbol{X}_s\boldsymbol{\beta}_0$ provides a good approximation to $\boldsymbol{Z}_s\boldsymbol{\gamma}_0$, the middle term will be small. But if it provides a poor approximation, the middle term may be large, and the expected squared

Copyright © 2021, Russell Davidson and James G. MacKinnon

forecast error may be much larger than expression (4.46) would suggest, on account of **model uncertainty**.

Our discussion of forecast errors has been extremely brief. Because forecasting is a very important application of econometric methods, there is a large literature on it. Important topics include forecast evaluation, methods for comparing competing forecasts, and methods for combining forecasts. Elliott and Timmermann (2016) provides a comprehensive treatment of forecasting in economics. Other useful references include Clements and Hendry (2002), Elliott and Timmermann (2008), and Pesaran (2015, Chapter 17). One perhaps surprising result, which has been observed many times, is that simple models often produce better forecasts than more complicated models, even when the latter appear to fit better within sample.

4.6 Efficiency of the OLS Estimator

One of the reasons for the popularity of ordinary least squares is that, under certain conditions, the OLS estimator can be shown to be more **efficient** than many competing estimators. One estimator is said to be more efficient than another if, on average, it yields more accurate estimates than the other. The reason for the terminology is that an estimator which yields more accurate estimates can be thought of as utilizing the information available in the sample more efficiently.

For scalar parameters, one estimator of a parameter is more efficient than another if the precision of the former is larger than that of the latter. For parameter vectors, there is a natural way to generalize this idea. Suppose that $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are two unbiased estimators of a k–vector of parameters $\boldsymbol{\beta}$, with covariance matrices $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ and $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$, respectively. Then, if efficiency is measured in terms of precision, $\tilde{\boldsymbol{\beta}}$ is said to be more efficient than $\hat{\boldsymbol{\beta}}$ if and only if the difference between their precision matrices, $\mathrm{Var}(\tilde{\boldsymbol{\beta}})^{-1} - \mathrm{Var}(\hat{\boldsymbol{\beta}})^{-1}$, is a nonzero positive semidefinite matrix.

Since it is more usual to work in terms of variance than of precision, it is convenient to express the efficiency condition directly in terms of covariance matrices. As readers are asked to show in Exercise 4.14, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive definite matrices of the same dimensions, then the matrix $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite if and only if $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$ is positive semidefinite. Thus the efficiency condition expressed above in terms of precision matrices is equivalent to saying that $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$ if and only if $\mathrm{Var}(\hat{\boldsymbol{\beta}}) - \mathrm{Var}(\tilde{\boldsymbol{\beta}})$ is a nonzero positive semidefinite matrix.

If $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$ in this sense, then every individual parameter in the vector $\boldsymbol{\beta}$, and every linear combination of those parameters, is estimated at least as efficiently by using $\tilde{\boldsymbol{\beta}}$ as by using $\hat{\boldsymbol{\beta}}$. Consider an arbitrary linear combination of the parameters in $\boldsymbol{\beta}$, say $\gamma = \boldsymbol{w}^\top\boldsymbol{\beta}$, for any k–vector $\boldsymbol{w}$ that

Copyright © 2021, Russell Davidson and James G. MacKinnon

we choose. As we saw in the preceding section, $\mathrm{Var}(\tilde{\gamma}) = \boldsymbol{w}^\top \mathrm{Var}(\tilde{\boldsymbol{\beta}})\boldsymbol{w}$, and similarly for $\mathrm{Var}(\hat{\gamma})$. Therefore, the difference between $\mathrm{Var}(\hat{\gamma})$ and $\mathrm{Var}(\tilde{\gamma})$ is

$$\boldsymbol{w}^\top \mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{w} - \boldsymbol{w}^\top \mathrm{Var}(\tilde{\boldsymbol{\beta}})\boldsymbol{w} = \boldsymbol{w}^\top\big(\mathrm{Var}(\hat{\boldsymbol{\beta}}) - \mathrm{Var}(\tilde{\boldsymbol{\beta}})\big)\boldsymbol{w}. \tag{4.50}$$

The right-hand side of equation (4.50) must be either positive or zero whenever the matrix $\mathrm{Var}(\hat{\boldsymbol{\beta}}) - \mathrm{Var}(\tilde{\boldsymbol{\beta}})$ is positive semidefinite. Thus, if $\tilde{\boldsymbol{\beta}}$ is a more efficient estimator than $\hat{\boldsymbol{\beta}}$, we can be sure that $\tilde{\gamma}$ is estimated with no more variance than $\hat{\gamma}$. In practice, when one estimator is more efficient than another, the difference between the covariance matrices is very often positive definite. When that is the case, every parameter or linear combination of parameters is estimated more efficiently using $\tilde{\boldsymbol{\beta}}$ than using $\hat{\boldsymbol{\beta}}$.

We now let $\tilde{\boldsymbol{\beta}}$ denote the vector of OLS parameter estimates. As we are about to show, this estimator is more efficient than any other **linear unbiased estimator**. In Section 4.2, we discussed what it means for an estimator to be unbiased, but we have not yet discussed what it means for an estimator to be linear. It simply means that we can write the estimator as a linear function of $\boldsymbol{y}$, the vector of observations on the dependent variable. It is clear that $\tilde{\boldsymbol{\beta}}$ itself is a linear estimator, because it is equal to the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ times the vector $\boldsymbol{y}$.

If $\hat{\boldsymbol{\beta}}$ now denotes any linear estimator that is not the OLS estimator, we can always write

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{C}\boldsymbol{y}, \tag{4.51}$$

where $\boldsymbol{A}$ and $\boldsymbol{C}$ are non-random or exogenous $k \times n$ matrices that usually depend on $\boldsymbol{X}$. The first equality here just says that $\hat{\boldsymbol{\beta}}$ is a linear estimator. The second follows from our definition of $\boldsymbol{C}$:

$$\boldsymbol{C} \equiv \boldsymbol{A} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top. \tag{4.52}$$

So far, least squares is the only estimator for linear regression models that we have encountered. Thus it may be difficult to imagine what kind of estimator $\hat{\boldsymbol{\beta}}$ might be. In fact, there are many estimators of this type, including **instrumental variables** estimators (Chapter 8)) and **generalized least-squares** estimators (Chapter 9). An alternative way of writing the class of linear unbiased estimators is explored in Exercise 4.25.

The principal theoretical result on the efficiency of the OLS estimator is called the **Gauss-Markov Theorem**. An informal way of stating this theorem is to say that $\tilde{\boldsymbol{\beta}}$ is the **best linear unbiased estimator**, or **BLUE** for short. In other words, the OLS estimator is more efficient than any other linear unbiased estimator.

Theorem 4.1. (Gauss-Markov Theorem)

If it is assumed that $\mathrm{E}(\boldsymbol{u} \,|\, \boldsymbol{X}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top \,|\, \boldsymbol{X}) = \sigma^2\mathbf{I}$ in the linear regression model (4.03), then the OLS estimator $\tilde{\boldsymbol{\beta}}$ is more efficient than any other linear unbiased estimator $\hat{\boldsymbol{\beta}}$, in the sense that $\mathrm{Var}(\hat{\boldsymbol{\beta}}) - \mathrm{Var}(\tilde{\boldsymbol{\beta}})$ is a positive semidefinite matrix.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Proof:

We assume that the DGP is a special case of (4.03), with parameters $\boldsymbol{\beta}_0$ and σ_0^2. Substituting for $\boldsymbol{y}$ in (4.51), we find that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{A}\boldsymbol{u}. \tag{4.53}$$

Since we want $\hat{\boldsymbol{\beta}}$ to be unbiased, we require that the expectation of the rightmost expression in (4.53), conditional on $\boldsymbol{X}$, should be $\boldsymbol{\beta}_0$. The second term in that expression has conditional expectation $\boldsymbol{0}$, and so the first term must have conditional expectation $\boldsymbol{\beta}_0$. This is the case for all $\boldsymbol{\beta}_0$ if and only if $\boldsymbol{A}\boldsymbol{X} = \mathbf{I}$, the $k \times k$ identity matrix. From (4.52), this condition is equivalent to $\boldsymbol{C}\boldsymbol{X} = \mathbf{O}$. Thus requiring $\hat{\boldsymbol{\beta}}$ to be unbiased imposes a strong condition on the matrix $\boldsymbol{C}$.

The unbiasedness condition that $\boldsymbol{C}\boldsymbol{X} = \mathbf{O}$ implies that $\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\boldsymbol{u}$. Since, from (4.51), $\boldsymbol{C}\boldsymbol{y} = \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$, this makes it clear that $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$ has conditional expectation zero. The unbiasedness condition also implies that the matrix of the covariances of elements of the matrix of $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$ and those of $\tilde{\boldsymbol{\beta}}$ is a zero matrix. To see this, observe that

$$\begin{aligned} \mathrm{E}\big((\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big) &= \mathrm{E}\big(\boldsymbol{C}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big) \\ &= \boldsymbol{C}\sigma_0^2\mathbf{I}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\ &= \sigma_0^2\boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \mathbf{O}. \end{aligned} \tag{4.54}$$

Consequently, equation (4.51) says that the unbiased linear estimator $\hat{\boldsymbol{\beta}}$ is equal to the least-squares estimator $\tilde{\boldsymbol{\beta}}$ plus a random component $\boldsymbol{C}\boldsymbol{u}$ which has expectation zero and is uncorrelated with $\tilde{\boldsymbol{\beta}}$. The random component simply adds noise to the efficient estimator $\tilde{\boldsymbol{\beta}}$. This makes it clear that $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$. To complete the proof, we note that

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}\big(\tilde{\boldsymbol{\beta}} + (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})\big) \\ &= \mathrm{Var}\big(\tilde{\boldsymbol{\beta}} + \boldsymbol{C}\boldsymbol{u}\big) \\ &= \mathrm{Var}(\tilde{\boldsymbol{\beta}}) + \mathrm{Var}(\boldsymbol{C}\boldsymbol{u}), \end{aligned} \tag{4.55}$$

because, from (4.54), the covariance of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{C}\boldsymbol{u}$ is zero. Thus the difference between $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ and $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$ is $\mathrm{Var}(\boldsymbol{C}\boldsymbol{u})$. Since it is a covariance matrix, this difference is necessarily positive semidefinite. ■

Remark: For the Gauss-Markov theorem to hold, it is *not* necessary to suppose that the disturbances are normally distributed.

We will encounter many cases in which an inefficient estimator is equal to an efficient estimator plus a random variable that has expectation zero and is uncorrelated with the efficient estimator. The zero correlation ensures that the

Copyright © 2021, Russell Davidson and James G. MacKinnon

covariance matrix of the inefficient estimator is equal to the covariance matrix of the efficient estimator plus another matrix that is positive semidefinite, as in the last line of (4.55). If the correlation were not zero, this sort of proof would not work.

The Gauss-Markov Theorem that the OLS estimator is BLUE is one of the most famous results in statistics. However, it is important to keep in mind the limitations of this theorem. The theorem applies only to a correctly specified model with exogenous regressors and disturbances with a scalar covariance matrix. Moreover, it does *not* say that the OLS estimator $\hat{\boldsymbol{\beta}}$ is more efficient than every imaginable estimator. Estimators which are nonlinear and/or biased may well perform better than ordinary least squares.

4.7 Residuals and Disturbances

Once we have obtained the OLS estimates $\hat{\boldsymbol{\beta}}$, it is easy to calculate the vector of least-squares residuals, $\hat{\boldsymbol{u}} \equiv \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.[1] The numerical properties of $\hat{\boldsymbol{u}}$ were discussed in Section 3.3. These properties include the fact that $\hat{\boldsymbol{u}}$ is orthogonal to $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and to every vector that lies in $\mathcal{S}(\boldsymbol{X})$. In this section, we turn our attention to the statistical properties of $\hat{\boldsymbol{u}}$ as an estimator of $\boldsymbol{u}$. These properties are very important, because we will want to use $\hat{\boldsymbol{u}}$ for a number of purposes. In particular, we will want to use it to estimate σ^2, the variance of the disturbances. We need an estimate of σ^2 if we are to estimate the covariance matrix (4.38) of $\hat{\boldsymbol{\beta}}$. As we will see in later chapters, the residuals can also be used to test some of the strong assumptions that are often made about the distribution of the disturbances and to implement more sophisticated estimation methods that require weaker assumptions.

The consistency of $\hat{\boldsymbol{\beta}}$ implies that $\hat{\boldsymbol{u}} \to \boldsymbol{u}$ as $n \to \infty$, but the finite-sample properties of $\hat{\boldsymbol{u}}$ differ from those of $\boldsymbol{u}$. As we saw in Section 3.3, the vector of residuals $\hat{\boldsymbol{u}}$ is what remains after we project the regressand $\boldsymbol{y}$ off $\mathcal{S}(\boldsymbol{X})$. Suppose we are estimating the linear regression model (4.01). If we assume that the DGP belongs to this model, as (4.02) does, then

$$\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}.$$

The first term in the middle expression here vanishes because $\boldsymbol{M}_{\boldsymbol{X}}$ annihilates everything that lies in $\mathcal{S}(\boldsymbol{X})$. The statistical properties of $\hat{\boldsymbol{u}}$ as an estimator of $\boldsymbol{u}$ when the model (4.01) is correctly specified follow directly from the key result that $\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$.

This result implies that each of the residuals is equal to a linear combination of every one of the disturbances. Consider a single row of the matrix product

[1] For the remainder of this chapter, we revert to letting $\hat{\boldsymbol{\beta}}$ rather than $\tilde{\boldsymbol{\beta}}$ denote the OLS estimator.

Copyright © 2021, Russell Davidson and James G. MacKinnon

$\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$. Since the product has dimensions $n \times 1$, this row has just one element, and this element is one of the residuals. Recalling the result on partitioned matrices in Exercise 2.16, which allows us to select any row of a matrix product by selecting the corresponding row of the leftmost factor, we can write the t^{th} residual as

$$\begin{aligned}\hat{u}_t &= u_t - \boldsymbol{X}_t(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u} \\ &= u_t - \sum_{s=1}^{n} \boldsymbol{X}_t(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_s^\top u_s. \end{aligned} \tag{4.56}$$

Thus, even if each of the disturbances u_t is independent of all the others, as we have been assuming, each of the $\hat{u}_t$ is not independent of all the other residuals. In general, there is some dependence between every pair of residuals. However, this dependence generally diminishes as the sample size n increases.

Let us now assume that $\mathrm{E}(\boldsymbol{u} \,|\, \boldsymbol{X}) = \boldsymbol{0}$. This is assumption (4.11), which we made in Section 4.2 in order to prove that $\hat{\boldsymbol{\beta}}$ is unbiased. According to this assumption, $\mathrm{E}(u_t \,|\, \boldsymbol{X}) = 0$ for all t. All the expectations we will take in the remainder of this section will be conditional on $\boldsymbol{X}$. Since, by (4.56), $\hat{u}_t$ is just a linear combination of all the u_t, the expectation of $\hat{u}_t$ conditional on $\boldsymbol{X}$ must be zero. Thus, in this respect, the residuals $\hat{u}_t$ behave just like the disturbances u_t.

In other respects, however, the residuals do not have the same properties as the disturbances. Consider $\mathrm{Var}(\hat{u}_t)$, the variance of $\hat{u}_t$. Since $\mathrm{E}(\hat{u}_t) = 0$, this variance is just $\mathrm{E}(\hat{u}_t^2)$. As we saw in Section 3.3, the Euclidean length of the vector of least-squares residuals, $\hat{\boldsymbol{u}}$, is always smaller than that of the vector of residuals evaluated at any other value, $\boldsymbol{u}(\boldsymbol{\beta})$. In particular, $\hat{\boldsymbol{u}}$ must be shorter than the vector of disturbances $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{\beta}_0)$. Thus we know that $\|\hat{\boldsymbol{u}}\|^2 \leq \|\boldsymbol{u}\|^2$. This implies that $\mathrm{E}\big(\|\hat{\boldsymbol{u}}\|^2\big) \leq \mathrm{E}\big(\|\boldsymbol{u}\|^2\big)$. If, as usual, we assume that the variance of the disturbances is σ_0^2 under the true DGP, we see that

$$\begin{aligned}\sum_{t=1}^{n} \mathrm{Var}(\hat{u}_t) &= \sum_{t=1}^{n} \mathrm{E}(\hat{u}_t^2) = \mathrm{E}\left(\sum_{t=1}^{n} \hat{u}_t^2\right) = \mathrm{E}\big(\|\hat{\boldsymbol{u}}\|^2\big) \\ &\leq \mathrm{E}\big(\|\boldsymbol{u}\|^2\big) = \mathrm{E}\left(\sum_{t=1}^{n} u_t^2\right) = \sum_{t=1}^{n} \mathrm{E}(u_t^2) = n\sigma_0^2. \end{aligned}$$

This suggests that, at least for most observations, the variance of $\hat{u}_t$ must be less than σ_0^2. In fact, as we are about to see, $\mathrm{Var}(\hat{u}_t)$ is less than σ_0^2 for every observation.

The easiest way to calculate the variance of $\hat{u}_t$ is to calculate the covariance matrix of the entire vector $\hat{\boldsymbol{u}}$:

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{u}}) &= \mathrm{Var}(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}) = \mathrm{E}(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{M}_{\boldsymbol{X}}) \\ &= \boldsymbol{M}_{\boldsymbol{X}}\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{M}_{\boldsymbol{X}}\mathrm{Var}(\boldsymbol{u})\boldsymbol{M}_{\boldsymbol{X}} \\ &= \boldsymbol{M}_{\boldsymbol{X}}(\sigma_0^2\mathbf{I})\boldsymbol{M}_{\boldsymbol{X}} = \sigma_0^2\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{X}} = \sigma_0^2\boldsymbol{M}_{\boldsymbol{X}}. \end{aligned} \tag{4.57}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The second equality in the first line here uses the fact that $\boldsymbol{M_X u}$ has expectation $\mathbf{0}$. The third equality in the last line uses the fact that $\boldsymbol{M_X}$ is idempotent. From the result (4.57), we see immediately that, in general, $\mathrm{E}(\hat{u}_t \hat{u}_s) \neq 0$ for $t \neq s$. Thus, even though the original disturbances are assumed to be uncorrelated, the residuals are not uncorrelated.

From equations (4.57), it can also be seen that the residuals do not have a constant variance, and that the variance of every residual must always be smaller than σ_0^2. Recall from Section 3.6 that h_t denotes the t^{th} diagonal element of the projection matrix $\boldsymbol{P_X}$. Thus a typical diagonal element of $\boldsymbol{M_X}$ is $1 - h_t$. Therefore, it follows from (4.57) that

$$\mathrm{Var}(\hat{u}_t) = \mathrm{E}(\hat{u}_t^2) = (1 - h_t)\sigma_0^2. \tag{4.58}$$

Since $0 \leq 1 - h_t < 1$, this equation implies that $\mathrm{E}(\hat{u}_t^2)$ must always be smaller than σ_0^2. Just how much smaller depends on h_t. It is clear that high-leverage observations, for which h_t is relatively large, must have residuals with smaller variance than low-leverage observations, for which h_t is relatively small. This makes sense, since high-leverage observations have more impact on the parameter values. As a consequence, the residuals for high-leverage observations tend to be shrunk more, relative to the disturbances, than the residuals for low-leverage observations.

Estimating the Variance of the Disturbances

The method of least squares provides estimates of the regression coefficients, but it does not directly provide an estimate of σ^2, the variance of the disturbances. The method of moments suggests that we can estimate σ^2 by using the corresponding sample moment. If we actually observed the u_t, this sample moment would be

$$\frac{1}{n}\sum_{t=1}^{n} u_t^2. \tag{4.59}$$

We do not observe the u_t, but we do observe the $\hat{u}_t$. Thus the simplest possible MM estimator is

$$\hat{\sigma}^2 \equiv \frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2. \tag{4.60}$$

This estimator is just the average of n squared residuals. It can be shown to be consistent; see Exercise 4.21. However, because each squared residual has expectation less than σ_0^2, by (4.58), $\hat{\sigma}^2$ must be biased downward.

It is easy to calculate the bias of $\hat{\sigma}^2$. We saw in Section 3.6 that $\sum_{t=1}^{n} h_t = k$. Therefore, from (4.58) and (4.60),

$$\mathrm{E}(\hat{\sigma}^2) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(\hat{u}_t^2) = \frac{1}{n}\sum_{t=1}^{n} (1 - h_t)\sigma_0^2 = \frac{n-k}{n}\sigma_0^2. \tag{4.61}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Since $\hat{\boldsymbol{u}} = \boldsymbol{M_X u}$ and $\boldsymbol{M_X}$ is idempotent, the sum of squared residuals is just $\boldsymbol{u}^\top \boldsymbol{M_X u}$. The result (4.61) implies that

$$\mathrm{E}\big(\boldsymbol{u}^\top \boldsymbol{M_X u}\big) = \mathrm{E}\big(\mathrm{SSR}(\hat{\boldsymbol{\beta}})\big) = \mathrm{E}\Big(\sum_{t=1}^{n} \hat{u}_t^2\Big) = (n-k)\sigma_0^2. \tag{4.62}$$

Readers are asked to show this in a different way in Exercise 4.22. Notice, from (4.62), that adding one more regressor has exactly the same effect on the expectation of the SSR as taking away one observation.

The result (4.61) suggests another estimator, which is unbiased:

$$s^2 \equiv \frac{1}{n-k}\sum_{t=1}^{n} \hat{u}_t^2. \tag{4.63}$$

The only difference between $\hat{\sigma}^2$ and s^2 is that the former divides the SSR by n and the latter divides it by $n - k$. As a result, s^2 is unbiased whenever $\hat{\boldsymbol{\beta}}$ is. Ideally, if we were able to observe the disturbances, our estimator would be (4.59), which would be unbiased. When we replace the disturbances u_t by the residuals $\hat{u}_t$, we introduce a downward bias. Dividing by $n - k$ instead of by n eliminates this bias.

Virtually all OLS regression programs report s^2 as the estimated variance of the disturbances. The square root of this estimate, s, is called the **standard error of the regression** or the **regression standard error**. It is important to remember that, even though s^2 provides an unbiased estimate of σ_0^2, s itself does not provide an unbiased estimate of σ_0, because taking the square root of s^2 is a nonlinear operation. If we replace σ_0^2 by s^2 in expression (4.38), we can obtain an unbiased estimate of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$,

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{4.64}$$

This is the usual estimate of the covariance matrix of the OLS parameter estimates under the assumption of IID disturbances.

It is easy to see that s^2 is not only unbiased, but also consistent under any reasonable asymptotic construction. We know from Exercise 4.21 that $\hat{\sigma}^2$ is consistent, and, since $\lim_{n\to\infty}(n-k)/n = 1$, the consistency of s^2 follows immediately from (4.63).

4.8 Misspecification of Linear Regression Models

Up to this point, we have (with one exception) assumed that the DGP belongs to the model that is being estimated, or, in other words, that the model is correctly specified. This is obviously a very strong assumption indeed. It is therefore important to know something about the statistical properties of $\hat{\boldsymbol{\beta}}$ when the model is not correctly specified. In this section, we consider a simple case of misspecification, namely, **underspecification**. In order to understand underspecification better, we begin by discussing its opposite, **overspecification**.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Overspecification

A model is said to be **overspecified** if some variables that rightly belong to the information set Ω_t, but do not appear in the DGP, are mistakenly included in the model. Overspecification is *not* a form of misspecification. Including irrelevant explanatory variables in a model makes the model larger than it need have been, but, since the DGP remains a special case of the model, there is no misspecification. Consider the case of an overspecified linear regression model. Suppose that we estimate the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{4.65}$$

when the data are actually generated by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_0^2\mathbf{I}). \tag{4.66}$$

It is assumed that $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$, the t^{th} rows of $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively, belong to Ω_t. Recall the discussion of information sets in Section 2.3. The overspecified model (4.65) is not misspecified, since the DGP (4.66) is a special case of it, with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma} = \boldsymbol{0}$, and $\sigma^2 = \sigma_0^2$.

Suppose now that we run the linear regression (4.65). By the FWL Theorem, the estimates $\hat{\boldsymbol{\beta}}$ from (4.65) are the same as those from the regression

$$\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}\boldsymbol{\beta} + \text{residuals},$$

where, as usual, $\boldsymbol{M}_{\boldsymbol{Z}} = \mathbf{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$. Thus we see that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}. \tag{4.67}$$

Since $\hat{\boldsymbol{\beta}}$ is part of the OLS estimator of a correctly specified model, it must be unbiased if $\boldsymbol{X}$ and $\boldsymbol{Z}$ are exogenous. Indeed, if we replace $\boldsymbol{y}$ by $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$, we find from (4.67) that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{u}. \tag{4.68}$$

The conditional expectation of the second term on the right-hand side of (4.68) is $\boldsymbol{0}$, provided we take expectations conditional on $\boldsymbol{Z}$ as well as on $\boldsymbol{X}$; see Section 4.2. Since $\boldsymbol{Z}_t$ is assumed to belong to Ω_t, it is perfectly legitimate to do this.

If we had estimated (4.65) subject to the valid restriction that $\boldsymbol{\gamma} = \boldsymbol{0}$, we would have obtained the OLS estimate $\tilde{\boldsymbol{\beta}}$, expression (4.04), which is unbiased and has covariance matrix (4.38). We see that both $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ are unbiased estimators, linear in $\boldsymbol{y}$. Both are OLS estimators, and so it seems that we should be able to apply the Gauss-Markov Theorem to both of them. This is in fact correct, but we must be careful to apply the theorem in the context of the appropriate model for each of the estimators.

Copyright © 2021, Russell Davidson and James G. MacKinnon

For $\tilde{\boldsymbol{\beta}}$, the appropriate model is the **restricted model**,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{4.69}$$

in which the restriction $\boldsymbol{\gamma} = \boldsymbol{0}$ is explicitly imposed. Provided this restriction is correct, as it is if the true DGP takes the form (4.66), $\tilde{\boldsymbol{\beta}}$ must be more efficient than any other linear unbiased estimator of $\boldsymbol{\beta}$. Thus we should find that the matrix $\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\tilde{\boldsymbol{\beta}})$ is positive semidefinite.

For $\hat{\boldsymbol{\beta}}$, the appropriate model is the **unrestricted model** (4.65). In this context, the Gauss-Markov Theorem says that, when we do not know the true value of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. It is important to note here that $\tilde{\boldsymbol{\beta}}$ is *not* an unbiased estimator of $\boldsymbol{\beta}$ for the unrestricted model, and so it cannot be included in the class of estimators covered by the Gauss-Markov Theorem for that model. We will make this point more fully in the next subsection, when we discuss underspecification.

It is illuminating to check these consequences of the Gauss-Markov Theorem explicitly. From equation (4.68), it follows that

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big) \\
&= (\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\text{E}(\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1} \\
&= \sigma_0^2(\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\mathbf{I}\,\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1} \\
&= \sigma_0^2(\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}.
\end{aligned} \tag{4.70}$$

The situation is clear in the case in which there is only one parameter, β, corresponding to a single regressor, $\boldsymbol{x}$. Since $\boldsymbol{M}_{\boldsymbol{Z}}$ is a projection matrix, the Euclidean length of $\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{x}$ must be smaller (or at least, no larger) than the Euclidean length of $\boldsymbol{x}$; recall (3.27). Thus $\boldsymbol{x}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{x} \leq \boldsymbol{x}^\top\boldsymbol{x}$, which implies that

$$\sigma_0^2(\boldsymbol{x}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{x})^{-1} \geq \sigma_0^2(\boldsymbol{x}^\top\boldsymbol{x})^{-1}. \tag{4.71}$$

The inequality in (4.71) almost always holds strictly. The only exception is the special case in which $\boldsymbol{x}$ lies in $\mathcal{S}^\perp(\boldsymbol{Z})$, which implies that the regression of $\boldsymbol{x}$ on $\boldsymbol{Z}$ has no explanatory power at all.

In general, we wish to show that $\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\tilde{\boldsymbol{\beta}})$ is a positive semidefinite matrix. As we saw in Section 4.5, this is equivalent to showing that the matrix $\text{Var}(\tilde{\boldsymbol{\beta}})^{-1} - \text{Var}(\hat{\boldsymbol{\beta}})^{-1}$ is positive semidefinite. A little algebra shows that

$$\begin{aligned}
\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X} &= \boldsymbol{X}^\top(\mathbf{I} - \boldsymbol{M}_{\boldsymbol{Z}})\boldsymbol{X} \\
&= \boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} = (\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^\top\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}.
\end{aligned} \tag{4.72}$$

Since $\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}$ can be written as the transpose of a matrix times itself, it must be positive semidefinite. Dividing by σ_0^2 gives the desired result.

Copyright © 2021, Russell Davidson and James G. MacKinnon

We have established that the OLS estimator of $\boldsymbol{\beta}$ in the overspecified regression model (4.65) is at most as efficient as the OLS estimator in the restricted model (4.69), provided the restrictions are true. Therefore, adding additional variables that do not really belong in a model normally leads to less accurate estimates. Only in certain very special cases is there no loss of efficiency. In such cases, the covariance matrices of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ must be the same, which implies that the matrix difference computed in (4.72) must be zero.

The last expression in (4.72) is a zero matrix whenever $\boldsymbol{P_Z X} = \mathbf{O}$. This condition holds whenever the two sets of regressors $\boldsymbol{X}$ and $\boldsymbol{Z}$ are mutually orthogonal, so that $\boldsymbol{Z}^\top \boldsymbol{X} = \mathbf{O}$. In this special case, $\hat{\boldsymbol{\beta}}$ is just as efficient as $\tilde{\boldsymbol{\beta}}$. In general, however, including regressors that do not belong in a model increases the variance of the estimates of the coefficients on the regressors that do belong, and the increase can be very great in many cases. As can be seen from the left-hand side of (4.71), the variance of the estimated coefficient $\hat{\beta}$ associated with any regressor $\boldsymbol{x}$ is proportional to the inverse of the SSR from a regression of $\boldsymbol{x}$ on all the other regressors. The more other regressors there are, whether they truly belong in the model or not, the smaller is this SSR.

Underspecification

The opposite of overspecification is underspecification, in which we omit some variables that actually do appear in the DGP. To avoid any new notation, let us suppose that the model we estimate is (4.69), which yields the estimator $\tilde{\boldsymbol{\beta}}$, but that the DGP is really

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\mathbf{0}, \sigma_0^2 \mathbf{I}). \tag{4.73}$$

Thus the situation is precisely the opposite of the one considered above. The estimator $\hat{\boldsymbol{\beta}}$, based on regression (4.65), is now the "correct" one to use, while the estimator $\tilde{\boldsymbol{\beta}}$ is based on an underspecified model. It is clear that underspecification, unlike overspecification, *is* a form of misspecification, because the DGP (4.73) does not belong to the model (4.69).

The first point to recognize about $\tilde{\boldsymbol{\beta}}$ is that it is now, in general, biased. Substituting the right-hand side of (4.73) for $\boldsymbol{y}$ in (4.04), and taking expectations conditional on $\boldsymbol{X}$ and $\boldsymbol{Z}$, we find that

$$\begin{aligned} \mathrm{E}(\tilde{\boldsymbol{\beta}}) &= \mathrm{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{Z}\boldsymbol{\gamma}_0 + \boldsymbol{u})\big) \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0 + \mathrm{E}\big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\big) \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0. \end{aligned} \tag{4.74}$$

The second term in the last line of (4.74) is equal to zero only when $\boldsymbol{X}^\top \boldsymbol{Z} = \mathbf{O}$ or $\boldsymbol{\gamma}_0 = \mathbf{0}$. The first possibility arises when the two sets of regressors are mutually orthogonal, the second when (4.69) is not in fact underspecified. Except in these very special cases, $\tilde{\boldsymbol{\beta}}$ is generally biased. The magnitude of

Copyright © 2021, Russell Davidson and James G. MacKinnon

the bias depends on the parameter vector $\boldsymbol{\gamma}_0$ and on the $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices. Because this bias does not vanish as $n \to \infty$, $\tilde{\boldsymbol{\beta}}$ is also generally inconsistent.

Since $\tilde{\boldsymbol{\beta}}$ is biased, we cannot reasonably use its covariance matrix to evaluate its accuracy. Instead, we can use the **mean squared error matrix**, or **MSE matrix**, of $\tilde{\boldsymbol{\beta}}$. This matrix is defined as

$$\mathrm{MSE}(\tilde{\boldsymbol{\beta}}) \equiv \mathrm{E}\big((\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big). \tag{4.75}$$

The MSE matrix is equal to $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$ if $\tilde{\boldsymbol{\beta}}$ is unbiased, but not otherwise. For a scalar parameter $\tilde{\beta}$, the MSE is equal to the square of the bias plus the variance:

$$\mathrm{MSE}(\tilde{\beta}) = \big(\mathrm{E}(\tilde{\beta}) - \beta_0\big)^2 + \mathrm{Var}(\tilde{\beta}).$$

Thus, when we use MSE to evaluate the accuracy of an estimator, we are choosing to give equal weight to random errors and to systematic errors that arise from bias.[2]

From equations (4.74), we can see that

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}.$$

Therefore, $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ times itself transposed is equal to

$$\begin{aligned} &(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0^\top \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &+ (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0 \boldsymbol{u}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}\boldsymbol{\gamma}_0^\top \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \end{aligned}$$

The second term here has expectation $\sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and the third and fourth terms, one of which is the transpose of the other, have expectation zero. Thus we conclude that

$$\mathrm{MSE}(\tilde{\boldsymbol{\beta}}) = \sigma_0^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0^\top \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{4.76}$$

The first term is what the covariance matrix would be if we were estimating a correctly specified model, and the second term arises because the restricted estimator $\tilde{\boldsymbol{\beta}}$ is biased.

We would like to compare $\mathrm{MSE}(\tilde{\boldsymbol{\beta}})$, expression (4.76), with $\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(\hat{\boldsymbol{\beta}})$, which is given by expression (4.70). However, no unambiguous comparison is possible. The first term in (4.76) cannot be larger, in the matrix sense, than (4.70). If the bias is small, the second term must be small, and it may well be that $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$. However, if the bias is large, the second term is necessarily large, and $\tilde{\boldsymbol{\beta}}$ must be less efficient than $\hat{\boldsymbol{\beta}}$. Of course, it is quite possible that some parameters may be estimated more efficiently by $\tilde{\boldsymbol{\beta}}$ and others more efficiently by $\hat{\boldsymbol{\beta}}$.

[2] For a scalar parameter, it is common to report the square root of the MSE, called the **root mean squared error**, or **RMSE**, instead of the MSE itself.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Whether or not the restricted estimator $\tilde{\boldsymbol{\beta}}$ happens to be more efficient than the unrestricted estimator $\hat{\boldsymbol{\beta}}$, the covariance matrix for $\tilde{\boldsymbol{\beta}}$ that is calculated by a least-squares regression program is incorrect. The program attempts to estimate the first term in (4.76), but it ignores the second. However, s^2 is typically larger than σ_0^2 if some regressors have been incorrectly omitted. Thus, the program yields a biased estimate of the first term.

It is tempting to conclude that underspecification is a more severe problem than overspecification. After all, the former constitutes misspecification, but the latter does not. In consequence, as we have seen, underspecification leads to biased estimates and an estimated covariance matrix that may be severely misleading, while overspecification merely leads to inefficiency. Therefore, it would seem that we should always err on the side of overspecification. If all samples were extremely large, this might be a reasonable conclusion. The bias caused by underspecification does not go away as the sample size increases, but the variances of all consistent estimators tend to zero. Therefore, in sufficiently large samples, it makes sense to avoid underspecification at all costs. However, in samples of modest size, the gain in efficiency from omitting some variables, even if their coefficients are not actually zero, may be very large relative to the bias that is caused by their omission.

4.9 Measures of Goodness of Fit

A natural question to ask about any regression is: How well does it fit? There is more than one way to answer this question, and none of the answers may be entirely satisfactory in every case.

One possibility might be to use s, the estimated standard error of the regression. But s can be rather hard to interpret, since it depends on the scale of the y_t. When the regressand is in logarithms, however, s is meaningful and easy to interpret. Consider the loglinear model

$$\log y_t = \beta_1 + \beta_2 \log x_{t2} + \beta_3 \log x_{t3} + u_t. \tag{4.77}$$

As we saw in Section 2.3, this model can be obtained by taking logarithms of both sides of the model

$$y_t = \mathrm{e}^{\beta_1} x_{t2}^{\beta_2} x_{t3}^{\beta_3} \mathrm{e}^{u_t}. \tag{4.78}$$

The factor e^{u_t} is, for u_t small, approximately equal to $1 + u_t$. Thus the standard deviation of u_t in (4.77) is, approximately, the standard deviation of the proportional disturbance in the regression (4.78). Therefore, for any regression where the dependent variable is in logs, we can simply interpret $100s$, provided it is small, as an estimate of the percentage error in the regression.

When the regressand is not in logarithms, we could divide s by $\bar{y}$, the average of the y_t, or perhaps by the average absolute value of y_t if they were not all of the same sign. This would provide a measure of how large are the disturbances

Copyright © 2021, Russell Davidson and James G. MacKinnon

in the regression relative to the magnitude of the dependent variable. In many cases, $s/\bar{y}$ (for a model in levels) or s (for a model in logarithms) provides a useful measure of how well a regression fits. However, these measures are not entirely satisfactory. They are bounded from below, since they cannot be negative, but they are not bounded from above. Moreover, $s/\bar{y}$ is very hard to interpret if y_t can be either positive or negative.

A much more commonly used (and misused) measure of goodness of fit is the **coefficient of determination**, or $\boldsymbol{R^2}$. There are several versions of R^2. The most fundamental is the **uncentered $\boldsymbol{R^2}$**, denoted R_u^2, which is the ratio of the explained sum of squares (ESS) of the regression to the total sum of squares (TSS). Recall that, for the regression $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$,

$$\mathrm{TSS} = \|\boldsymbol{y}\|^2 = \|\boldsymbol{P_X y}\|^2 + \|\boldsymbol{M_X y}\|^2 = \mathrm{ESS} + \mathrm{SSR}.$$

This is a consequence of Pythagoras' Theorem; see equation (3.26). Thus

$$R_u^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\|\boldsymbol{P_X y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\|\boldsymbol{M_X y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\mathrm{SSR}}{\mathrm{TSS}} = \cos^2\theta, \tag{4.79}$$

where θ is the angle between $\boldsymbol{y}$ and $\boldsymbol{P_X y}$; see Figure 3.10. For any angle θ, we know that $-1 \leq \cos\theta \leq 1$. Consequently, $0 \leq R_u^2 \leq 1$. If the angle θ were zero, $\boldsymbol{y}$ and $\boldsymbol{X\hat{\beta}}$ would coincide, the residual vector $\hat{\boldsymbol{u}}$ would vanish, and we would have what is called a **perfect fit**, with $R_u^2 = 1$. At the other extreme, if $R_u^2 = 0$, the fitted value vector would vanish, and $\boldsymbol{y}$ would coincide with the residual vector $\hat{\boldsymbol{u}}$.

Because R_u^2 depends on the data only through the residuals and fitted values, it is invariant under nonsingular linear transformations of the regressors. In addition, because it is defined as a ratio, the value of R_u^2 is invariant to changes in the scale of $\boldsymbol{y}$. For example, we could change the units in which the regressand is measured from dollars to thousands of dollars without affecting the value of R_u^2.

The **centered $\boldsymbol{R^2}$**, denoted R_c^2, is much more commonly encountered than the uncentered one. For this version, all variables are centered, that is, expressed as deviations from their respective means, before ESS and TSS are calculated. The advantage of R_c^2 is that it is invariant to changes in the mean of the regressand. By adding a large enough constant to all the y_t, we could always make R_u^2 become arbitrarily close to 1, at least if the regression included a constant, since the SSR would stay the same and the TSS would increase without limit; see Exercise 4.30.

One important limitation of both versions of R^2 is that they are valid only if a regression model is estimated by least squares, since otherwise it would not be true that TSS = ESS + SSR. Moreover, the centered version is not valid if the regressors do not include a constant term or the equivalent, that is, if $\boldsymbol{\iota}$, the vector of 1s, does not belong to $\mathcal{S}(\boldsymbol{X})$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Another, possibly undesirable, feature of both R_u^2 and R_c^2 as measures of goodness of fit is that both increase whenever more regressors are added. To demonstrate this, we argue in terms of R_u^2, but the FWL Theorem can be used to show that the same results hold for R_c^2. Consider once more the restricted and unrestricted models, (4.69) and (4.65), respectively. Since both regressions have the same dependent variable, they have the same TSS. Thus the regression with the larger ESS must also have the larger R^2. The ESS from (4.65) is $\|\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{y}\|^2$ and that from (4.69) is $\|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}\|^2$, and so the difference between them is

$$\boldsymbol{y}^\top(\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}. \tag{4.80}$$

Clearly, $\mathcal{S}(\boldsymbol{X}) \subset \mathcal{S}(\boldsymbol{X},\boldsymbol{Z})$. Thus $\boldsymbol{P}_{\boldsymbol{X}}$ projects on to a subspace of the image of $\boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}}$. This implies that the matrix in the middle of (4.80), say $\boldsymbol{Q}$, is an orthogonal projection matrix; see Exercise 3.18. Consequently, (4.80) takes the form $\boldsymbol{y}^\top\boldsymbol{Q}\boldsymbol{y} = \|\boldsymbol{Q}\boldsymbol{y}\|^2 \geq 0$. The ESS from (4.65) is therefore no less than that from (4.69), and so the R^2 from (4.65) is no less than that from (4.69).

The R^2 can be modified so that adding additional regressors does not necessarily increase its value. If $\boldsymbol{\iota} \in \mathcal{S}(\boldsymbol{X})$, the centered R^2 can be written as

$$R_c^2 = 1 - \frac{\sum_{t=1}^n \hat{u}_t^2}{\sum_{t=1}^n (y_t - \bar{y})^2}. \tag{4.81}$$

The numerator of the second term is just the SSR. As we saw in Section 4.6, it has expectation $(n-k)\sigma_0^2$ under standard assumptions. The denominator is $n-1$ times an unbiased estimator of the variance of y_t about its true mean. As such, it has expectation $(n-1)\mathrm{Var}(y)$. Thus the second term of (4.81) can be thought of as the ratio of two biased estimators. If we replace these biased estimators by unbiased estimators, we obtain the **adjusted $\boldsymbol{R^2}$**,

$$\bar{R}^2 \equiv 1 - \frac{\frac{1}{n-k}\sum_{t=1}^n \hat{u}_t^2}{\frac{1}{n-1}\sum_{t=1}^n (y_t - \bar{y})^2} = 1 - \frac{(n-1)\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{(n-k)\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{\iota}}\boldsymbol{y}}. \tag{4.82}$$

The adjusted R^2 is reported by virtually all regression packages, often in preference to R_c^2. However, $\bar{R}^2$ is really no more informative than R_c^2. The two are generally very similar, except when $(n-k)/(n-1)$ is noticeably less than 1.

One nice feature of R_u^2 and R_c^2 is that they are constrained to lie between 0 and 1. In contrast, $\bar{R}^2$ can actually be negative. If a model has very little explanatory power, it is conceivable that $(n-1)/(n-k)$ may be greater than $\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{\iota}}\boldsymbol{y}/\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$. When that happens, $\bar{R}^2 < 0$.

The widespread use of $\bar{R}^2$ dates from the early days of econometrics, when sample sizes were often small, and investigators were easily impressed by models that yielded large values of R_c^2. As we saw above, adding an extra regressor to a linear regression always increases R_c^2. This increase can be quite noticeable when the sample size is small, even if the added regressor does not really

Copyright © 2021, Russell Davidson and James G. MacKinnon

belong in the regression. In contrast, adding an extra regressor increases $\bar{R}^2$ only if the proportional reduction in the SSR is greater than the proportional reduction in $n-k$. Therefore, a naive investigator who tries to maximize $\bar{R}^2$ is less likely to end up choosing a severely overspecified model than one who tries to maximize R_c^2.

It can be extremely misleading to compare any form of R^2 for models that are estimated using different data sets. Suppose, for example, that we estimate Model 1 using a set of data for which the regressors, and consequently the regressand, vary a lot, and we estimate Model 2 using a second set of data for which both the regressors and the regressand vary much less. Then, even if both models fit equally well, in the sense that their residuals have just about the same variance, Model 1 has a much larger R^2 than Model 2. This can most easily be seen from equation (4.81). Increasing the denominator of the second term while holding the numerator constant evidently increases the R^2.

4.10 Final Remarks

In this chapter, we have dealt with many of the best-known and most fundamental statistical properties of ordinary least squares. In particular, we discussed the properties of $\hat{\boldsymbol{\beta}}$ as an estimator of $\boldsymbol{\beta}$ in several sections and of s^2 as an estimator of σ_0^2 in Section 4.7. We introduced some of the key concepts of asymptotic analysis, including laws of large numbers, the same-order notation, and consistency in Section 4.3. We also proved the famous Gauss-Markov Theorem that ordinary least squares is the best linear unbiased estimator in Section 4.6.

In addition, we derived $\mathrm{Var}(\hat{\boldsymbol{\beta}})$, the covariance matrix of $\hat{\boldsymbol{\beta}}$, in Section 4.4, and we showed how to estimate it when the error covariance matrix is a scalar matrix in Section 4.7. However, we have not yet said anything about how to use $\hat{\boldsymbol{\beta}}$ and the estimate of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ to make inferences about $\boldsymbol{\beta}$. This important topic will be taken up in the next two chapters.

4.11 Exercises

4.1 Generate a sample of size 25 from the autoregressive model (4.14), with $\beta_1 = 1$ and $\beta_2 = 0.8$. For simplicity, assume that $y_0 = 0$ and that the u_t are NID(0, 1). Use this sample to compute the OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. Repeat at least 1000 times, and find the averages of the $\hat{\beta}_1$ and the $\hat{\beta}_2$. Use these averages to estimate the bias of the OLS estimators of β_1 and β_2.

Repeat this exercise for sample sizes of 50, 100, and 200. What happens to the bias of $\hat{\beta}_1$ and $\hat{\beta}_2$ as the sample size is increased?

4.2 Show that the unconditional distribution of y_t in the model (4.14) has expectation $\beta_1/(1-\beta_2)$ and variance $\sigma^2/(1-\beta_2^2)$. Repeat the simulations of

Copyright © 2021, Russell Davidson and James G. MacKinnon

Exercise 4.1 under the assumption that y_0 is drawn from a normal distribution with that expectation and variance. Are $\hat{\beta}_1$ and $\hat{\beta}_2$ more or less biased when the data are generated in this more realistic way?

4.3 Consider a sequence of random variables Y_t, $t = 1, \ldots, \infty$, which are such that $\mathrm{E}(Y_t) = \mu_t$. By considering the centered variables $Y_t - \mu_t$, show that the law of large numbers can be formulated as

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} Y_t = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mu_t.$$

4.4 Let the scalar random variable Y_n have CDF

$$F_n(x) = \begin{cases} 0 & \text{for } x < 0, \\ nx & \text{for } 0 \le x \le 1/n, \\ 1 & \text{for } x > 1/n. \end{cases}$$

Y_n is said to have the uniform distribution $\mathrm{U}(0, 1/n)$, since its density is constant and equal to n on the interval $[0, 1/n]$, and zero elsewhere.

Show that the sequence $\{Y_n\}$ converges in distribution. What is the limiting CDF F_∞? Show that F_∞ has a point of discontinuity at 0, and that $\lim_{n\to\infty} F_n(0) \neq F_\infty(0)$.

4.5 Consider the model (4.18). Show that the matrix $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ is singular under the asymptotic construction that uses the model unchanged for all sample sizes. In order to do so, it may be helpful to know that

$$\sum_{t=1}^{\infty} 1/t^2 = \pi^2/6.$$

4.6 Use the FWL Theorem to obtain an explicit expression for $\hat{\beta}_1$ in the model (4.18). Under the asymptotic construction in which the model is unchanged for all sample sizes, show that $\hat{\beta}_1$ has a nonstochastic plim equal to the true value of β_1.

4.7 Using the data on consumption and personal disposable income for the United States in the file **consumption-data.txt**, estimate the following model for the period 1948:1 to 2014:4:

$$c_t = \beta_1 + \beta_2 y_t + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2).$$

Here $c_t = \log C_t$ is the log of consumption, and $y_t = \log Y_t$ is the log of disposable income. Then use a random number generator and the estimates of β_1, β_2, and σ to obtain 268 simulated observations for c_t.

Begin by regressing your simulated log consumption variable on the log of income and a constant using just the first 4 observations. Save the estimates of β_2 and σ. Repeat this exercise for sample sizes of $5, 6, \ldots, 268$. Plot the estimates of β_2 and σ as functions of the sample size. What happens to these estimates as the sample size grows?

Repeat the complete exercise using a different set of simulated consumption data. Which features of the paths of the parameter estimates are common to the two experiments, and which are different?

Copyright © 2021, Russell Davidson and James G. MacKinnon

4.8 Plot the EDF (empirical distribution function) of the residuals from OLS estimation using one of the sets of simulated data, for the entire sample period, that you obtained in the last exercise; see Exercise 2.1 for a definition of the EDF. On the same graph, plot the CDF of the $\mathrm{N}(0, \sigma^2)$ distribution, where σ^2 now denotes the variance you used to simulate the log of consumption.

Show that the distributions characterized by the EDF and the normal CDF have the same expectation but different variances. How could you modify the residuals so that the EDF of the modified residuals would have the same variance, σ^2, as the normal CDF?

4.9 In Section 4.4, it is stated that the covariance matrix $\mathrm{Var}(\boldsymbol{b})$ of any random k–vector $\boldsymbol{b}$ is positive semidefinite. Prove this fact by considering arbitrary linear combinations $\boldsymbol{w}^\top\boldsymbol{b}$ of the components of $\boldsymbol{b}$ with nonrandom $\boldsymbol{w}$. If $\mathrm{Var}(\boldsymbol{b})$ is positive semidefinite without being positive definite, what can you say about $\boldsymbol{b}$?

4.10 For any pair of random variables, b_1 and b_2, show, by using the fact that the covariance matrix of $\boldsymbol{b} \equiv [b_1 \vdots b_2]$ is positive semidefinite, that

$$\mathrm{Cov}(b_1, b_2)^2 \le \mathrm{Var}(b_1)\mathrm{Var}(b_2).$$

Use this result to show that the correlation of b_1 and b_2 lies between -1 and 1.

4.11 Consider the linear regression model with n observations,

$$\boldsymbol{y} = \delta_1\boldsymbol{d}_1 + \delta_2\boldsymbol{d}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}). \tag{4.83}$$

The two regressors are dummy variables, with every element of $\boldsymbol{d}_2$ equal to 1 minus the corresponding element of $\boldsymbol{d}_1$. The vector $\boldsymbol{d}_1$ has n_1 elements equal to 1, and the vector $\boldsymbol{d}_2$ has $n_2 = n - n_1$ elements equal to 1.[3]

The parameter of interest is $\gamma \equiv \delta_2 - \delta_1$. Find the standard deviation of $\hat{\gamma}$ (that is, the positive square root of its true variance) and write it as a function of σ, n, and either n_1 or n_2.

Suppose the data for regression (4.83) come from an experiment that you design and administer. If you can only afford to collect 800 observations, how should you choose n_1 and n_2 in order to estimate γ as efficiently as possible?

4.12 Suppose that $\boldsymbol{X}$, a matrix of regressors that do not vary systematically with the sample size n, is added to regression (4.83), so that it becomes

$$\boldsymbol{y} = \delta_1\boldsymbol{d}_1 + \delta_2\boldsymbol{d}_2 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{4.84}$$

What is the true variance of $\hat{\gamma}$ in this case? Write this variance as a function of n and n_2 using same-order notation. Will this variance tend to 0 as $n \to \infty$ if n_2 is held fixed?

4.13 If $\boldsymbol{A}$ is a positive definite matrix, show that $\boldsymbol{A}^{-1}$ is also positive definite.

4.14 If $\boldsymbol{A}$ is a symmetric positive definite $k \times k$ matrix, then $\mathbf{I} - \boldsymbol{A}$ is positive definite if and only if $\boldsymbol{A}^{-1} - \mathbf{I}$ is positive definite, where $\mathbf{I}$ is the $k \times k$ identity

[3] Equations like (4.83) are frequently used to study **treatment effects**. $d_{t2} = 1$ corresponds to the t^{th} observation being treated, and $d_{t2} = 0$ (which implies that $d_{t1} = 1$) corresponds to its not being treated.

Copyright © 2021, Russell Davidson and James G. MacKinnon

matrix. Prove this result by considering the quadratic form $\boldsymbol{x}^\top(\mathbf{I}-\boldsymbol{A})\boldsymbol{x}$ and expressing $\boldsymbol{x}$ as $\boldsymbol{R}^{-1}\boldsymbol{z}$, where $\boldsymbol{R}$ is a symmetric matrix such that $\boldsymbol{A}=\boldsymbol{R}^2$.

Extend the above result to show that, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric positive definite matrices of the same dimensions, then $\boldsymbol{A}-\boldsymbol{B}$ is positive definite if and only if $\boldsymbol{B}^{-1}-\boldsymbol{A}^{-1}$ is positive definite.

4.15 Show that the variance of a sum of random variables z_t, $t=1,\dots,n$, with $\mathrm{Cov}(z_t,z_s)=0$ for $t\neq s$, equals the sum of their individual variances, whatever their expectations may be.

4.16 If $\gamma\equiv\boldsymbol{w}^\top\boldsymbol{\beta}=\sum_{i=1}^k w_i\beta_i$, show that $\mathrm{Var}(\hat\gamma)$, which is given by (4.44), can also be written as

$$\sum_{i=1}^k w_i^2\,\mathrm{Var}(\hat\beta_i)+2\sum_{i=2}^k\sum_{j=1}^{i-1}w_iw_j\,\mathrm{Cov}(\hat\beta_i,\hat\beta_j). \tag{4.85}$$

4.17 Use the result (4.58) on the variance of the OLS residual $\hat u_t$ to construct an unbiased estimating equation for the parameter σ^2 that is linear in σ^2. Show that solving this estimating equation yields the unbiased estimator of σ^2.

4.18 Using the data in the file **consumption-data.txt**, construct the variables c_t, the logarithm of consumption, and y_t, the logarithm of income, and their first differences $\Delta c_t\equiv c_t-c_{t-1}$ and $\Delta y_t\equiv y_t-y_{t-1}$. Use these data to estimate the following model for the period 1948:1 to 2014:4:

$$\Delta c_t=\beta_1+\beta_2\Delta y_t+\beta_3\Delta y_{t-1}+\beta_4\Delta y_{t-2}+u_t. \tag{4.86}$$

Let $\gamma=\sum_{i=2}^4\beta_i$. Calculate $\hat\gamma$ and its standard error in two different ways. One method should explicitly use the result (4.44), and the other should use a transformation of regression (4.86) which allows $\hat\gamma$ and its standard error to be read off directly from the regression output.

4.19 Using the data in the file **house-price-data.txt**, regress the logarithm of the price on a constant, the logarithm of lot size, and the variables `baths`, `sty`, `ffin`, `ca`, `gar`, and `reg`. What is s, the standard error of the regression?

Now estimate the model again using data for only the first 540 observations, and use those estimates to forecast the log prices for the last 6 observations. What are the standard errors of these forecasts? How are they related to the value of s for the regression with 540 observations?

Hint: It is possible to obtain both the forecasts and the standard errors by running a single regression with 546 observations and 14 regressors.

⋆4.20 Starting from equation (4.56) and using the result proved in Exercise 4.15, but without using (4.57), prove that, if $\mathrm{E}(u_t^2)=\sigma_0^2$ and $\mathrm{E}(u_su_t)=0$ for all $s\neq t$, then $\mathrm{Var}(\hat u_t)=(1-h_t)\sigma_0^2$. This is the result (4.58).

4.21 Use the result (4.58) to show that the MM estimator $\hat\sigma^2$ of (4.60) is consistent. You may assume that a LLN applies to the average in that equation.

4.22 Prove that $\mathrm{E}(\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{u}})=(n-k)\sigma_0^2$. This is the result (4.62). The proof should make use of the fact that the trace of a product of matrices is invariant to cyclic permutations; see Section 4.6.

Copyright © 2021, Russell Davidson and James G. MacKinnon

4.23 Consider two linear regressions, one restricted and the other unrestricted:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u},\ \text{and}$$
$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\gamma}+\boldsymbol{u}.$$

Show that, in the case of mutually orthogonal regressors, with $\boldsymbol{X}^\top\boldsymbol{Z}=\mathbf{O}$, the estimates of $\boldsymbol{\beta}$ from the two regressions are identical.

4.24 Suppose that you use the OLS estimates $\hat{\boldsymbol{\beta}}$, obtained by regressing the $n\times1$ vector $\boldsymbol{y}$ on the $n\times k$ matrix $\boldsymbol{X}$, to forecast the $n_*\times1$ vector $\boldsymbol{y}_*$ using the $n_*\times k$ matrix $\boldsymbol{X}_*$. Assuming that the disturbances, both within the sample used to estimate the parameters $\boldsymbol{\beta}$ and outside the sample in the forecast period, are $\mathrm{IID}(0,\sigma^2)$, and that the model is correctly specified, what is the covariance matrix of the vector of forecast errors?

4.25 The class of estimators considered by the Gauss-Markov Theorem can be written as $\hat{\boldsymbol{\beta}}=\boldsymbol{A}\boldsymbol{y}$, with $\boldsymbol{A}\boldsymbol{X}=\mathbf{I}$. Show that this class of estimators is in fact identical to the class of estimators of the form

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{W}^\top\boldsymbol{y}, \tag{4.87}$$

where $\boldsymbol{W}$ is a matrix of exogenous variables such that $\boldsymbol{W}^\top\boldsymbol{X}$ is nonsingular.

4.26 Show that the estimator (4.87) is unchanged if $\boldsymbol{W}$ is replaced by any other matrix $\boldsymbol{W}'$ of the same dimensions such that $\boldsymbol{P}_{\boldsymbol{W}}=\boldsymbol{P}_{\boldsymbol{W}'}$, or, equivalently, such that $\mathcal{S}(\boldsymbol{W})=\mathcal{S}(\boldsymbol{W}')$. In particular, show that the estimator (4.87) is the OLS estimator if $\boldsymbol{P}_{\boldsymbol{X}}=\boldsymbol{P}_{\boldsymbol{W}}$.

4.27 Show that the difference between the unrestricted estimator $\hat{\boldsymbol{\beta}}$ of model (4.65) and the restricted estimator $\tilde{\boldsymbol{\beta}}$ of model (4.69) is given by

$$\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}.$$

Hint: In order to prove this result, it is convenient to premultiply the difference by $(\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}$.

4.28 Consider the linear regression model

$$y_t=\beta_1+\beta_2x_{t2}+\beta_3x_{t3}+u_t.$$

Explain how you could estimate this model subject to the restriction that $\beta_2+\beta_3=1$ by running a regression that imposes the restriction. Also, explain how you could estimate the unrestricted model in such a way that the value of one of the coefficients would be zero if the restriction held exactly for your data.

4.29 Prove that, for a linear regression model with a constant term, the uncentered R_u^2 is always greater than the centered R_c^2.

4.30 Consider a linear regression model for a dependent variable y_t that has a sample mean of 17.21. Suppose that we create a new variable $y_t'=y_t+10$ and run the same linear regression using y_t' instead of y_t as the regressand. How are R_c^2, R_u^2, and the estimate of the constant term related in the two regressions? What if instead $y_t'=y_t-10$?

Copyright © 2021, Russell Davidson and James G. MacKinnon

4.31 Using the data in the file **consumption-data.txt**, construct the variables c_t, the logarithm of consumption, and y_t, the logarithm of income. Use them to estimate the following two models for the period 1948:1 to 2014:4:

$$c_t = \beta_1 + \beta_2 y_t + u_t, \text{ and} \tag{4.88}$$
$$\Delta c_t = \gamma_1 + \gamma_2 \Delta y_t + v_t. \tag{4.89}$$

Here Δ denotes the first difference operator, so that $\Delta c_t \equiv c_t - c_{t-1}$, and $\Delta y_t \equiv y_t - y_{t-1}$. Which model has the larger R^2? In your opinion, which model has the better fit? Explain.

4.32 Using the data in the file **consumption-data.txt**, construct the variables c_t, the logarithm of consumption, and y_t, the logarithm of income. Use them to estimate, for the period 1948:1 to 2014:4, the following **autoregressive distributed lag**, or **ADL**, model:

$$c_t = \alpha + \beta c_{t-1} + \gamma_0 y_t + \gamma_1 y_{t-1} + u_t. \tag{4.90}$$

Such models are often expressed in first-difference form, that is, as

$$\Delta c_t = \delta + \phi c_{t-1} + \theta \Delta y_t + \psi y_{t-1} + u_t, \tag{4.91}$$

where the first-difference operator Δ is defined so that $\Delta c_t = c_t - c_{t-1}$. Estimate the first-difference model (4.91), and then, without using the results of (4.90), rederive the estimates of α, β, γ_0, and γ_1 solely on the basis of your results from (4.91).

4.33 Simulate model (4.90) of the previous question, using your estimates of α, β, γ_0, γ_1, and the variance σ^2 of the disturbances. Perform the simulation conditional on the income series and the first observation c_1 of consumption. Plot the residuals from running (4.90) on the simulated data, and compare the plot with that of the residuals from the real data. Comments?

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 5

Hypothesis Testing in Linear Regression Models

5.1 Introduction

As we saw in Section 4.2, the vector of OLS parameter estimates $\hat{\boldsymbol{\beta}}$ is a random vector. Since it would be an astonishing coincidence if $\hat{\boldsymbol{\beta}}$ were equal to the true parameter vector $\boldsymbol{\beta}_0$ in any finite sample, we must take the randomness of $\hat{\boldsymbol{\beta}}$ into account if we are to make inferences about $\boldsymbol{\beta}$. In classical econometrics, the two principal ways of doing this are performing **hypothesis tests** and constructing **confidence intervals** or, more generally, **confidence regions**. We discuss hypothesis testing in this chapter and confidence intervals in the next one. We start with hypothesis testing because it typically plays a larger role than confidence intervals in applied econometrics and because it is essential to have a thorough grasp of hypothesis testing if the construction of confidence intervals is to be understood at anything more than a very superficial level.

In the next section, we develop the fundamental ideas of hypothesis testing in the context of a very simple special case. In Section 5.3, we review some of the properties of a number of important distributions, all related to the standard normal distribution, which are commonly encountered in the context of hypothesis testing. This material is needed for Section 5.4, in which we develop a number of results about hypothesis tests in the classical normal linear model. In Section 5.5, we relax some of the assumptions of that model and develop the asymptotic theory of linear regression models. That theory is then used to study large-sample tests in Section 5.6.

The remainder of the chapter deals with more advanced topics. In Section 5.7, we discuss some of the rather tricky issues associated with performing two or more tests at the same time. In Section 5.8, we discuss the **power** of a test, that is, what determines the ability of a test to reject a hypothesis that is false. Finally, in Section 5.9, we introduce the important concept of **pretesting**, in which the results of a test are used to determine which of two or more estimators to use.

5.2 Basic Ideas

When we conduct hypothesis tests, we must do so in the context of a model. The hypotheses considered in this chapter are about a parameter or parameters of the model. As with the parameter estimators considered in previous chapters, we work with a data set, consisting of observations on a dependent variable and some explanatory variables.

The very simplest sort of hypothesis test concerns the (population) mean from which a **random sample** has been drawn. By saying that the sample is random, we mean that the observations are independent and identically distributed (IID), and are realizations drawn from some underlying distribution. The term "population mean", borrowed from biostatistics, here refers simply to the expectation of that distribution.

Suppose that we wish to test the hypothesis that the expectation is equal to some value that we specify. A suitable model for this test is the following regression model

$$y_t = \beta + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{5.01}$$

where y_t is an observation on the dependent variable, β is the expectation of each of the y_t, and is the only parameter of the regression function, and σ^2 is the variance of the disturbance u_t. Let β_0 be the specified value of the expectation, so that we can express the hypothesis to be tested as $\beta = \beta_0$.[1]

The least-squares estimator of β is just the sample mean. If we denote it by $\hat{\beta}$, then it follows that, for a sample of size n,

$$\hat{\beta} = \frac{1}{n}\sum_{t=1}^{n} y_t \qquad \text{and} \qquad \text{Var}(\hat{\beta}) = \frac{1}{n}\sigma^2. \tag{5.02}$$

These formulas can either be obtained from first principles or as special cases of the general results for OLS estimation. In this case, the regressor matrix $\boldsymbol{X}$ is just $\boldsymbol{\iota}$, an n–vector of 1s. Thus, for the model (5.01), the standard formulas $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ and $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ yield the two formulas given in (5.02).

The hypothesis to be tested is called, for historical reasons, the **null hypothesis**. It is often given the label H_0 for short. In order to test H_0, we need a **test statistic**, which is a random variable that has a known distribution when the null hypothesis is true and some other distribution when the null hypothesis is false. If the value of this test statistic is one that might frequently be encountered by chance under the null hypothesis, then the test provides no evidence against the null. On the other hand, if the value of the test statistic is an extreme one that would rarely be encountered by chance under the null, then the test does provide evidence against the null. If this evidence is sufficiently convincing, we may decide to **reject** the null hypothesis that $\beta = \beta_0$.

[1] It may be slightly confusing that a 0 subscript is used here to denote the value of a parameter under the hypothesis being tested as well as its true value. So long as it is assumed that the hypothesis is true, however, there should be no possible confusion.

Copyright © 2021, Russell Davidson and James G. MacKinnon

For the moment, we will restrict the model (5.01) by making two very strong assumptions. The first is that u_t is normally distributed, and the second is that σ is known. Under these assumptions, a test of the hypothesis that $\beta = \beta_0$ can be based on the test statistic

$$z = \frac{\hat{\beta} - \beta_0}{\big(\text{Var}(\hat{\beta})\big)^{1/2}} = \frac{n^{1/2}}{\sigma}(\hat{\beta} - \beta_0). \tag{5.03}$$

It turns out that, under the null hypothesis, z is distributed as $\text{N}(0,1)$. It has expectation 0 because $\hat{\beta}$ is an unbiased estimator of β, and $\beta = \beta_0$ under the null. It has variance unity because, by (5.02),

$$\text{E}(z^2) = \frac{n}{\sigma^2}\text{E}\big((\hat{\beta} - \beta_0)^2\big) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1.$$

Finally, to see that z is normally distributed, note that $\hat{\beta}$ is just the average of the y_t, each of which is normally distributed if the corresponding u_t is; see Exercise 2.7. As we will see in the next section, this implies that z is also normally distributed. Thus z has the first property that we would like a test statistic to possess: It has a known distribution under the null hypothesis.

For every null hypothesis there is, at least implicitly, an **alternative hypothesis**, which is often given the label H_1. The alternative hypothesis is what we are testing the null against. Note that, if we consider the model that results from imposing the condition of the null hypothesis on the model (5.01), we get

$$y_t - \beta_0 = u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

The parameter β does not appear in this model; rather it appears only in the model (5.01), in which the null hypothesis is not imposed. In this case, the model (5.01) represents the alternative hypothesis, which can be thought of as providing a framework in which the null hypothesis can be expressed as a restriction on one or more parameters, here just $\beta = \beta_0$. As important as the fact that z has the $\text{N}(0,1)$ distribution under the null is the fact that z does *not* have this distribution under the alternative. Suppose that β takes on some other value, say β_1. Then it is clear that $\hat{\beta} = \beta_1 + \hat{\gamma}$, where $\hat{\gamma}$ has expectation 0 and variance σ^2/n; recall equation (4.05). In fact, $\hat{\gamma}$ is normal under our assumption that the u_t are normal, just like $\hat{\beta}$, and so $\hat{\gamma} \sim \text{N}(0, \sigma^2/n)$. It follows that z is also normal (see Exercise 2.7 again), and we find from (5.03) that

$$z \sim \text{N}(\lambda, 1), \quad \text{with} \quad \lambda = \frac{n^{1/2}}{\sigma}(\beta_1 - \beta_0). \tag{5.04}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The expectation λ is called the **non-centrality parameter**, or **NCP** of the distribution of z. Provided n is large enough, we would expect λ to be large and positive if $\beta_1 > \beta_0$ and large and negative if $\beta_1 < \beta_0$. Thus we reject the null hypothesis whenever z is sufficiently far from 0. Just how we can decide what "sufficiently far" means will be discussed shortly.

If we want to test the null that $\beta = \beta_0$ against the alternative that $\beta \neq \beta_0$, we must perform a **two-tailed test** and reject the null whenever the absolute value of z is sufficiently large. If instead we were interested in testing the null hypothesis that $\beta \leq \beta_0$ against the alternative that $\beta > \beta_0$, we would perform a **one-tailed test** and reject the null whenever z was sufficiently large and positive. In general, tests of equality restrictions are two-tailed tests, and tests of inequality restrictions are one-tailed tests.

Since z is a random variable that can, in principle, take on any value on the real line, no value of z is absolutely incompatible with the null hypothesis, and so we can never be absolutely certain that the null hypothesis is false. One way to deal with this situation is to decide in advance on a **rejection rule**, according to which we choose to reject the null hypothesis if and only if the value of z falls into the **rejection region** of the rule. For two-tailed tests, the appropriate rejection region is the union of two sets, one containing all values of z greater than some positive value, the other all values of z less than some negative value. For a one-tailed test, the rejection region would consist of just one set, containing either sufficiently positive or sufficiently negative values of z, according to the sign of the inequality we wish to test.

A test statistic combined with a rejection rule is generally simply called a **test**. A test returns a binary result, namely, reject or do-not-reject. We can never reach a conclusion that a null hypothesis is true on the basis of statistical evidence, and so our conclusion if a test fails to reject must simply be that the test provides no evidence against the null. Other tests, or other data sets, may well provide strong evidence against it.

If the test incorrectly leads us to reject a null hypothesis that is true, we are said to make a **Type I error**. The probability of making such an error is, by construction, the probability, *under the null hypothesis*, that z falls into the rejection region. A property of any given test is its **significance level**, or just **level**, and it is defined as the probability, under the null, of making a Type I error, that is, the probability of rejecting the null when it is true. A common notation for this is α. Like all probabilities, α is a number between 0 and 1, although, in practice, it is generally chosen to be much closer to 0 than 1. Popular values of α include .05 and .01.

In order to construct the rejection region for a test at level α based on the test statistic z, the first step is to calculate the **critical value** associated with the level α. We begin with the simplest case, which is when we want to test an inequality restriction of the form $\beta \geq \beta_0$. Evidence against this null is provided by a value of z that is negative and large enough in absolute value. The rejection region is thus an infinite interval containing everything to the

Copyright © 2021, Russell Davidson and James G. MacKinnon

left of the critical value appropriate for level α, say c_α. In order to attain this level, the probability under the null of a realization in this interval must be α. We continue to suppose that z is standard normal under the null, and so the critical value c_α has to satisfy the equation

$$\Phi(c_\alpha) = \alpha; \tag{5.05}$$

recall that Φ denotes the CDF of the standard normal distribution. We can solve (5.05) in terms of the inverse function Φ^{-1}, and we find that

$$c_\alpha = \Phi^{-1}(\alpha).$$

Note that, for $\alpha < 1/2$, c_α, being in the left-hand tail of the standard normal distribution, is negative.

In order to test an equality restriction, we use a two-tailed test. This means that we need both a negative and a positive critical value. In this case, the commonest sort of test is an **equal-tail test**, with the same probability mass in the rejection regions on the left and on the right. For level α, then, that means that we want a probabilty mass of $\alpha/2$ in both tails. For the left-hand tail, we must have

$$\Phi(-c_\alpha) = \alpha/2,$$

We know that the left-tail critical value is negative, hence the minus sign. On account of the symmetry of the N(0,1) distribution, the right-tail critical value is just $+c_\alpha$. We could equally well have defined c_α by the equation

$$\Phi(c_\alpha) = 1 - \alpha/2, \tag{5.06}$$

which allocates a probability mass of $\alpha/2$ to the right of c_α. Solving equation (5.06) for c_α gives

$$c_\alpha = \Phi^{-1}(1 - \alpha/2). \tag{5.07}$$

Clearly, the critical value c_α increases as α approaches 0. As an example, when $\alpha = .05$, we see from equation (5.07) that the critical value for a two-tailed test is $\Phi^{-1}(.975) = 1.96$. We would reject the null at the .05 level whenever the observed absolute value of the test statistic exceeds 1.96.

Until now, we have assumed that the distribution of the test statistic under the null hypothesis is known exactly, so that we have what is called an **exact test**. In econometrics, however, the distribution of a test statistic is often known only approximately. In this case, we need to draw a distinction between the **nominal level** of the test, that is, the probability of making a Type I error according to whatever approximate distribution we are using to determine the rejection region, and the actual **rejection probability**, which may differ greatly

Copyright © 2021, Russell Davidson and James G. MacKinnon

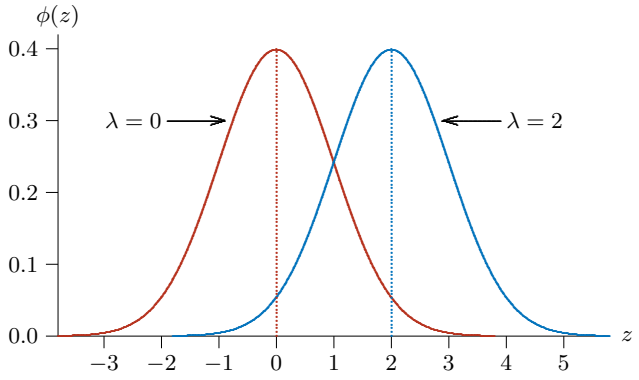


Figure 5.1 The normal distribution centered and uncentered

from the nominal level. The rejection probability is generally unknowable in practice, because it typically depends on unknown features of the DGP.[2]

The probability that a test rejects the null is called the **power** of the test. If the data are generated by a DGP that satisfies the null hypothesis, the power of an exact test is equal to its level. In general, power depends on precisely how the data were generated and on the sample size. We can see from (5.04) that the distribution of z is entirely determined by the value of the non-centrality parameter λ, with $\lambda = 0$ under the null, and that the value of λ depends on the parameters of the DGP. In this example, λ is proportional to $\beta_1 - \beta_0$ and to the square root of the sample size, and it is inversely proportional to σ.

Values of λ different from 0 move the probability mass of the $\mathrm{N}(\lambda, 1)$ distribution away from the center of the $\mathrm{N}(0, 1)$ distribution and into its tails. This can be seen in Figure 5.1, which graphs the $\mathrm{N}(0, 1)$ density and the $\mathrm{N}(\lambda, 1)$ density for $\lambda = 2$. The second density places much more probability than the first on values of z greater than 2. Thus, if the rejection region for our test were the interval from 2 to $+\infty$, there would be a much higher probability in that region for $\lambda = 2$ than for $\lambda = 0$. Therefore, we would reject the null hypothesis more often when the null hypothesis is false, with $\lambda = 2$, than when it is true, with $\lambda = 0$.

Mistakenly failing to reject a false null hypothesis is called making a **Type II error**. The probability of making such a mistake is equal to 1 minus the power of the test. It is not hard to see that, quite generally, the probability

[2] Another term that often arises in the discussion of hypothesis testing is the **size** of a test. Technically, this is the supremum of the rejection probability over all DGPs that satisfy the null hypothesis. For an exact test, the size equals the level. For an approximate test, the size is typically difficult or impossible to calculate. It is often, but by no means always, greater than the nominal level of the test.

Copyright © 2021, Russell Davidson and James G. MacKinnon

of rejecting the null with a two-tailed test based on z increases with the absolute value of λ. Consequently, the power of such a test increases as $\beta_1 - \beta_0$ increases, as σ decreases, and as the sample size increases. We will discuss what determines the power of a test in more detail in Section 5.8.

P Values

As we have defined it, the result of a test is yes or no: Reject or do not reject. The result depends on the chosen level, and it is to that extent subjective: different people can be expected to have different tolerances for Type I error. A more sophisticated approach to deciding whether or not to reject the null hypothesis is to calculate the ***P* value**, or **marginal significance level**, associated with a test statistic. The P value for the statistic z is defined as the greatest level for which a test based on z fails to reject the null. Equivalently, at least if the statistic z has a continuous distribution, it is the smallest level for which the test rejects. Thus, the test rejects for all levels greater than the P value, and it fails to reject for all levels smaller than the P value. The P value is given as a deterministic function of the (random) statistic z by finding the level for which z is equal to the critical value for that level. Therefore, if the P value determined by z is denoted $p(z)$, we must be prepared to accept a probability $p(z)$ of Type I error if we choose to reject the null. But the P value itself, being a purely objective quantity, is the same for everyone, and it allows different people to draw their own subjective conclusions.

The consequences of the definition of the P value are a little trickier for the equal-tail test we have been discussing. We find that

$$p(z) = 2\big(1 - \Phi(|z|)\big). \tag{5.08}$$

To see this, note that the test based on z rejects at level α if and only if $|z| > c_\alpha$. This inequality is equivalent to $\Phi(|z|) > \Phi(c_\alpha)$, because $\Phi(\cdot)$ is a strictly increasing function. Further, for this equal-tail test, $\Phi(c_\alpha) = 1 - \alpha/2$, by equation (5.06). The smallest value of α for which the inequality holds is thus obtained by solving the equation

$$\Phi(|z|) = 1 - \alpha/2,$$

and the solution is easily seen to be the right-hand side of equation (5.08).

One advantage of using P values is that they preserve all the information conveyed by a test statistic, while presenting it in a way that is directly interpretable. For example, the test statistics 2.02 and 5.77 would both lead us to reject the null at the .05 level using a two-tailed test. The second of these obviously provides more evidence against the null than does the first, but it is only after they are converted to P values that the magnitude of the difference becomes apparent. The P value for the first test statistic is .0434, while the P value for the second is 7.93×10^{-9}, an extremely small number.

Copyright © 2021, Russell Davidson and James G. MacKinnon

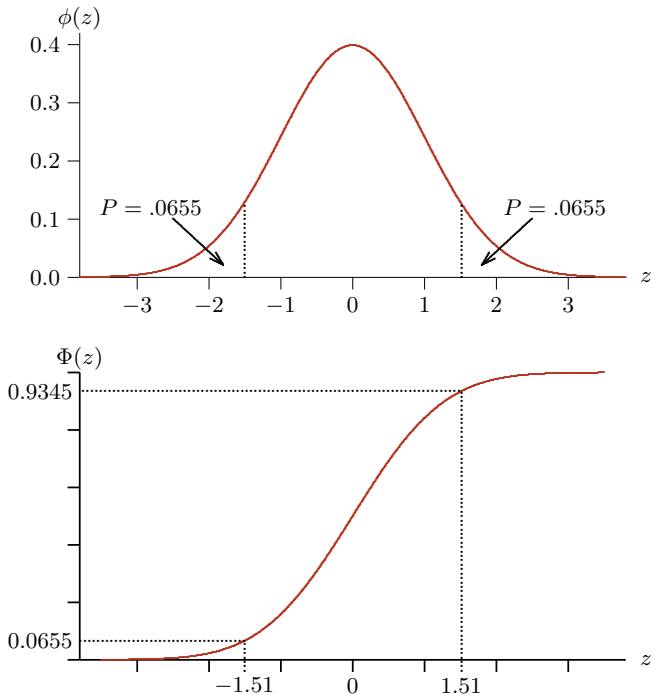


Figure 5.2 P values for a two-tailed test

Computing a P value transforms z from a random variable with the $N(0,1)$ distribution into a new random variable $p(z)$ with the uniform $U(0,1)$ distribution. In Exercise 5.1, readers are invited to prove this fact. It is quite possible to think of $p(z)$ as a test statistic, of which the observed realization is $p(\hat{z})$ whenever the realization of z is $\hat{z}$. A test at level α rejects whenever $p(\hat{z}) < \alpha$. Note that the sign of this inequality is the opposite of that in the condition $|\hat{z}| > c_\alpha$. Generally, one rejects for *large* values of test statistics, but for *small* P values.

Figure 5.2 illustrates how the test statistic z is related to its P value $p(z)$. Suppose that the value of the test statistic is 1.51. Then

$$\Pr(z > 1.51) = \Pr(z < -1.51) = .0655. \tag{5.09}$$

This implies, by equation (5.08), that the P value for an equal-tail test based on z is .1310. The top panel of the figure illustrates (5.09) in terms of the standard normal density, and the bottom panel illustrates it in terms of the CDF. To avoid clutter, no critical values are shown on the figure, but it is clear that a test based on z does not reject at any level smaller than .131. From the figure, it is also easy to see that the P value for a one-tailed test of

Copyright © 2021, Russell Davidson and James G. MacKinnon

the hypothesis that $\beta \leq \beta_0$ is .0655. This is just $\Pr(z > 1.51)$. Similarly, the P value for a one-tailed test of the hypothesis that $\beta \geq \beta_0$ is $\Pr(z < 1.51) = .9345$.

The P values discussed above, whether for one-tailed or two-tailed tests, are based on the symmetric $N(0,1)$ distribution. In Exercise 5.19, readers are asked to show how to compute P values for two-tailed tests based on an asymmetric distribution.

In this section, we have introduced the basic ideas of hypothesis testing. However, we had to make two very restrictive assumptions. The first is that the disturbances are normally distributed, and the second, which is grossly unrealistic, is that the variance of the disturbances is known. In addition, we limited our attention to a single restriction on a single parameter. In Section 5.4, we will discuss the more general case of linear restrictions on the parameters of a linear regression model with unknown disturbance variance. Before we can do so, however, we need to review the properties of the normal distribution and of several distributions that are closely related to it.

5.3 Some Common Distributions

Most test statistics in econometrics have one of four well-known distributions, at least approximately. These are the standard normal distribution, the chi-squared (or χ^2) distribution, Student's t distribution, and the F distribution. The most basic of these is the normal distribution, since the other three distributions can be derived from it. In this section, we discuss the standard, or **central**, versions of these distributions. Later, in Section 5.8, we will have occasion to introduce **noncentral** versions of all these distributions.

The Normal Distribution

The **normal distribution**, which is sometimes called the **Gaussian distribution** in honor of the celebrated German mathematician and astronomer Carl Friedrich Gauß (1777–1855)[3], even though he did not invent it, is certainly the most famous distribution in statistics. As we saw in Section 2.2, there is a whole family of normal distributions, all based on the **standard normal distribution**, so called because it has expectation 0 and variance 1.

The density of the standard normal distribution, which is usually denoted by $\phi(\cdot)$, was defined in equation (2.06). No elementary closed-form expression exists for its CDF, which is usually denoted by $\Phi(\cdot)$. Although there is no closed form, it is perfectly easy to evaluate Φ numerically, and virtually every program for econometrics and statistics can do this. Thus it is straightforward

[3] Here, but only here, we have used the German spelling of the name that, elsewhere, is conventionally spelt "Gauss".

Copyright © 2021, Russell Davidson and James G. MacKinnon

to compute the P value for any test statistic that is distributed as standard normal. The graphs of the functions ϕ and Φ were first shown in Figure 2.1 and have just reappeared in Figure 5.2. In both tails, as can be seen in the top panel of the figure, the density rapidly approaches 0. Thus, although a standard normal r.v. can, in principle, take on any value on the real line, values greater than about 4 in absolute value occur extremely rarely.

The full normal family of distributions of scalar random variables is what is called a **location-scale** family. Any such family can be generated by varying two parameters, the expectation, which is the location parameter, and the variance.[4] The base member of the family is usually chosen to have expectation zero and variance, and so also standard deviation, one. For the normal family, the base is the standard normal distribution.

A random variable X that is normally distributed with expectation μ and variance σ^2 can be generated by the formula

$$X = \mu + \sigma Z, \tag{5.10}$$

where Z is standard normal. The distribution of X, that is, the normal distribution with expectation μ and variance σ^2, is denoted $\mathrm{N}(\mu, \sigma^2)$. Thus the standard normal distribution is the $\mathrm{N}(0, 1)$ distribution. As readers were asked to show in Exercise 2.8, the density of the $\mathrm{N}(\mu, \sigma^2)$ distribution, evaluated at x, is

$$\frac{1}{\sigma}\phi\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big). \tag{5.11}$$

The formulas (5.10) and (5.11) illustrate an important point. In principle, there are two equivalent ways of characterizing a probability distribution. One is analytic, where the CDF or density of the distribution is given; this is the case with (5.11). The other is to provide a **recipe for simulation**, by which the distribution is specified by means of a random variable that has that distribution, and is defined by a formula, usually one involving either uniform U(0,1) or standard normal N(0,1) variables; that is the case with (5.10). Although these two sorts of characterization are entirely equivalent in principle, it is often much easier to implement the recipe for a simulation, and it may be difficult to derive an analytic formula from the recipe. In what follows in this section, we rely exclusively on recipes for simulation.

In expression (5.10), as in Section 2.2, we have distinguished between the random variable X and a value x that it can take on. However, for the following discussion, this distinction is more confusing than illuminating. For the rest of this section, we therefore use lower-case letters to denote both random variables and the arguments of their densities or CDFs, depending on

[4] It is the standard deviation, the square root of the variance, that is called the scale parameter.

Copyright © 2021, Russell Davidson and James G. MacKinnon

context. No confusion should result. Adopting this convention, then, we see that, if x is distributed as $\mathrm{N}(\mu, \sigma^2)$, we can invert equation (5.10) and obtain $z = (x - \mu)/\sigma$, where z is standard normal. Note also that z is the argument of ϕ in the expression (5.11) of the density of x. In general, the density of a normal variable x with expectation μ and variance σ^2 is $1/\sigma$ times ϕ evaluated at the corresponding standard normal variable, which is $z = (x - \mu)/\sigma$.

Although the normal distribution is fully characterized by its first two moments, the higher moments are also important. Because the distribution is symmetric around its expectation, the third central moment, which measures the **skewness** of the distribution, is always zero.[5] This is true for all of the odd central moments. The fourth moment of a symmetric distribution provides a way to measure its **kurtosis**, which essentially means how thick the tails are. In the case of the $\mathrm{N}(\mu, \sigma^2)$ distribution, the fourth central moment is $3\sigma^4$; see Exercise 5.2.

The Multivariate Normal Distribution

As its name suggests, the **multivariate normal distribution** is a family of distributions for random *vectors*, with the scalar normal distribution being a special case of it. When the vector has just two members, the two random variables are said to have the **bivariate normal distribution**. As we will see in a moment, all these distributions, like the scalar normal distribution, are completely characterized by their first two moments.

In order to construct the multivariate normal distribution, we begin with a set of m mutually independent standard normal variables, z_i, $i = 1, \ldots, m$, which we can assemble into a random m–vector. We write $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, where the m-vector $\mathbf{0}$ is the vector of the expectations of the components of $\boldsymbol{z}$, and $\mathbf{I}$ is the $m \times m$ covariance matrix, because by construction the variance of each component of $\boldsymbol{z}$ is 1, and, since the z_i are mutually independent, all the covariances are 0; see Exercise 2.13. Then, by definition, any m–vector $\boldsymbol{x}$ of linearly independent linear combinations of the components of $\boldsymbol{z}$ has a multivariate normal distribution. Such a vector $\boldsymbol{x}$ can always be written as $\boldsymbol{Az}$, for some nonsingular nonrandom $m \times m$ matrix $\boldsymbol{A}$.

One of the most important properties of the multivariate normal distribution, which we used in Section 5.2, is that any linear combination of the elements of a multivariate normal vector is itself normally distributed. The proof of this result, even for the special case in which the random variables are independent, requires some effort. It is therefore relegated to an appendix; see Section 5.11.

We denote the components of $\boldsymbol{x}$ as x_i, $i = 1, \ldots, m$. From the result proved in Section 5.11, it follows that each x_i is normally distributed, with (unconditional) expectation zero. Therefore, from results proved in Section 4.4, we

[5] A distribution is said to be skewed to the right if the third central moment is positive, and to the left if the third central moment is negative.

Copyright © 2021, Russell Davidson and James G. MacKinnon

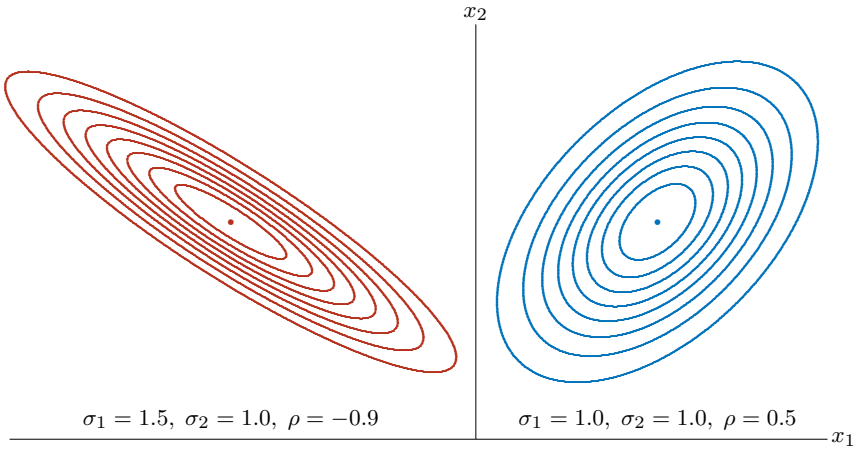


Figure 5.3 Contours of two bivariate normal densities

can see that the covariance matrix of $\boldsymbol{x}$ is

$$\mathrm{Var}(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^\top) = \boldsymbol{A}\mathrm{E}(\boldsymbol{z}\boldsymbol{z}^\top)\boldsymbol{A}^\top = \boldsymbol{A}\mathbf{I}\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top.$$

Here we have used the fact that the covariance matrix of $\boldsymbol{z}$ is the identity matrix $\mathbf{I}$.

Let us denote the covariance matrix of $\boldsymbol{x}$ by $\boldsymbol{\Omega}$. Recall that, according to a result mentioned in Section 4.4 in connection with Crout's algorithm, for any positive definite matrix $\boldsymbol{\Omega}$, we can always find a lower-triangular matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Omega}$. Thus, in order to construct a vector $\boldsymbol{x}$ with covariance matrix $\boldsymbol{\Omega}$, we may always choose the matrix $\boldsymbol{A}$ to be lower-triangular. The distribution of $\boldsymbol{x}$ is multivariate normal with expectation vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$. We write this as $\boldsymbol{x} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$. If we add an m–vector $\boldsymbol{\mu}$ of constants to $\boldsymbol{x}$, the resulting vector has the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ distribution.

It is clear from this argument that any linear combination of random variables that are jointly multivariate normal must itself be normally distributed. Thus, if $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, any scalar $\boldsymbol{a}^\top\boldsymbol{x}$, where $\boldsymbol{a}$ is an m–vector of fixed coefficients, is normally distributed with expectation $\boldsymbol{a}^\top\boldsymbol{\mu}$ and variance $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}$.

We saw above that $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ whenever the components of the vector $\boldsymbol{z}$ are independent. Another crucial property of the multivariate normal distribution is that the converse of this result is also true: If $\boldsymbol{x}$ is any multivariate normal vector with zero covariances, the components of $\boldsymbol{x}$ are mutually independent. This is a very special property of the multivariate normal distribution, and readers are asked to prove it, for the bivariate case, in Exercise 5.5. In general, a zero covariance between two random variables does *not* imply that they are independent.

Copyright © 2021, Russell Davidson and James G. MacKinnon

It is important to note that the results of the last two paragraphs do not hold unless the vector $\boldsymbol{x}$ is *multivariate* normal, that is, constructed as a set of linear combinations of *independent* normal variables. In most cases, when we have to deal with linear combinations of two or more normal random variables, it is reasonable to assume that they are jointly distributed as multivariate normal. However, as Exercise 2.14 illustrates, it is possible for two or more random variables not to be multivariate normal even though each one individually has a normal distribution.

Figure 5.3 illustrates the bivariate normal distribution, of which the density is given in Exercise 5.5 in terms of the variances σ_1^2 and σ_2^2 of the two variables, and their correlation ρ. Contours of the density are plotted, on the right for $\sigma_1 = \sigma_2 = 1.0$ and $\rho = 0.5$, on the left for $\sigma_1 = 1.5$, $\sigma_2 = 1.0$, and $\rho = -0.9$. The contours of the bivariate normal density can be seen to be elliptical. The ellipses slope upward when $\rho > 0$ and downward when $\rho < 0$. They do so more steeply the larger is the ratio σ_2/σ_1. The closer $|\rho|$ is to 1, for given values of σ_1 and σ_2, the more elongated are the elliptical contours.

The Chi-Squared Distribution

Suppose, as in our discussion of the multivariate normal distribution, that the random vector $\boldsymbol{z}$ is such that its components $z_1, \ldots, z_m$ are mutually independent standard normal random variables, that is, $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. Then the random variable

$$y \equiv \|\boldsymbol{z}\|^2 = \boldsymbol{z}^\top\boldsymbol{z} = \sum_{i=1}^{m} z_i^2 \tag{5.12}$$

is said to have the **chi-squared distribution** with m **degrees of freedom**. A compact way of writing this is: $y \sim \chi^2(m)$. From (5.12), it is clear that m must be a positive integer. In the case of a test statistic, it will turn out to be equal to the number of restrictions being tested.

The expectation and variance of the $\chi^2(m)$ distribution can easily be obtained from the definition (5.12). The expectation is

$$\mathrm{E}(y) = \sum_{i=1}^{m} \mathrm{E}(z_i^2) = \sum_{i=1}^{m} 1 = m. \tag{5.13}$$

Since the z_i are independent, the variance of the sum of the z_i^2 is just the sum of the (identical) variances:

$$\begin{aligned} \mathrm{Var}(y) &= \sum_{i=1}^{m} \mathrm{Var}(z_i^2) = m\mathrm{E}\big((z_i^2 - 1)^2\big) \\ &= m\mathrm{E}(z_i^4 - 2z_i^2 + 1) = m(3 - 2 + 1) = 2m. \end{aligned} \tag{5.14}$$

The third equality here uses the fact that $\mathrm{E}(z_i^4) = 3$; see Exercise 5.2.

Copyright © 2021, Russell Davidson and James G. MacKinnon

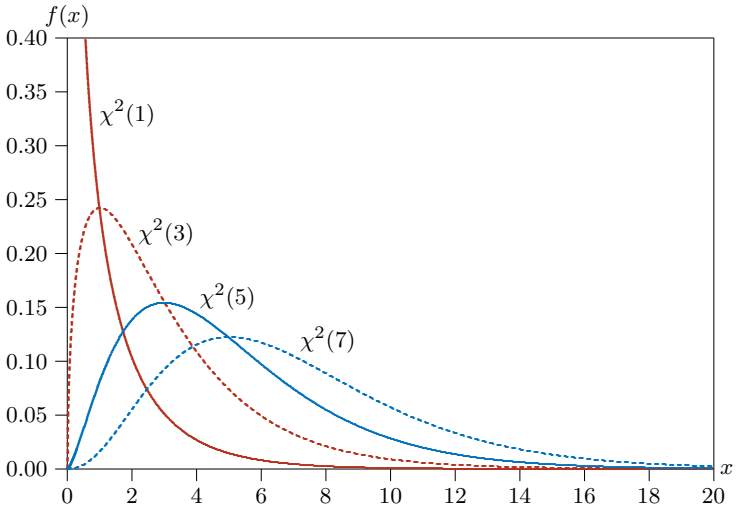


Figure 5.4 Various chi-squared PDFs

Another important property of the chi-squared distribution, which follows immediately from (5.12), is that, if $y_1 \sim \chi^2(m_1)$ and $y_2 \sim \chi^2(m_2)$, and y_1 and y_2 are independent, then $y_1 + y_2 \sim \chi^2(m_1 + m_2)$. To see this, rewrite (5.12) as

$$y = y_1 + y_2 = \sum_{i=1}^{m_1} z_i^2 + \sum_{i=m_1+1}^{m_1+m_2} z_i^2 = \sum_{i=1}^{m_1+m_2} z_i^2,$$

from which the result follows.

Figure 5.4 shows the density of the $\chi^2(m)$ distribution for $m = 1$, $m = 3$, $m = 5$, and $m = 7$. The changes in the location and height of the density function as m increases are what we should expect from the results (5.13) and (5.14) about its expectation and variance. In addition, the density, which is extremely skewed to the right for $m = 1$, becomes less skewed as m increases. In fact, as we will see in Section 5.5, the $\chi^2(m)$ distribution approaches the $\mathrm{N}(m, 2m)$ distribution as m becomes large.

In Section 4.4, we introduced quadratic forms. As we will see, many test statistics can be written as quadratic forms in normal vectors, or as functions of such quadratic forms. The following theorem states two results about quadratic forms in normal vectors that will prove to be extremely useful.

Theorem 5.1.

1. If the m–vector $\boldsymbol{x}$ is distributed as $\mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$, then the quadratic form $\boldsymbol{x}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{x}$ is distributed as $\chi^2(m)$;

Copyright © 2021, Russell Davidson and James G. MacKinnon

2. If $\boldsymbol{P}$ is an $n \times n$ orthogonal projection matrix with rank $r < n$ and $\boldsymbol{z}$ is an n–vector that is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$, then the quadratic form $\boldsymbol{z}^\top \boldsymbol{P} \boldsymbol{z}$ is distributed as $\chi^2(r)$.

Proof:

Since the covariance matrix $\boldsymbol{\Omega}$ is positive definite, as before we can find an $m \times m$ nonrandom nonsingular matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Omega}$. Since the vector $\boldsymbol{x}$ is multivariate normal with expectation vector $\mathbf{0}$, so is the vector $\boldsymbol{A}^{-1}\boldsymbol{x}$. Moreover, the covariance matrix of $\boldsymbol{A}^{-1}\boldsymbol{x}$ is

$$\mathrm{E}\big(\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{A}^\top)^{-1}\big) = \boldsymbol{A}^{-1}\boldsymbol{\Omega}(\boldsymbol{A}^\top)^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^\top(\boldsymbol{A}^\top)^{-1} = \mathbf{I}_m.$$

Thus we have shown that the vector $\boldsymbol{z} \equiv \boldsymbol{A}^{-1}\boldsymbol{x}$ is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$.

The quadratic form $\boldsymbol{x}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{x}$ is equal to $\boldsymbol{x}^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{A}^{-1}\boldsymbol{x} = \boldsymbol{z}^\top\boldsymbol{z}$. As we have just shown, this is equal to the sum of m independent, squared, standard normal random variables. From the definition of the chi-squared distribution, we know that such a sum is distributed as $\chi^2(m)$. This proves the first part of the theorem.

Since $\boldsymbol{P}$ is an orthogonal projection matrix, it projects orthogonally on to some subspace of E^n. Suppose, then, that $\boldsymbol{P}$ projects on to the span of the columns of an $n \times r$ matrix $\boldsymbol{Z}$. This allows us to write

$$\boldsymbol{z}^\top \boldsymbol{P} \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{z}.$$

The r–vector $\boldsymbol{x} \equiv \boldsymbol{Z}^\top \boldsymbol{z}$ evidently has the $\mathrm{N}(\mathbf{0}, \boldsymbol{Z}^\top\boldsymbol{Z})$ distribution. Therefore, $\boldsymbol{z}^\top \boldsymbol{P} \boldsymbol{z}$ is seen to be a quadratic form in the multivariate normal r–vector $\boldsymbol{x}$ and $(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}$, which is the inverse of its covariance matrix. That this quadratic form is distributed as $\chi^2(r)$ follows immediately from the the first part of the theorem. ■

Student's t Distribution

If $z \sim \mathrm{N}(0, 1)$ and $y \sim \chi^2(m)$, and z and y are independent, then the random variable

$$t \equiv \frac{z}{(y/m)^{1/2}} \tag{5.15}$$

is said to have **Student's t distribution** with m degrees of freedom. A compact way of writing this is: $t \sim t(m)$. The density of Student's t distribution looks very much like that of the standard normal distribution, since both are bell-shaped and symmetric around 0.[6]

[6] "Student" was the pen name of W. S. Gosset, who worked for the Guinness brewery in Dublin. He used a pen name because he did not want his employers to know that he was wasting his time on statistics.

Copyright © 2021, Russell Davidson and James G. MacKinnon

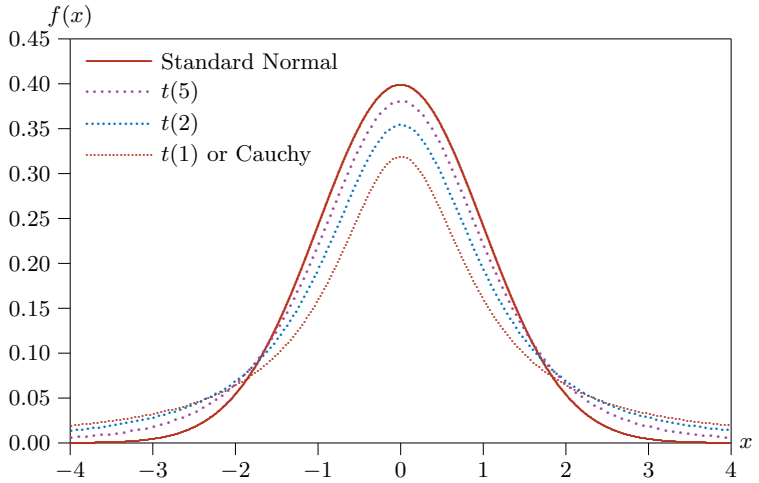


Figure 5.5 PDFs of the Student's t distribution

The moments of the t distribution depend on m, and only the first $m-1$ moments exist. Thus the $t(1)$ distribution, which is also called the **Cauchy distribution**, has no moments at all, and the $t(2)$ distribution has no variance. From (5.15), we see that, for the Cauchy distribution, the denominator of t is just the absolute value of a standard normal random variable. Whenever this denominator happens to be close to zero, the ratio is likely to be a very big number, even if the numerator is not particularly large. Thus the Cauchy distribution has very thick tails. As m increases, the chance that the denominator of (5.15) is close to zero diminishes (see Figure 5.4), and so the tails become thinner.

In general, if t is distributed as $t(m)$ with $m > 2$, then $\mathrm{Var}(t) = m/(m-2)$. Thus, as $m \to \infty$, the variance tends to 1, the variance of the standard normal distribution. In fact, the entire $t(m)$ distribution tends to the standard normal distribution as $m \to \infty$. By (5.12), the chi-squared variable y can be expressed as $\sum_{i=1}^{m} z_i^2$, where the z_i are independent standard normal variables. Therefore, by a law of large numbers, such as (4.22), y/m, which is the average of the z_i^2, tends to its expectation as $m \to \infty$. By (5.13), this expectation is just $m/m = 1$. It follows that the denominator of (5.15), $(y/m)^{1/2}$, also tends to 1, and hence that $t \to z \sim \mathrm{N}(0,1)$ as $m \to \infty$. Note that we are dealing here with convergence in distribution.

Figure 5.5 shows the densities of the standard normal, $t(1)$, $t(2)$, and $t(5)$ distributions. In order to make the differences among the various densities in the figure apparent, all the values of m are chosen to be very small. However, it is clear from the figure that, for larger values of m, the density of $t(m)$ is very similar to the density of the standard normal distribution.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The F Distribution

If y_1 and y_2 are independent random variables distributed as $\chi^2(m_1)$ and $\chi^2(m_2)$, respectively, then the random variable

$$F \equiv \frac{y_1/m_1}{y_2/m_2} \tag{5.16}$$

is said to have the ***F* distribution** with m_1 and m_2 degrees of freedom. A compact way of writing this is: $F \sim F(m_1, m_2)$.[7] The $F(m_1, m_2)$ distribution looks a lot like a rescaled version of the $\chi^2(m_1)$ distribution. As for the t distribution, the denominator of (5.16) tends to unity as $m_2 \to \infty$, and so $m_1 F \to y_1 \sim \chi^2(m_1)$ as $m_2 \to \infty$. Therefore, for large values of m_2, a random variable that is distributed as $F(m_1, m_2)$ behaves very much like $1/m_1$ times a random variable that is distributed as $\chi^2(m_1)$.

The F distribution is very closely related to Student's t distribution. It is evident from (5.16) and (5.15) that the square of a random variable which is distributed as $t(m_2)$ is distributed as $F(1, m_2)$. In the next section, we will see how these two distributions arise in the context of hypothesis testing in linear regression models.

5.4 Exact Tests in the Classical Normal Linear Model

In the example of Section 5.2, we were able to obtain a test statistic z that is distributed as $\mathrm{N}(0,1)$. Tests based on this statistic are exact. Unfortunately, it is possible to perform exact tests only in certain special cases. One very important special case of this type arises when we test linear restrictions on the parameters of the classical normal linear model, which was introduced in Section 4.1. This model may be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \tag{5.17}$$

where $\boldsymbol{X}$ is an $n \times k$ matrix of regressors, so that there are n observations and k regressors, and it is assumed that the disturbance vector $\boldsymbol{u}$ is statistically independent of the matrix $\boldsymbol{X}$. Notice that in (5.17) the assumption which in Section 4.1 was written as $u_t \sim \mathrm{NID}(0, \sigma^2)$ is now expressed in matrix notation using the multivariate normal distribution. In addition, since the assumption that $\boldsymbol{u}$ and $\boldsymbol{X}$ are independent means that the generating process for $\boldsymbol{X}$ is independent of that for $\boldsymbol{y}$, we can express this independence assumption by saying that the regressors $\boldsymbol{X}$ are **exogenous** in the model (5.17); the concept of exogeneity[8] was introduced in Section 2.3 and discussed in Section 4.2.

[7] The F distribution was introduced by Snedecor (1934). The notation F is used in honor of the well-known statistician R. A. Fisher.

[8] This assumption is usually called **strict exogeneity** in the literature, but, since we will not discuss any other sort of exogeneity in this book, it is convenient to drop the word "strict."

Copyright © 2021, Russell Davidson and James G. MacKinnon

Tests of a Single Restriction

We begin by considering a single, linear restriction on $\boldsymbol{\beta}$. This could, in principle, be any sort of linear restriction, for example, that $\beta_1 = 5$ or $\beta_3 = \beta_4$. However, it simplifies the analysis, and involves no loss of generality, if we confine our attention to a restriction that one of the coefficients should equal 0. If a restriction does not naturally have the form of a zero restriction, we can always apply suitable linear transformations to $\boldsymbol{y}$ and $\boldsymbol{X}$, of the sort considered in Sections 2.3 and 2.4, in order to rewrite the model so that it does; see Exercise 5.8 and Exercise 5.9.

Let us partition $\boldsymbol{\beta}$ as $[\boldsymbol{\beta}_1 \vdots \beta_2]$, where $\boldsymbol{\beta}_1$ is a $(k-1)$–vector and β_2 is a scalar, and consider a restriction of the form $\beta_2 = 0$. When $\boldsymbol{X}$ is partitioned conformably with $\boldsymbol{\beta}$, the model (5.17) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{5.18}$$

where $\boldsymbol{X}_1$ denotes an $n \times (k-1)$ matrix and $\boldsymbol{x}_2$ denotes an n–vector, with $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{x}_2]$.

By the FWL Theorem, the least-squares estimate of β_2 from (5.18) is the same as the least-squares estimate from the FWL regression

$$\boldsymbol{M}_1\boldsymbol{y} = \beta_2\boldsymbol{M}_1\boldsymbol{x}_2 + \text{residuals}, \tag{5.19}$$

where $\boldsymbol{M}_1 \equiv \mathbf{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ is the matrix that projects on to $\mathcal{S}^\perp(\boldsymbol{X}_1)$. By applying the standard formulas for the OLS estimator and covariance matrix to regression (5.19), under the assumption that the model (5.18) is correctly specified, we find that

$$\hat{\beta}_2 = \frac{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{y}}{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2} \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_2) = \sigma^2(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{-1}.$$

In order to test the hypothesis that β_2 equals any specified value, say β_2^0, we have to subtract β_2^0 from $\hat{\beta}_2$ and divide by the square root of the variance. For the null hypothesis that $\beta_2 = 0$, this yields a test statistic analogous to (5.03),

$$z_{\beta_2} \equiv \frac{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{y}}{\sigma(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}}, \tag{5.20}$$

which can be computed only under the unrealistic assumption that σ is known.

If the data are actually generated by the model (5.18) with $\beta_2 = 0$, then

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}) = \boldsymbol{M}_1\boldsymbol{u}.$$

Therefore, the right-hand side of equation (5.20) becomes

$$\frac{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{u}}{\sigma(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}}. \tag{5.21}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

It is now easy to see that z_{β_2} is distributed as $\mathrm{N}(0,1)$. Because we can condition on $\boldsymbol{X}$, the only thing left in (5.21) that is stochastic is $\boldsymbol{u}$. Since the numerator is just a linear combination of the components of $\boldsymbol{u}$, which is multivariate normal, the entire test statistic is normally distributed. The variance of the numerator is

$$\begin{aligned}\mathrm{E}(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{M}_1\boldsymbol{x}_2) &= \boldsymbol{x}_2^\top\boldsymbol{M}_1\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{M}_1\boldsymbol{x}_2 \\ &= \boldsymbol{x}_2^\top\boldsymbol{M}_1\sigma^2\mathbf{I}\boldsymbol{M}_1\boldsymbol{x}_2 = \sigma^2\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2.\end{aligned}$$

Since the denominator of (5.21) is just the square root of the variance of the numerator, we conclude that z_{β_2} is distributed as $\mathrm{N}(0,1)$ under the null hypothesis.

The test statistic z_{β_2} defined in equation (5.20) has exactly the same distribution under the null hypothesis as the test statistic z defined in (5.03). The analysis of Section 5.2 therefore applies to it without any change. Thus we now know how to test the hypothesis that any coefficient in the classical normal linear model is equal to 0, or to any specified value, but only if we know the variance of the disturbances.

In order to handle the more realistic case in which the variance of the disturbances is unknown, we need to replace σ in equation (5.20) by s, the standard error of the regression (5.18), which was defined in equation (4.63). If, as usual, $\boldsymbol{M}_{\boldsymbol{X}}$ is the orthogonal projection on to $\mathcal{S}^\perp(\boldsymbol{X})$, then we have $s^2 = \boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}/(n-k)$, and so we obtain the test statistic

$$t_{\beta_2} \equiv \frac{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{y}}{s(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}} = \left(\frac{\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{n-k}\right)^{-1/2}\frac{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{y}}{(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}}. \tag{5.22}$$

As we will now demonstrate, this test statistic is distributed as $t(n-k)$ under the null hypothesis. Not surprisingly, it is called a ***t* statistic**.

As we discussed in the last section, for a test statistic to have the $t(n-k)$ distribution, it must be possible to write it as the ratio of a standard normal variable z to the square root of $\zeta/(n-k)$, where ζ is independent of z and distributed as $\chi^2(n-k)$. The t statistic defined in (5.22) can be rewritten as

$$t_{\beta_2} = \frac{z_{\beta_2}}{\left(\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}/((n-k)\sigma^2)\right)^{1/2}}, \tag{5.23}$$

which has the form of such a ratio. We have already shown that $z_{\beta_2} \sim \mathrm{N}(0,1)$. Thus it only remains to show that $\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}/\sigma^2 \sim \chi^2(n-k)$ and that the random variables in the numerator and denominator of (5.23) are independent.

Under any DGP that belongs to (5.18),

$$\frac{\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{\sigma^2} = \frac{\boldsymbol{u}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}}{\sigma^2} = \boldsymbol{\varepsilon}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}, \tag{5.24}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma$ is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$. Since $\boldsymbol{M}_{\boldsymbol{X}}$ is a projection matrix with rank $n - k$, the second part of Theorem 5.1 shows that the rightmost expression in (5.24) is distributed as $\chi^2(n-k)$.

To see that the random variables z_{β_2} and $\boldsymbol{\varepsilon}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{\varepsilon}$ are independent, we note first that $\boldsymbol{\varepsilon}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{\varepsilon}$ depends on $\boldsymbol{y}$ only through $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$. Second, from (5.20), it is not hard to see that z_{β_2} depends on $\boldsymbol{y}$ only through $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y}$, since

$$\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{x}_2^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{x}_2^\top (\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{P}_1) \boldsymbol{y} = \boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y};$$

the first equality here simply uses the fact that $\boldsymbol{x}_2 \in \mathcal{S}(\boldsymbol{X})$, and the third equality uses the result (3.35) that $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{P}_1 = \boldsymbol{P}_1 \boldsymbol{P}_{\boldsymbol{X}}$. Independence now follows because, as we will see directly, $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y}$ and $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$ are independent.

We saw above that $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$. Further, from (5.17), $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{u}$, from which it follows that the centered version of $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y}$ is $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{u}$. The $n \times n$ matrix of covariances of the components of $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{u}$ and $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$ is thus

$$\mathrm{E}(\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{u} \boldsymbol{u}^\top \boldsymbol{M}_{\boldsymbol{X}}) = \sigma^2 \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{M}_{\boldsymbol{X}} = \mathbf{O},$$

by (3.25), because $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}$ are complementary projections. These zero covariances imply that the vectors $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{u}$ and $\boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}$ are independent, since both are multivariate normal. Geometrically, these vectors have zero covariance because they lie in *orthogonal* subspaces, namely, the images of $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}$. Thus, even though the numerator and denominator of (5.23) both depend on $\boldsymbol{y}$, this orthogonality implies that they are independent.

We therefore conclude that the t statistic (5.23) for $\beta_2 = 0$ in the model (5.18) has the $t(n-k)$ distribution. Performing one-tailed and two-tailed tests based on t_{β_2} is almost the same as performing them based on z_{β_2}. We just have to use the $t(n-k)$ distribution instead of the $\mathrm{N}(0,1)$ distribution to compute P values or critical values. An interesting property of t statistics is explored in Exercise 5.10.

Tests of Several Restrictions

Economists frequently want to test more than one linear restriction. Let us suppose that there are r restrictions, with $r \leq k$, since there cannot be more equality restrictions than there are parameters in the unrestricted model. As before, there is no loss of generality if we assume that the restrictions take the form $\boldsymbol{\beta}_2 = \mathbf{0}$. The alternative hypothesis is the model (5.17), which has been rewritten as

$$H_1: \quad \boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{5.25}$$

Here $\boldsymbol{X}_1$ is an $n \times k_1$ matrix, $\boldsymbol{X}_2$ is an $n \times k_2$ matrix, $\boldsymbol{\beta}_1$ is a k_1–vector, $\boldsymbol{\beta}_2$ is a k_2–vector, $k = k_1 + k_2$, and the number of restrictions $r = k_2$. Unless $r = 1$, it is no longer possible to use a t test, because there is one t statistic

Copyright © 2021, Russell Davidson and James G. MacKinnon

for each element of $\boldsymbol{\beta}_2$, and we want to compute a single test statistic for all the restrictions at once.

It is natural to base a test on a comparison of how well the model fits when the restrictions are imposed with how well it fits when they are not imposed. The null hypothesis is the regression model

$$H_0: \quad \boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{5.26}$$

in which we impose the restriction that $\boldsymbol{\beta}_2 = \mathbf{0}$. As we saw in Section 4.8, the restricted model (5.26) must always fit worse than the unrestricted model (5.25), in the sense that the SSR from (5.26) cannot be smaller, and is almost always larger, than the SSR from (5.25). However, if the restrictions are true, the reduction in SSR from adding $\boldsymbol{X}_2$ to the regression should be relatively small. Therefore, it seems natural to base a test statistic on the difference between these two SSRs. If USSR denotes the **unrestricted sum of squared residuals**, from (5.25), and RSSR denotes the **restricted sum of squared residuals**, from (5.26), the appropriate test statistic is

$$F_{\boldsymbol{\beta}_2} \equiv \frac{(\mathrm{RSSR} - \mathrm{USSR})/r}{\mathrm{USSR}/(n-k)}. \tag{5.27}$$

Under the null hypothesis, as we will now demonstrate, this test statistic has the F distribution with r and $n - k$ degrees of freedom. Not surprisingly, it is called an ***F* statistic**.

The restricted SSR is $\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{y}$, and the unrestricted one is $\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$. One way to obtain a convenient expression for the difference between these two expressions is to use the FWL Theorem. By this theorem, the USSR is the SSR from the FWL regression

$$\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \text{residuals}. \tag{5.28}$$

The total sum of squares from (5.28) is $\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{y}$. The explained sum of squares can be expressed in terms of the orthogonal projection on to the r–dimensional subspace $\mathcal{S}(\boldsymbol{M}_1 \boldsymbol{X}_2)$, and so the difference is

$$\mathrm{USSR} = \boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{M}_1 \boldsymbol{y}. \tag{5.29}$$

Therefore, since $\boldsymbol{M}_1 \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{M}_1 = \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2}$,

$$\mathrm{RSSR} - \mathrm{USSR} = \boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{y},$$

and the F statistic (5.27) can be written as

$$F_{\boldsymbol{\beta}_2} = \frac{\boldsymbol{y}^\top \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{y}/r}{\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}/(n-k)}. \tag{5.30}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

In general, $\boldsymbol{M_X y} = \boldsymbol{M_X u}$. Under the null hypothesis, $\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{M}_1 \boldsymbol{u}$. Thus, under this hypothesis, the F statistic (5.30) reduces to

$$\frac{\boldsymbol{\varepsilon}^\top \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{\varepsilon} / r}{\boldsymbol{\varepsilon}^\top \boldsymbol{M_X} \boldsymbol{\varepsilon} / (n-k)}, \tag{5.31}$$

where, as before, $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma$. We saw in the last subsection that the quadratic form in the denominator of (5.31) is distributed as $\chi^2(n-k)$. Since the quadratic form in the numerator can be written as $\boldsymbol{\varepsilon}^\top \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{\varepsilon}$, it is distributed as $\chi^2(r)$. Moreover, the random variables in the numerator and denominator are independent, because $\boldsymbol{M_X}$ and $\boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2}$ project on to mutually orthogonal subspaces: $\boldsymbol{M_X M}_1 \boldsymbol{X}_2 = \boldsymbol{M_X}(\boldsymbol{X}_2 - \boldsymbol{P}_1 \boldsymbol{X}_2) = \boldsymbol{O}$. Thus it is apparent that the statistic (5.31) has the $F(r, n-k)$ distribution under the null hypothesis.

A Threefold Orthogonal Decomposition

Each of the restricted and unrestricted models generates an orthogonal decomposition of the dependent variable $\boldsymbol{y}$. It is illuminating to see how these two decompositions interact to produce a threefold orthogonal decomposition. It turns out that all three components of this decomposition have useful interpretations. From the two models, we find that

$$\boldsymbol{y} = \boldsymbol{P}_1 \boldsymbol{y} + \boldsymbol{M}_1 \boldsymbol{y} \quad \text{and} \quad \boldsymbol{y} = \boldsymbol{P_X y} + \boldsymbol{M_X y}. \tag{5.32}$$

In Exercises 3.18 and 3.19, $\boldsymbol{P_X} - \boldsymbol{P}_1$ was seen to be an orthogonal projection matrix, equal to $\boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2}$. It follows that

$$\boldsymbol{P_X} = \boldsymbol{P}_1 + \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2}, \tag{5.33}$$

where the two projections on the right-hand side of this equation are obviously mutually orthogonal, since $\boldsymbol{P}_1$ annihilates $\boldsymbol{M}_1 \boldsymbol{X}_2$. From (5.32) and (5.33), we obtain the threefold orthogonal decomposition

$$\boldsymbol{y} = \boldsymbol{P}_1 \boldsymbol{y} + \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{y} + \boldsymbol{M_X y}. \tag{5.34}$$

The first term is the vector of fitted values from the restricted model, $\boldsymbol{X}_1 \tilde{\boldsymbol{\beta}}_1$. In this and what follows, we use a tilde ($\tilde{}$) to denote the **restricted estimates**, and a hat ($\hat{}$) to denote the **unrestricted estimates**. The second term is the vector of fitted values from the FWL regression (5.28). It equals $\boldsymbol{M}_1 \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2$, where, by the FWL Theorem, $\hat{\boldsymbol{\beta}}_2$ is a subvector of estimates from the unrestricted model. Finally, $\boldsymbol{M_X y}$ is the vector of residuals from the unrestricted model.

Since $\boldsymbol{P_X y} = \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2$, the vector of fitted values from the unrestricted model, we see that

$$\boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1 \tilde{\boldsymbol{\beta}}_1 + \boldsymbol{M}_1 \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2. \tag{5.35}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

In Exercise 5.11, this result is exploited to show how to obtain the restricted estimates in terms of the unrestricted estimates.

The F statistic (5.30) can be written as the ratio of the squared norm of the second component in (5.34) to the squared norm of the third, each normalized by the appropriate number of degrees of freedom. Under both hypotheses, the third component, $\boldsymbol{M_X y}$, equals $\boldsymbol{M_X u}$, and so it just consists of random noise. Its squared norm is a $\chi^2(n-k)$ variable times σ^2, which serves as the (unrestricted) estimate of σ^2 and can be thought of as a measure of the scale of the random noise. Since $\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, every element of $\boldsymbol{u}$ has the same variance, and so every component of (5.34), if centered so as to leave only the random part, should have the same scale.

Under the null hypothesis, the second component is $\boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{u}$, which just consists of random noise. But, under the alternative, $\boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{y} = \boldsymbol{M}_1 \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{P}_{\boldsymbol{M}_1 \boldsymbol{X}_2} \boldsymbol{u}$, and it thus contains a systematic part related to $\boldsymbol{X}_2$. The length of the second component must be greater, on average, under the alternative than under the null, since the random part is there in all cases, but the systematic part is present only under the alternative. The F test compares the squared length of the second component with the squared length of the third. It thus serves to detect the possible presence of systematic variation, related to $\boldsymbol{X}_2$, in the second component of (5.34).

We want to reject the null whenever RSSR − USSR, the numerator of the F statistic, is relatively large. Consequently, the P value corresponding to a realized F statistic $F_{\boldsymbol{\beta}_2}$ is computed as $1 - F_{r,n-k}(F_{\boldsymbol{\beta}_2})$, where $F_{r,n-k}(\cdot)$ denotes the CDF of the F distribution with r and $n-k$ degrees of freedom. Although we compute the P value as if for a one-tailed test, F tests are really two-tailed tests, because they test equality restrictions, not inequality restrictions. An F test for $\boldsymbol{\beta}_2 = \boldsymbol{0}$ rejects the null hypothesis whenever $\hat{\boldsymbol{\beta}}_2$ is sufficiently far from $\boldsymbol{0}$, whether the individual elements of $\hat{\boldsymbol{\beta}}_2$ are positive or negative.

There is a very close relationship between F tests and t tests. In the previous section, we saw that the square of a random variable with the $t(n-k)$ distribution has the $F(1, n-k)$ distribution. The square of the t statistic t_{β_2}, defined in (5.22), is

$$t^2_{\beta_2} = \frac{\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{x}_2 (\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{-1} \boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{M_X y} / (n-k)}.$$

This test statistic is evidently a special case of (5.30), with the vector $\boldsymbol{x}_2$ replacing the matrix $\boldsymbol{X}_2$. Thus, when there is only one restriction, it makes no difference whether we use a two-tailed t test or an F test.

An Example of the F Test

The most familiar application of the F test is testing the hypothesis that all the coefficients in a classical normal linear model, except the constant term,

Copyright © 2021, Russell Davidson and James G. MacKinnon

are zero. The null hypothesis is that $\boldsymbol{\beta}_2 = \mathbf{0}$ in the model

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{5.36}$$

where $\boldsymbol{\iota}$ is an n–vector of 1s and $\boldsymbol{X}_2$ is $n \times (k-1)$. In this case, using (5.29), the test statistic (5.30) can be written as

$$F_{\boldsymbol{\beta}_2} = \frac{\boldsymbol{y}^\top \boldsymbol{M}_\iota \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_\iota \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_\iota \boldsymbol{y} / (k-1)}{\big(\boldsymbol{y}^\top \boldsymbol{M}_\iota \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{M}_\iota \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_\iota \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_\iota \boldsymbol{y}\big) / (n-k)}, \tag{5.37}$$

where $\boldsymbol{M}_\iota$ is the projection matrix that takes deviations from the mean, which was defined in (3.31). Thus the matrix expression in the numerator of (5.37) is just the explained sum of squares, or ESS, from the FWL regression

$$\boldsymbol{M}_\iota \boldsymbol{y} = \boldsymbol{M}_\iota \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \text{residuals}.$$

Similarly, the matrix expression in the denominator is the total sum of squares, or TSS, from this regression, minus the ESS. Since the centered R^2 from (5.36) is just the ratio of this ESS to this TSS, it requires only a little algebra to show that

$$F_{\boldsymbol{\beta}_2} = \frac{n-k}{k-1} \times \frac{R_c^2}{1 - R_c^2}.$$

Therefore, the F statistic (5.37) depends on the data only through the centered R^2, of which it is a monotonically increasing function.

Testing the Equality of Two Parameter Vectors

It is often natural to divide a sample into two, or possibly more than two, subsamples. These might correspond to periods of fixed exchange rates and floating exchange rates, large firms and small firms, rich countries and poor countries, or men and women, to name just a few examples. We may then ask whether a linear regression model has the same coefficients for both the subsamples. It is natural to use an F test for this purpose. Because the classic treatment of this problem is found in Chow (1960), the test is often called a **Chow test**; later treatments include Fisher (1970) and Dufour (1982).

Let us suppose, for simplicity, that there are only two subsamples, of lengths n_1 and n_2, with $n = n_1 + n_2$. We will assume that both n_1 and n_2 are greater than k, the number of regressors. If we separate the subsamples by partitioning the variables, we can write

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix},$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are, respectively, an n_1–vector and an n_2–vector, while $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $n_1 \times k$ and $n_2 \times k$ matrices. Even if we need different parameter vectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, for the two subsamples, we can nonetheless put the subsamples together in the following regression model:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{O} \\ \mathbf{O} & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{5.38}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The null hypothesis can now be written as $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Since it is preferable to express the null as a set of zero restrictions, a reformulation of (5.38) that achieves this is:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \mathbf{O} \\ \boldsymbol{X}_2 & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\gamma} \end{bmatrix} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{5.39}$$

It can readily be seen that, in the first subsample, the regression functions are the components of $\boldsymbol{X}_1 \boldsymbol{\beta}_1$, while, in the second, they are the components of $\boldsymbol{X}_2(\boldsymbol{\beta}_1 + \boldsymbol{\gamma})$. Thus $\boldsymbol{\gamma}$ is to be defined as $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$. If we define $\boldsymbol{Z}$ as an $n \times k$ matrix with $\mathbf{O}$ in its first n_1 rows and $\boldsymbol{X}_2$ in the remaining n_2 rows, then (5.39) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X} \boldsymbol{\beta}_1 + \boldsymbol{Z} \boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{5.40}$$

This is a regression model with n observations and $2k$ regressors. It has been constructed in such a way that $\boldsymbol{\beta}_1$ is estimated directly, while $\boldsymbol{\beta}_2$ is estimated using the relation $\boldsymbol{\beta}_2 = \boldsymbol{\gamma} + \boldsymbol{\beta}_1$. Since the restriction that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is equivalent to the restriction that $\boldsymbol{\gamma} = \mathbf{0}$ in (5.40), the null hypothesis has been expressed as a set of k zero restrictions. Since (5.40) is just a classical normal linear model with k linear restrictions to be tested, the F test provides the appropriate way to test those restrictions.

The F statistic can perfectly well be computed as usual, by running (5.40) to get the USSR and then running the restricted model, which is just the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, to get the RSSR. However, there is another way to compute the USSR. In Exercise 5.12, readers are invited to show that it is simply the sum of the two SSRs obtained by running two independent regressions on the two subsamples. If SSR_1 and SSR_2 denote the sums of squared residuals from these two regressions, and RSSR denotes the sum of squared residuals from regressing $\boldsymbol{y}$ on $\boldsymbol{X}$, the F statistic becomes

$$F_{\boldsymbol{\gamma}} = \frac{(\mathrm{RSSR} - \mathrm{SSR}_1 - \mathrm{SSR}_2)/k}{(\mathrm{SSR}_1 + \mathrm{SSR}_2)/(n-2k)}. \tag{5.41}$$

This **Chow statistic**, as it is often called, is distributed as $F(k, n-2k)$ under the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

5.5 Asymptotic Theory for Linear Regression Models

The t and F tests that we developed in the previous section are exact only under the strong assumptions of the classical normal linear model. If the disturbance vector were not normally distributed or not independent of the matrix of regressors, we could still compute t and F statistics, but they would not actually have their namesake distributions in finite samples. However,

Copyright © 2021, Russell Davidson and James G. MacKinnon

like a great many test statistics in econometrics that do not have any known distribution exactly, in many cases they would have known distributions approximately whenever the sample size was large enough. In such cases, we can perform what are called **large-sample tests** or **asymptotic tests**, using the approximate distributions to compute P values or critical values. In this section, we introduce several key results of asymptotic theory for linear regression models. These are then applied to large-sample tests in Section 5.6.

In general, asymptotic theory is concerned with the distributions of estimators and test statistics as the sample size n tends to infinity. Nevertheless, it often allows us to obtain simple results which provide useful approximations even when the sample size is far from infinite. Some of the basic ideas of asymptotic theory, in particular the concept of consistency, were introduced in Section 4.3. In this section, we investigate the asymptotic properties of the linear regression model. We show that, under much weaker assumptions than those of the classical normal linear model, the OLS estimator is asymptotically normally distributed with a familiar-looking covariance matrix.

Laws of Large Numbers

There are two types of fundamental results on which asymptotic theory is based. The first type, which we briefly discussed in Section 4.3, is called a **law of large numbers**, or **LLN**. A law of large numbers may apply to any quantity which can be written as an average of n random variables, that is, $1/n$ times their sum. Suppose, for example, that

$$\bar{x} \equiv \frac{1}{n}\sum_{t=1}^{n} x_t,$$

where the x_t are independent random variables, each with its own finite variance σ_t^2 and with a common expectation μ. We say that the variances σ_t^2 are **bounded** if there exists a finite real number $K > 0$ such that $\sigma_t^2 < K$ for all t. In that case, a fairly simple LLN assures us that, as $n \to \infty$, $\bar{x}$ tends to μ.

An example of how useful a law of large numbers can be is the **Fundamental Theorem of Statistics**, which concerns the **empirical distribution function**, or **EDF**, of a random sample. The EDF was introduced in Exercises 2.1 and 4.8. Let X be a random variable with CDF F, and suppose that we obtain a random sample of size n with typical element x_t, where each x_t is an independent realization of X. The **empirical distribution** defined by this sample is the discrete distribution that gives a weight of $1/n$ to each of the x_t for $t = 1, \ldots, n$. The EDF is the distribution function of the empirical distribution. It is defined algebraically as

$$\hat{F}(x) \equiv \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}(x_t \leq x), \tag{5.42}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

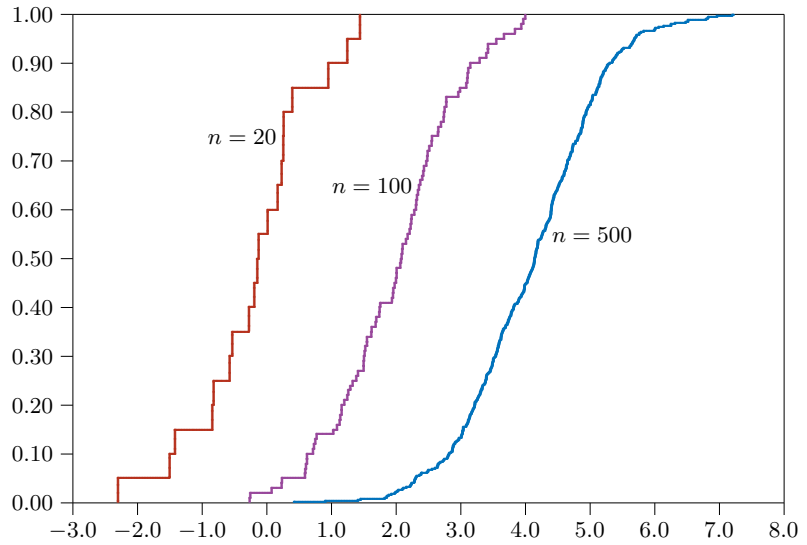


Figure 5.6 EDFs for several sample sizes

where $\mathbb{I}(\cdot)$ is the **indicator function**, which takes the value 1 when its argument is true and takes the value 0 otherwise. Thus, for a given argument x, the sum on the right-hand side of (5.42) counts the number of realizations x_t that are smaller than or equal to x.

The EDF has the form of a step function: The height of each step is $1/n$, and the width is equal to the difference between two successive values of x_t. As an illustration, Figure 5.6 shows the EDFs for three samples of sizes 20, 100, and 500 drawn from three normal distributions, each with variance 1 and with expectations 0, 2, and 4, respectively. These may be compared with the CDF of the standard normal distribution in the lower panel of Figure 5.2. There is not much resemblance between the EDF based on $n = 20$ and the normal CDF from which the sample was drawn, but the resemblance is somewhat stronger for $n = 100$ and very much stronger for $n = 500$. It is a simple matter to simulate data from an EDF, as we will see in Chapter 7, and this type of simulation can be very useful.

The Fundamental Theorem of Statistics tells us that the EDF consistently estimates the CDF of the random variable X. More formally, the theorem can be stated as

Theorem 5.2. (Fundamental Theorem of Statistics)

For the EDF $\hat{F}(x)$ defined in (5.42), for any x,

$$\operatorname*{plim}_{n\to\infty} \hat{F}(x) = F(x).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Proof:

For any real value of x, each term in the sum on the right-hand side of equation (5.42) depends only on x_t. The expectation of $\mathbb{I}(x_t \leq x)$ can be found by using the fact that it can take on only two values, 1 and 0. The expectation is

$$\begin{aligned}\mathrm{E}\big(\mathbb{I}(x_t \leq x)\big) &= 0 \cdot \Pr\big(\mathbb{I}(x_t \leq x) = 0\big) + 1 \cdot \Pr\big(\mathbb{I}(x_t \leq x) = 1\big) \\ &= \Pr\big(\mathbb{I}(x_t \leq x) = 1\big) = \Pr(x_t \leq x) = F(x).\end{aligned}$$

Since the x_t are mutually independent, so too are the terms $\mathbb{I}(x_t \leq x)$. Since the x_t all have the same distribution, so too must these terms. Thus $\hat{F}(x)$ is the mean of n IID random terms, each with finite expectation. The simplest of all LLNs (due to Khinchin) applies to such a mean. Thus we conclude that, for every x, $\hat{F}(x)$ is a consistent estimator of $F(x)$. ■

There are many different LLNs, some of which do not require that the individual random variables have a common expectation or be independent, although the amount of dependence must be limited. If we can apply a LLN to any random average, we can treat it as a nonrandom quantity for the purpose of asymptotic analysis. In many cases, as we saw in Section 4.3, this means that we must divide the quantity of interest by n. For example, the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ that appears in the OLS estimator generally does not converge to anything as $n \to \infty$. In contrast, the matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$, under many asymptotic constructions, tends to a nonstochastic limiting matrix $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ as $n \to \infty$.

Central Limit Theorems

The second type of fundamental result on which asymptotic theory is based is called a **central limit theorem**, or **CLT**. Central limit theorems are crucial in establishing the asymptotic distributions of estimators and test statistics. They tell us that, in many circumstances, $1/\sqrt{n}$ times the sum of n centered random variables has an approximately normal distribution when n is sufficiently large.

Suppose that the random variables x_t, $t = 1, \ldots, n$, are independently and identically distributed with expectation μ and variance σ^2. Then, according to the Lindeberg-Lévy central limit theorem, the quantity

$$z_n \equiv \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{x_t - \mu}{\sigma} \tag{5.43}$$

is **asymptotically distributed** as $\mathrm{N}(0, 1)$. This means that, as $n \to \infty$, the sequence of random variables z_n converges in distribution to the $\mathrm{N}(0, 1)$ distribution; recall the discussion of convergence in distribution in Section 4.3. We can write this result compactly as $z_n \xrightarrow{d} \mathrm{N}(0, 1)$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

It may seem curious that we divide by $\sqrt{n}$ instead of by n in (5.43), but this is an essential feature of every CLT. To see why, let us calculate the variance of z_n. Since the terms in the sum in (5.43) are independent, the variance of z_n is just the sum of the variances of the n terms:

$$\mathrm{Var}(z_n) = n\mathrm{Var}\Big(\frac{1}{\sqrt{n}}\frac{x_t - \mu}{\sigma}\Big) = \frac{n}{n} = 1.$$

If we had divided by n, we would, by a law of large numbers, have obtained a random variable with a plim of 0 instead of a random variable with a limiting standard normal distribution. Thus, whenever we want to use a CLT, we must ensure that a factor of $n^{-1/2} = 1/\sqrt{n}$ is present.

Just as there are many different LLNs, so too are there many different CLTs, almost all of which impose weaker conditions on the x_t than those imposed by the Lindeberg-Lévy CLT. The assumption that the x_t are identically distributed is easily relaxed, as is the assumption that they are independent. However, if there is either too much dependence or too much heterogeneity, a CLT may not apply. Several CLTs are discussed in Davidson and MacKinnon (1993, Section 4.7). Davidson (1994) provides a more advanced treatment.

In all cases of interest to us, the CLT says that, for a sequence of uncorrelated random variables x_t, $t = 1, \ldots, \infty$, with $\mathrm{E}(x_t) = 0$,

$$n^{-1/2}\sum_{t=1}^{n} x_t = x_n^0 \xrightarrow{d} \mathrm{N}\Big(0, \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{Var}(x_t)\Big).$$

We sometimes need vector, or **multivariate**, versions of CLTs. Suppose that we have a sequence of uncorrelated random m–vectors $\boldsymbol{x}_t$, for some fixed m, with $\mathrm{E}(\boldsymbol{x}_t) = \boldsymbol{0}$. Then the appropriate multivariate CLT tells us that

$$n^{-1/2}\sum_{t=1}^{n} \boldsymbol{x}_t = \boldsymbol{x}_n^0 \xrightarrow{d} \mathrm{N}\Big(\boldsymbol{0}, \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{Var}(\boldsymbol{x}_t)\Big), \tag{5.44}$$

where $\boldsymbol{x}_n^0$ is multivariate normal, and each $\mathrm{Var}(\boldsymbol{x}_t)$ is an $m \times m$ matrix.

Figure 5.7 illustrates the fact that CLTs often provide good approximations even when n is not very large. Both panels of the figure show the densities of various random variables z_n defined as in (5.43). In the top panel, the x_t are uniformly distributed, and we see that z_n is remarkably close to being distributed as standard normal even when n is as small as 8. This panel does not show results for larger values of n because they would have made it too hard to read. In the bottom panel, the x_t have the $\chi^2(1)$ distribution, which exhibits extreme right skewness. The mode[9] of the distribution is 0, there

[9] A **mode** of a distribution is a point at which the density achieves a local maximum. If there is just one such point, a density is said to be **unimodal**.

Copyright © 2021, Russell Davidson and James G. MacKinnon

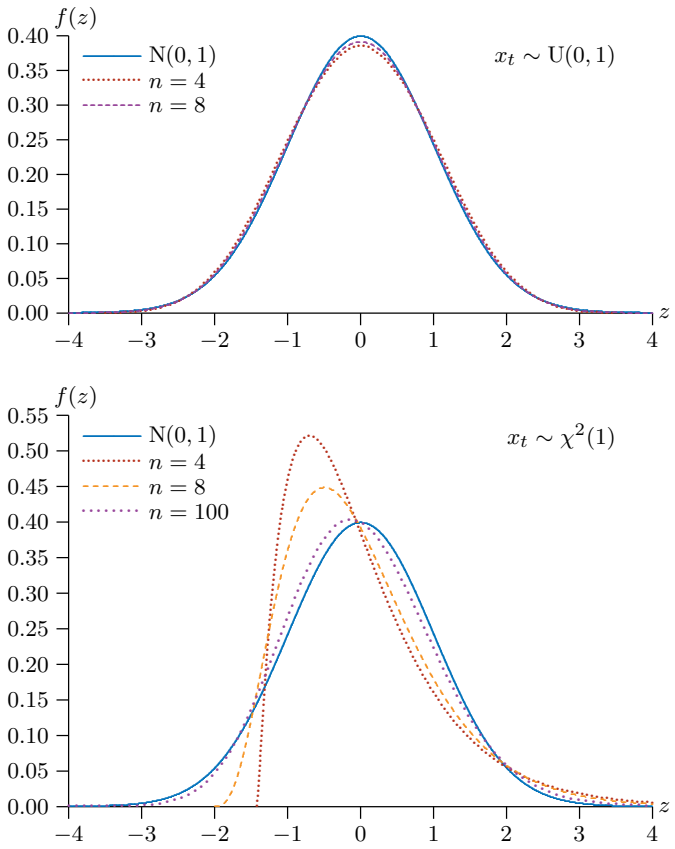


Figure 5.7 The normal approximation for different values of n

are no values less than 0, and there is a very long right-hand tail. For $n = 4$ and $n = 8$, the standard normal provides a poor approximation to the actual distribution of z_n. For $n = 100$, on the other hand, the approximation is not bad at all, although it is still noticeably skewed to the right.

Asymptotic Normality and Root-n Consistency

Although the notion of **asymptotic normality** is very general, for now we will introduce it for linear regression models only. Suppose that the data are generated by the DGP

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{\mathrm{I}}), \tag{5.45}$$

instead of the classical normal linear model (5.17). The disturbances here are drawn from some specific but unknown distribution with expectation 0

Copyright © 2021, Russell Davidson and James G. MacKinnon

and variance σ_0^2. Although some or all of the regressors may be exogenous, that assumption is stronger than we need. Instead, we allow $\boldsymbol{X}_t$ to contain lagged dependent variables, replacing the exogeneity assumption with assumption (4.13) from Section 4.2, plus an analogous assumption about the variance. These two assumptions can be written as

$$\mathrm{E}(u_t \,|\, \boldsymbol{X}_t) = 0 \quad \text{and} \quad \mathrm{E}(u_t^2 \,|\, \boldsymbol{X}_t) = \sigma_0^2. \tag{5.46}$$

The first equation here, which is assumption (4.13), can be referred to in two ways. From the point of view of the explanatory variables $\boldsymbol{X}_t$, it says that they are **predetermined** with respect to the disturbances, a terminology that was introduced in Section 4.2. From the point of view of the disturbances, however, it says that they are **innovations**. An innovation is a random variable of which the expectation is 0 conditional on the information in the explanatory variables, and so knowledge of the values taken by the latter is of no use in predicting the expectation of the innovation. We thus have two different ways of saying the same thing. Both can be useful, depending on the circumstances.

Although we have greatly weakened the assumptions of the classical normal linear model in equations (5.45) and (5.46), we now need to make an additional assumption in order to be able to use asymptotic results. We assume that the data-generating process for the explanatory variables is such that, under the asymptotic construction used in order to obtain asymptotic results,

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}, \tag{5.47}$$

where $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ is a finite, deterministic, positive definite matrix. We made this assumption previously, in Section 4.3, when we proved that the OLS estimator is consistent. Although it is often reasonable, condition (5.47) is violated by some asymptotic constructions, as we saw in Section 4.3. For example, it cannot hold if one of the columns of the $\boldsymbol{X}$ matrix is a linear time trend, because $\sum_{t=1}^n t^2$ grows at a rate faster than n.

Now consider the k–vector

$$\boldsymbol{v} \equiv n^{-1/2} \boldsymbol{X}^\top \boldsymbol{u} = n^{-1/2} \sum_{t=1}^n u_t \boldsymbol{X}_t^\top. \tag{5.48}$$

We wish to apply a multivariate CLT to this vector. By the first assumption in (5.46), $\mathrm{E}(u_t \,|\, \boldsymbol{X}_t) = 0$. This implies that $\mathrm{E}(u_t \boldsymbol{X}_t^\top) = \boldsymbol{0}$, as required for the CLT. Thus, assuming that the vectors $u_t \boldsymbol{X}_t^\top$ satisfy the technical assumptions for an appropriate multivariate CLT to apply, we have from (5.44) that

$$\boldsymbol{v} \xrightarrow{d} \mathrm{N}\Big(\boldsymbol{0}, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \mathrm{Var}(u_t \boldsymbol{X}_t^\top)\Big) = \mathrm{N}\Big(\boldsymbol{0}, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \mathrm{E}(u_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t)\Big).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Notice that, because $\boldsymbol{X}_t$ is a $1 \times k$ row vector, the covariance matrix here is $k \times k$, as it must be.

The second assumption in equations (5.46) says that the disturbances are conditionally homoskedastic. It allows us to simplify the limiting covariance matrix:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(u_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t) &= \lim_{n\to\infty} \sigma_0^2 \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(\boldsymbol{X}_t^\top \boldsymbol{X}_t) \\ &= \sigma_0^2 \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}_t^\top \boldsymbol{X}_t \\ &= \sigma_0^2 \operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}.\end{aligned} \tag{5.49}$$

We applied a LLN in reverse to go from the first line to the second, and the last equality follows from assumption (5.47). Thus we conclude that

$$\boldsymbol{v} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}). \tag{5.50}$$

Consider now the estimation error of the vector of OLS estimates. For the DGP (5.45), this is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}. \tag{5.51}$$

As we saw in Section 4.3, $\hat{\boldsymbol{\beta}}$ is consistent under any sensible asymptotic construction. If it is, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ must tend to a limit of $\boldsymbol{0}$ as the sample size $n \to \infty$. Therefore, its limiting covariance matrix is a zero matrix. Thus it would appear that asymptotic theory has nothing to say about limiting variances for consistent estimators. However, this is easily corrected by the usual device of introducing a few well-chosen powers of n.[10] If we rewrite equation (5.51) as

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (n^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1} n^{-1/2} \boldsymbol{X}^\top \boldsymbol{u},$$

For the first factor, we have $\operatorname{plim}_{n\to\infty}\big[(n^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1} - \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}\big] = \boldsymbol{O}$, and so

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\big[(n^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1} - \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}\big] n^{-1/2} \boldsymbol{X}^\top \boldsymbol{u} = \boldsymbol{0}, \quad \text{or} \\ \operatorname*{plim}_{n\to\infty}\big[n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{v}\big] = \boldsymbol{0}.\end{aligned} \tag{5.52}$$

[10] Under standard assumptions, sums of random variables that have nonzero expectations, like the elements of the matrix $\boldsymbol{X}^\top \boldsymbol{X}$, are $O_p(n)$, and weighted sums of random variables that have zero expectations, like the elements of the vector $\boldsymbol{X}^\top \boldsymbol{u}$, are $O_p(n^{1/2})$. We need to multiply the former by n^{-1} and the latter by $n^{-1/2}$ in order to obtain quantities that are $O_p(1)$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The relation (5.52) can be written as

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{v}, \tag{5.53}$$

where the symbol $\stackrel{a}{=}$ is used for **asymptotic equality**. By definition, it means that the plim of the difference between two things that are asymptotically equal is zero. Because $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ is deterministic, we find, using (5.50), that the variance of the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is

$$\sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}.$$

Moreover, since $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{v}$ is just a deterministic linear combination of the components of the multivariate normal random vector $\boldsymbol{v}$, we conclude from (5.53) that

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}).$$

Informally, we may say that the vector $\hat{\boldsymbol{\beta}}$ is **asymptotically normal**, or exhibits **asymptotic normality**.

It is convenient to collect the key results above into a theorem.

Theorem 5.3.

For the correctly specified linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2 \boldsymbol{\mathrm{I}}), \tag{5.54}$$

where the data are generated by the DGP (5.45), the regressors and disturbances satisfy assumptions (5.46), and the regressors satisfy assumption (5.47) for the chosen asymptotic construction, we have

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}), \tag{5.55}$$

and

$$\operatorname*{plim}_{n\to\infty} s^2 (n^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}. \tag{5.56}$$

Remark:

The first part of the theorem allows us to pretend that $\hat{\boldsymbol{\beta}}$ is normally distributed with expectation $\boldsymbol{0}$, and the second part, which follows on account of the consistency of s^2 for $\hat{\sigma}^2$ proved in Section 4.7, allows us to use $s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ to estimate $\mathrm{Var}(\hat{\boldsymbol{\beta}})$. The result (5.56) tells us that the **asymptotic covariance matrix** of the vector $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is the limit of the matrix $\sigma_0^2(n^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ as $n \to \infty$. Of course, both (5.55) and (5.56) just approximations. The theorem does not tell us that **asymptotic inference** based on these approximations will necessarily be reliable.

Copyright © 2021, Russell Davidson and James G. MacKinnon

It is important to remember that, whenever the matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ tends to $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ as $n \to \infty$, the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, without the factor of n, simply tends to a zero matrix. As we saw just below equation (5.51), this is simply a consequence of the fact that $\hat{\boldsymbol{\beta}}$ is consistent. Thus, although it would be convenient if we could dispense with powers of n when working out asymptotic approximations to covariance matrices, it would be mathematically incorrect and very risky to do so.

The result (5.55) gives us the **rate of convergence** of $\hat{\boldsymbol{\beta}}$ to its probability limit of $\boldsymbol{\beta}_0$. Since multiplying the estimation error by $n^{1/2}$ gives rise to an expression of zero expectation and finite covariance matrix, it follows that the estimation error itself tends to zero at the same rate as $n^{-1/2}$, that is, it is $O_p(n^{-1/2})$. This property is expressed by saying that the estimator $\hat{\boldsymbol{\beta}}$ is **root-n consistent**.

Quite generally, suppose that $\hat{\boldsymbol{\theta}}$ is a root-n consistent, asymptotically normal, estimator of a parameter vector $\boldsymbol{\theta}$. Any estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ must tend to zero as $n \to \infty$. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, and let $\boldsymbol{V}(\boldsymbol{\theta})$ denote the limiting covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Then an estimator $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ is said to be consistent for the covariance matrix of $\hat{\boldsymbol{\theta}}$ if

$$\operatorname*{plim}_{n\to\infty}\big(n\,\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\big) = \boldsymbol{V}(\boldsymbol{\theta}). \tag{5.57}$$

For every root-n consistent estimator, there is generally at least one such covariance matrix estimator.

5.6 Large-Sample Tests

Theorem 5.3 implies that the t test discussed in Section 5.4 is asymptotically valid under weaker conditions than those needed to prove that the t statistic actually has Student's t distribution in finite samples. Consider the linear regression model (5.18), but with IID disturbances and regressors that may be predetermined rather than exogenous. As before, we wish to test the hypothesis that $\beta_2 = \beta_2^0$. The t statistic for this hypothesis is simply

$$t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2^0}{\sqrt{s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}}} = \frac{n^{1/2}(\hat{\beta}_2 - \beta_2^0)}{\sqrt{s^2(n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}}}, \tag{5.58}$$

where $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}$ denotes the last element on the main diagonal of the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. The result (5.55) tells us that $n^{1/2}(\hat{\beta}_2 - \beta_2^0)$ is asymptotically normally distributed with variance the last element on the main diagonal of $\sigma_0^2\boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}}$. Equation (5.56) tells us that s^2 times $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}$ consistently estimates this variance. Therefore, t_{β_2} has the standard normal distribution asymptotically under the null hypothesis. We may write

$$t_{\beta_2} \overset{a}{\sim} \mathrm{N}(0,1). \tag{5.59}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The notation "$\overset{a}{\sim}$" means that t_{β_2} is **asymptotically distributed** as $\mathrm{N}(0,1)$. This is just a different way of saying that t_{β_2} converges in distribution to $\mathrm{N}(0,1)$, and this implies that $t_{\beta_2} = O_p(1)$.

The result (5.59) justifies the use of t tests outside the confines of the classical normal linear model. We can compute asymptotic P values or critical values using either the standard normal or t distributions. Of course, these **asymptotic *t* tests** are not exact in finite samples, and they may or may not be reliable. It is often possible to perform more reliable tests by using bootstrap methods, which will be introduced in Chapter 7.

Asymptotic *F* Tests

In view of the result (5.59) for the asymptotic t statistic, it should not be surprising that the F statistic (5.30) for the null hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ in the model (5.25) is also valid asymptotically when the DGP is (5.45) and the disturbances satisfy assumptions (5.46). Under the null, $F_{\boldsymbol{\beta}_2}$ is equal to expression (5.31). Rewriting this expression in terms of quantities that are $O_p(1)$, we obtain

$$F_{\boldsymbol{\beta}_2} = \frac{n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{M}_1\boldsymbol{X}_2(n^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon}/r}{\boldsymbol{\varepsilon}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}/(n-k)}, \tag{5.60}$$

where $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma_0$ and $r = k_2$, the dimension of $\boldsymbol{\beta}_2$. We now show that $rF_{\boldsymbol{\beta}_2}$ is asymptotically distributed as $\chi^2(r)$. This result follows from Theorem 5.3, but it is not entirely obvious.

The denominator of the F statistic (5.60) is $\boldsymbol{\varepsilon}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}/(n-k)$, which is just s^2 times $1/\sigma_0^2$. Since s^2 is consistent for σ_0^2, it is evident that the denominator of expression (5.60) tends to 1 asymptotically.

The numerator of the F statistic, multiplied by r, is

$$n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{M}_1\boldsymbol{X}_2(n^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\,n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon}. \tag{5.61}$$

Let $\boldsymbol{v} = n^{-1/2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$. Then a central limit theorem shows that $\boldsymbol{v} \overset{a}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})$, as in the previous section. Now

$$n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon} = n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon} - n^{-1}\boldsymbol{X}_2^\top\boldsymbol{X}_1(n^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}n^{-1/2}\boldsymbol{X}_1^\top\boldsymbol{\varepsilon}. \tag{5.62}$$

If we partition $\boldsymbol{v}$, conformably with the partition of $\boldsymbol{X}$, into two subvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, it is clear that the right-hand side of (5.62) tends to the vector $\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1$ as $n \to \infty$. Here $\boldsymbol{S}_{11}$ and $\boldsymbol{S}_{21}$ are submatrices of $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$, so that $n^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_1$ tends to $\boldsymbol{S}_{11}$ and $n^{-1}\boldsymbol{X}_2^\top\boldsymbol{X}_1$ tends to $\boldsymbol{S}_{21}$. Since $\boldsymbol{v}$ is asymptotically multivariate normal, and $\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1$ is just a linear combination of the elements of $\boldsymbol{v}$, this vector must itself be asymptotically multivariate normal.

The vector (5.62) evidently has expectation $\boldsymbol{0}$. Thus its covariance matrix is the expectation of

$$n^{-1}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}_2 + \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}n^{-1}\boldsymbol{X}_1^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}_1\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} - 2n^{-1}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}_1\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

We can replace $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top$ by its expectation, which is $\mathbf{I}$. Then, by the second part of Theorem 5.3, we can replace $n^{-1}\boldsymbol{X}_i^\top\boldsymbol{X}_j$ by $\boldsymbol{S}_{ij}$, for $i, j = 1, 2$, which is what each of those submatrices tends to asymptotically. This yields an expression that can be simplified, allowing us to conclude that

$$\operatorname{Var}(\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1) = \boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}.$$

Thus the numerator of the F statistic, expression (5.61), is asymptotically equal to

$$(\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1)^\top(\boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12})^{-1}(\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1). \tag{5.63}$$

This is simply a quadratic form in the r–vector $\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1$, which is asymptotically multivariate normal, and the inverse of its covariance matrix. By Theorem 5.1, it follows that expression (5.63) is asymptotically distributed as $\chi^2(r)$. Because the denominator of the F statistic tends to 1 asymptotically, we conclude that

$$rF_{\boldsymbol{\beta}_2} \overset{a}{\sim} \chi^2(r) \tag{5.64}$$

under the null hypothesis with predetermined regressors. Since $1/r$ times a random variable that has the $\chi^2(r)$ distribution is distributed as $F(r, \infty)$, we may also conclude that $F_{\boldsymbol{\beta}_2} \overset{a}{\sim} F(r, n-k)$.

The result (5.64) justifies the use of **asymptotic F tests** when the disturbances are not normally distributed and some of the regressors are predetermined rather than exogenous. We can compute the P value associated with an F statistic using either the χ^2 or F distributions. Of course, if we use the χ^2 distribution, we have to multiply the F statistic by r.

Wald Tests

A vector of r linear restrictions on a parameter vector $\boldsymbol{\beta}$ can always be written in the form

$$\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}, \tag{5.65}$$

where $\boldsymbol{R}$ is an $r \times k$ matrix and $\boldsymbol{r}$ is an r–vector. For example, if $k = 3$ and the restrictions were that $\beta_1 = 0$ and $\beta_2 = -1$, equations (5.65) would be

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\boldsymbol{\beta} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

The elements of the matrix $\boldsymbol{R}$ and the vector $\boldsymbol{r}$ must be known. They are not functions of the data, and, as in this example, they are very often integers.

Now suppose that we have obtained a k–vector of unrestricted parameter estimates $\hat{\boldsymbol{\beta}}$, of which the covariance matrix is $\operatorname{Var}(\hat{\boldsymbol{\beta}})$. By a slight generalization of the result (4.44), the covariance matrix of the vector $\boldsymbol{R}\boldsymbol{\beta}$ is $\boldsymbol{R}\operatorname{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Then the simplest way to test the restrictions (5.65) is to calculate the **Wald statistic**

$$W(\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top\big(\boldsymbol{R}\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \tag{5.66}$$

where $\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}})$ estimates $\operatorname{Var}(\hat{\boldsymbol{\beta}})$ consistently. Inserting appropriate powers of n so that each factor is $O_p(1)$, equation (5.66) can be rewritten as

$$W(\hat{\boldsymbol{\beta}}) = \big(n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\big)^\top\big(\boldsymbol{R}n\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top\big)^{-1}\big(n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\big). \tag{5.67}$$

Theorem 5.3 implies that the vector $n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ is asymptotically multivariate normal. Therefore, the right-hand side of equation (5.67) is asymptotically a quadratic form in an r–vector that is multivariate normal and the inverse of its covariance matrix. It follows that, by Theorem 5.1, the Wald statistic is asymptotically distributed as $\chi^2(r)$ under the null hypothesis.

These results are much more general than the ones for asymptotic t tests and F tests. Equation (5.66) would still define a Wald statistic for the hypothesis (5.65) if $\hat{\boldsymbol{\beta}}$ were any root-n consistent estimator and $\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}})$ were any consistent estimator of its covariance matrix. Thus we will encounter Wald tests several times throughout this book. For the specific case of a linear regression model with zero restrictions on some of the parameters, Wald tests turn out to be very closely related to t tests and F tests. In fact, the square of the t statistic (5.22) is a Wald statistic, and r times the F statistic (5.30) is a Wald statistic. Readers are asked to demonstrate these results in Exercise 5.15 and Exercise 5.16, respectively.

Asymptotic tests cannot be exact in finite samples, because they are necessarily based on P values, or critical values, that are approximate. By itself, asymptotic theory cannot tell us just how accurate such tests are. If we decide to use a nominal level of α for a test, we reject if the asymptotic P value is less than α. In many cases, but certainly not all, asymptotic tests are probably quite accurate, committing Type I errors with probability reasonably close to α. They may either **overreject**, that is, reject the null hypothesis more than $100\alpha\%$ of the time when it is true, or **underreject**, that is, reject the null hypothesis less than $100\alpha\%$ of the time. Whether they overreject or underreject, and how severely, depends on many things, including the sample size, the distribution of the disturbances, the number of regressors and their properties, the number of restrictions, and the relationship between the disturbances and the regressors.

In the next section, we will see a test set up deliberately so as to underreject for any given nominal level. The motivation for this is to be sure of not committing a Type I error more frequently than at the rate chosen as the level. Such tests are called **conservative**.

Copyright © 2021, Russell Davidson and James G. MacKinnon

5.7 Performing Multiple Hypothesis Tests

Up to this point, we have implicitly assumed that just one hypothesis test is performed at a time. This allows test statistics and P values to be interpreted in the usual way. In practice, however, investigators almost always perform several tests simultaneously. For example, whenever an econometrics package is used to run an OLS regression, the package will normally report a t statistic for every coefficient. Unless the investigator consciously chooses to ignore all but one of these t statistics, which most people would find it almost impossible to do, he or she is implicitly (and often explicitly) engaged in **multiple testing**. This simply means performing two or more hypothesis tests of different null hypotheses as part of the same investigation. It is not to be confused with testing two or more restrictions via a single test, such as an F test or a Wald test.

The problem with multiple testing is that an unusually large test statistic is much more likely to be obtained by chance when several tests are performed rather than just one, even when all of the null hypotheses being tested severally are true. This is easiest to see if the test statistics are independent. Suppose that we perform m independent exact tests at level α. Let α_m denote the **familywise error rate**, which is the probability that *at least* one of the tests rejects. Because the tests are independent, the familywise error rate is simply one minus the probability that none of the tests rejects:

$$\alpha_m = 1 - (1 - \alpha)^m. \tag{5.68}$$

When m is large, α_m can be much larger than α. For example, if $\alpha = 0.05$, then $\alpha_2 = 0.0975$, $\alpha_4 = 0.18549$, $\alpha_8 = 0.33658$, and $\alpha_{16} = 0.55987$. It is evident from (5.68) that the familywise error rate can be very much larger than the level of each individual test when the number of tests is large.

The simplest method for controlling the familywise error rate is known as the **Bonferroni procedure**. Instead of rejecting any of the hypotheses when the corresponding P value is less than α, rejection occurs only if the P value is less than α/m. In this way, as we will see, the familywise error rate is bounded above by α. This procedure is based on the Bonferroni inequality

$$\Pr\Big(\bigcup_{i=1}^{m}(P_i \le \alpha/m)\Big) \le \alpha, \tag{5.69}$$

where P_i is the P value for the i^{th} test. The event on the left-hand side of (5.69) is just the probability that at least one of the hypotheses is rejected at level α/m, that is, the familywise error rate. The inequality ensures that this rate is less than α. The Bonferroni inequality is easy to prove. Consider the union of two events A and B, as illustrated in Figure 2.3. It is clear from the figure that

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B), \text{ so that } \Pr(A \cup B) \le \Pr(A) + \Pr(B).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

More generally, if we have a set of events A_i, $i = 1, \dots, m$, it is easy to see that

$$\Pr\Big(\bigcup_{i=1}^{m} A_i\Big) \le \sum_{i=1}^{m} \Pr(A_i). \tag{5.70}$$

If the tests are all exact, then we have $\Pr(P_i \le \alpha/m) = \alpha/m$, and so (5.69) follows from (5.70).

The Bonferroni procedure is very easy to implement, but it can be extremely conservative. For large m, α/m is very much smaller than α. It is even smaller than the value α that solves equation (5.68) for a given α_m, which would be appropriate if all the tests were independent. When the P values are positively correlated, as is often the case in practice, α/m can be much too small. Consider the extreme case in which there is perfect dependence and all the tests yield identical P values. In that case, the familywise error rate for individual tests at level α is just α, and no correction is needed.

There is a large literature on multiple testing in statistics. For example, Simes (1986) and Hochberg (1988) proposed improved Bonferroni procedures that are less conservative. They both use all the P values, not just the smallest one. The Simes procedure is quite simple. We first order the P values from the smallest, $P_{(1)}$, to the largest, $P_{(m)}$. Then the rejection rule for the individual tests becomes

$$P_{(j)} \le j\alpha/m \text{ for any } j = 1, \dots, m, \tag{5.71}$$

where α is the desired familywise error rate. If the smallest P value is less than α/m, both this procedure and the Bonferroni procedure reject the corresponding hypothesis. But the Simes procedure can also reject when the second-smallest P value is less than $2\alpha/m$, the third-smallest is less than $3\alpha/m$, and so on. Thus it is always less conservative than the Bonferroni procedure. Because it is based on an inequality that may not always hold, the Simes procedure can conceivably yield misleading results, but it seems to work well in practice.

A more recent approach is to control the **false discovery rate** instead of the familywise error rate; see Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). The false discovery rate is the expected proportion of erroneous rejections among all rejections. The idea is that some of the tested hypotheses may be true and others may be false, and so we want to reject the false nulls but not the true ones. As in (5.71), we order the P values and compare them to $j\alpha/m$. Let J denote the largest j for which the inequality holds. Then we reject the first J hypotheses and do not reject the remaining ones. If the inequality is never satisfied, then the Benjamini-Hochberg and Simes procedures yield the same results.

In this section, we have given a very brief introduction to testing multiple hypotheses. We have assumed that little or nothing is known about the joint distribution of the test statistics. If we knew that distribution, then we could

Copyright © 2021, Russell Davidson and James G. MacKinnon

in principle do better than any of the procedures discussed above. This suggests that bootstrap-based procedures may be attractive, and these will be discussed in Chapter 7.

5.8 The Power of Hypothesis Tests

To be useful, hypothesis tests must be able to discriminate between the null hypothesis and the alternative. Thus, as we saw in Section 5.2, the distribution of a useful test statistic under the null is different from its distribution when the DGP does not belong to the null. Whenever a DGP places most of the probability mass of the test statistic in the rejection region of a test, the test has high power, that is, a high probability of rejecting the null.

For a variety of reasons, it is important to know something about the power of the tests we employ. For example, if it expensive to obtain the data, as it might well be in an experimental setting, it will often make sense to verify in advance that the sample will be large enough for the tests that we are proposing to perform to have enough power to discriminate between the null and alternative hypotheses. If a test fails to reject the null, this tells us more if the test had high power against plausible alternatives than if the test had low power. In practice, more than one test of a given null hypothesis is usually available. Of two equally reliable tests, if one has more power than the other against the alternatives in which we are interested, then we would surely prefer to employ the more powerful one.

In Section 5.4, we saw that an F statistic is a ratio of the squared norms of two vectors, each divided by its appropriate number of degrees of freedom. In the notation of that section, these vectors are $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{y}$ for the numerator and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ for the denominator. If the null and alternative hypotheses are classical normal linear models, as we assume throughout this subsection, then, under the null, both the numerator and the denominator of this ratio are independent χ^2 variables, divided by their respective degrees of freedom; see expression (5.31). Under the alternative hypothesis, the distribution of the denominator is unchanged, because, under either hypothesis, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$. Consequently, the difference in distribution under the null and the alternative that gives the test its power must come from the numerator alone.

From equation (5.30), r/σ^2 times the numerator of the F statistic $F_{\boldsymbol{\beta}_2}$ is

$$\frac{1}{\sigma^2}\boldsymbol{y}^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y}. \tag{5.72}$$

The vector $\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y}$ is normal under both the null and the alternative. Its expectation is $\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2$, which vanishes under the null when $\boldsymbol{\beta}_2 = \boldsymbol{0}$, and its covariance matrix is $\sigma^2\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2$. We can use these facts to determine the distribution of the quadratic form (5.72). To do so, we must introduce the

Copyright © 2021, Russell Davidson and James G. MacKinnon

noncentral chi-squared distribution, which is a generalization of the ordinary, or **central**, chi-squared distribution.

We saw in Section 5.3 that, if the m–vector $\boldsymbol{z}$ is distributed as $\mathrm{N}(\boldsymbol{0}, \boldsymbol{\mathrm{I}})$, then $\|\boldsymbol{z}\|^2 = \boldsymbol{z}^\top\boldsymbol{z}$ is distributed as (central) chi-squared with m degrees of freedom. Similarly, if $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, then $\boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x} \sim \chi^2(m)$. If instead $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\mathrm{I}})$, then $\boldsymbol{z}^\top\boldsymbol{z}$ has the noncentral chi-squared distribution with m degrees of freedom and **noncentrality parameter**, or **NCP**, $\Lambda \equiv \boldsymbol{\mu}^\top\boldsymbol{\mu}$. This distribution is written as $\chi^2(m, \Lambda)$. It is easy to see that its expectation is $m + \Lambda$; see Exercise 5.22. Likewise, if $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then $\boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x} \sim \chi^2(m, \boldsymbol{\mu}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\mu})$. Although we will not prove it, the distribution depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ only through the quadratic form $\boldsymbol{\mu}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}$. If we set $\boldsymbol{\mu} = \boldsymbol{0}$, we see that the $\chi^2(m, 0)$ distribution is just the central $\chi^2(m)$ distribution.

Under either the null or the alternative hypothesis, therefore, the distribution of expression (5.72) is noncentral chi-squared, with r degrees of freedom, and with noncentrality parameter given by

$$\begin{aligned}\Lambda &\equiv \frac{1}{\sigma^2}\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 \\ &= \frac{1}{\sigma^2}\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2. \end{aligned}\tag{5.73}$$

Under the null, $\Lambda = 0$. Under either hypothesis, the distribution of the denominator of the F statistic, divided by σ^2, is central chi-squared with $n-k$ degrees of freedom, and it is independent of the numerator. The F statistic therefore has a distribution that we can write as

$$\frac{\chi^2(r, \Lambda)/r}{\chi^2(n-k)/(n-k)},$$

with numerator and denominator mutually independent. This distribution is called the **noncentral F distribution**, with r and $n-k$ degrees of freedom and noncentrality parameter Λ. In any given testing situation, r and $n-k$ are given, and so the difference between the distributions of the F statistic under the null and under the alternative depends only on the NCP Λ.

To illustrate this fact, we limit our attention to expression (5.72), that is, r/σ^2 times the numerator of the F statistic, which is distributed as $\chi^2(r, \Lambda)$. As Λ increases, the distribution moves to the right and becomes more spread out. This happens because, under the alternative, expression (5.72) is equal to

$$\begin{aligned}&\frac{1}{\sigma^2}\boldsymbol{u}^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{u} \\ &\qquad + \frac{2}{\sigma^2}\boldsymbol{u}^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \frac{1}{\sigma^2}\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2.\end{aligned}$$

The first term here has the central $\chi^2(r)$ distribution. The third term is the noncentrality parameter Λ. The second term is a random scalar which

Copyright © 2021, Russell Davidson and James G. MacKinnon

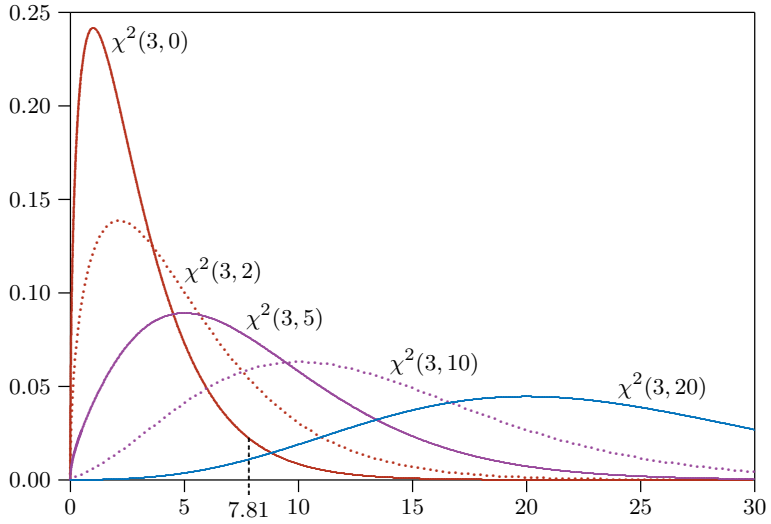


Figure 5.8 Densities of noncentral χ^2 distributions

is normally distributed with expectation zero and has variance equal to four times the NCP. The third term is what causes the $\chi^2(r, \Lambda)$ distribution to move to the right as Λ increases, and the second term is what causes it to become more spread out.

The way in which the noncentral χ^2 distribution depends on Λ is illustrated in Figure 5.8, which shows the density of the $\chi^2(3, \Lambda)$ distribution for noncentrality parameters of 0, 2, 5, 10, and 20. The .05 critical value for the central $\chi^2(3)$ distribution, which is 7.81, is also shown. If a test statistic has the noncentral $\chi^2(3, \Lambda)$ distribution, the probability that the null hypothesis is rejected at the .05 level is the probability mass to the right of 7.81. It is evident from the figure that this probability is small for small values of the NCP and large for large ones.

In Figure 5.8, the number of degrees of freedom r is held constant as Λ is increased. If, instead, we held Λ constant, the density functions would move to the right as r was increased, as they do in Figure 5.4 for the special case with $\Lambda = 0$. Thus, at any given level, the critical value of a χ^2 or F test increases as r increases. It has been shown by Das Gupta and Perlman (1974) that this rightward shift of the critical value has a greater effect than the rightward shift of the density for any positive Λ. Specifically, Das Gupta and Perlman show that, for a given NCP, the power of a χ^2 or F test at any given level is strictly decreasing in r, as well as being strictly increasing in Λ, as we indicated in the previous paragraph.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The square of a t statistic for a single restriction is just the F test for that restriction, and so the above analysis applies equally well to t tests. Things can be made a little simpler, however. From equation (5.22), the t statistic t_{β_2} is $1/s$ times

$$\frac{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}}. \tag{5.74}$$

The numerator of this expression, $\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}$, is normally distributed under both the null and the alternative, with variance $\sigma^2 \boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2$ and expectation $\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2 \beta_2$. Thus $1/\sigma$ times expression (5.74) is normal with variance 1 and expectation

$$\lambda \equiv \frac{1}{\sigma} (\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2} \beta_2. \tag{5.75}$$

It follows that t_{β_2} has a distribution which can be written as

$$\frac{\mathrm{N}(\lambda, 1)}{\left(\chi^2(n-k)/(n-k)\right)^{1/2}},$$

with independent numerator and denominator. This distribution is known as the **noncentral t distribution**, with $n-k$ degrees of freedom and noncentrality parameter λ; it is written as $t(n-k, \lambda)$. Note that $\lambda^2 = \Lambda$, where Λ is the NCP of the corresponding F statistic. Except for very small sample sizes, the $t(n-k, \lambda)$ distribution is quite similar to the $\mathrm{N}(\lambda, 1)$ distribution. It is also very much like an ordinary, or **central**, t distribution with its expectation shifted from the origin to (5.75), but it has a bit more variance, because of the stochastic denominator.

When we know the distribution of a test statistic under the alternative hypothesis, we can determine the power of a test at any given level as a function of the parameters of that hypothesis. This function is called the **power function** of the test. The distribution of t_{β_2} under the alternative depends only on the NCP λ. For a given regressor matrix $\boldsymbol{X}$ and sample size n, λ in turn depends on the parameters only through the ratio β_2/σ; see (5.75). Therefore, the power of the t test depends only on this ratio. According to assumption (5.47), as $n \to \infty$, $n^{-1}\boldsymbol{X}^\top \boldsymbol{X}$ tends to a nonstochastic limiting matrix $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$. Thus both the factor $(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}$ and λ itself are evidently $O(n^{1/2})$.

Figure 5.9 shows power functions for tests at the .05 level for a very simple model, in which $\boldsymbol{x}_2$, the only regressor, is a constant. Power is plotted as a function of β_2/σ for three sample sizes: $n = 25$, $n = 100$, and $n = 400$. Since the test is exact, all the power functions are equal to .05 when $\beta_2 = 0$. Power then increases as β_2 moves away from 0. As we would expect, the power when $n = 400$ exceeds the power when $n = 100$, which in turn exceeds the power when $n = 25$, for every value of $\beta_2 \neq 0$. It is clear that, as $n \to \infty$, the power function converges to the shape of a T, with the foot of the vertical segment at .05 and the horizontal segment at 1.0. Thus, asymptotically, the test rejects the null with probability 1 whenever it is false. In finite samples, however, we

Copyright © 2021, Russell Davidson and James G. MacKinnon

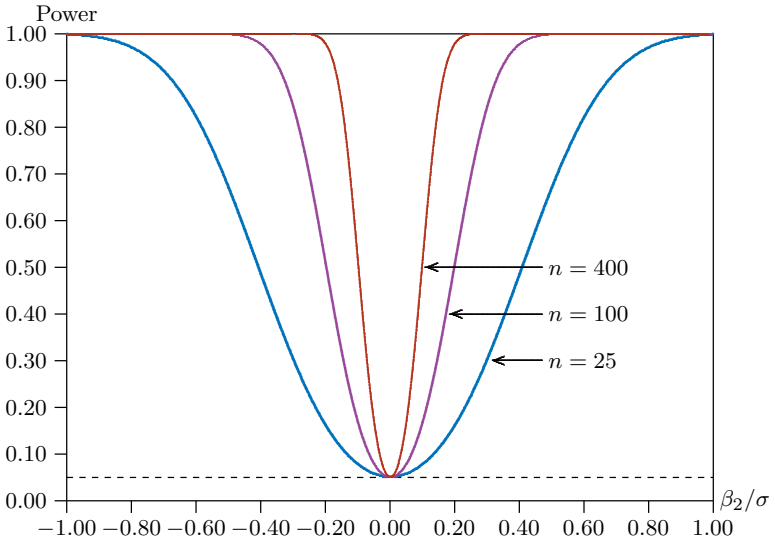


Figure 5.9 Power functions for t tests at the .05 level

can see from the figure that a false hypothesis is very unlikely to be rejected if $n^{1/2}\beta_2/\sigma$ is sufficiently small.

Because t tests in the classical normal linear regression model are exact, the case shown in Figure 5.9 is an ideal one. Tests that are valid only asymptotically may have power functions that look quite different from the ones in the figure. Power may be greater or less than .05 when the null hypothesis holds, depending on whether the test overrejects or underrejects, and it may well be minimized at a parameter value that does not correspond to the null. Instead of being a symmetric inverted bell shape, the power function may be quite asymmetrical, and in some cases power may not even tend to unity as the parameter under test becomes infinitely far from the null hypothesis. Readers are asked to investigate a less than ideal case in Exercise 5.24.

5.9 Pretesting

In regression analysis, interest often centers on certain explanatory variables only. The other explanatory variables are generally included solely to avoid possible misspecification. Consider the linear regression model (4.65), which was discussed in Section 4.8 and is rewritten here for convenience:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\boldsymbol{I}). \tag{5.76}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Here $\boldsymbol{\beta}$ is a k–vector, $\boldsymbol{\gamma}$ is an r–vector, and the regressors in $\boldsymbol{X}$ and $\boldsymbol{Z}$ are assumed, for simplicity, to be exogenous, so that, in all that follows, we condition on them. The parameters of interest are the k elements of $\boldsymbol{\beta}$. We would like to estimate them as well as possible, but we do not care about $\boldsymbol{\gamma}$. Instead, we would like to choose between the unrestricted estimator $\hat{\boldsymbol{\beta}}$ obtained by running the regression (5.76) and the restricted estimator $\tilde{\boldsymbol{\beta}}$ from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ alone, setting $\boldsymbol{\gamma}$ equal to zero.

Except in the very special case in which the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are orthogonal, the restricted estimator $\tilde{\boldsymbol{\beta}}$ is more efficient than the unrestricted estimator $\hat{\boldsymbol{\beta}}$. However, because the estimator $\tilde{\boldsymbol{\beta}}$ is biased if $\boldsymbol{\gamma} \neq \boldsymbol{0}$, its mean squared error matrix is larger than its covariance matrix in that case; recall equation (4.76).

Since $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\gamma}$ is zero, it seems natural to test the hypothesis that $\boldsymbol{\gamma} = \boldsymbol{0}$ and use the latter estimator when the test rejects and the former when it does not. This test is called a **preliminary test**, or **pretest** for short. Such a procedure implicitly defines a new estimator, which is called a **pretest estimator**. Formally, we can write

$$\acute{\boldsymbol{\beta}} = \mathbb{I}(F_{\boldsymbol{\gamma}=\boldsymbol{0}} > c_\alpha)\hat{\boldsymbol{\beta}} + \mathbb{I}(F_{\boldsymbol{\gamma}=\boldsymbol{0}} \leq c_\alpha)\tilde{\boldsymbol{\beta}}, \tag{5.77}$$

where $F_{\boldsymbol{\gamma}=\boldsymbol{0}}$ is the F statistic for the hypothesis that $\boldsymbol{\gamma} = \boldsymbol{0}$, and c_α is the critical value for an F test with r and $n-k-r$ degrees of freedom at level α. Thus $\acute{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ when the pretest rejects, and $\acute{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ when the pretest does not reject.

Equation (5.77) for the pretest estimator can be written in a simpler form as

$$\acute{\boldsymbol{\beta}} = \hat{\lambda}\hat{\boldsymbol{\beta}} + (1-\hat{\lambda})\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \hat{\lambda}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}), \tag{5.78}$$

where $\hat{\lambda} = 1$ when $F_{\boldsymbol{\gamma}=\boldsymbol{0}} > c_\alpha$ and $\hat{\lambda} = 0$ when $F_{\boldsymbol{\gamma}=\boldsymbol{0}} \leq c_\alpha$. In this form, $\acute{\boldsymbol{\beta}}$ looks like a weighted average of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, but with random weights that can only equal 0 or 1.

The MSE Matrix of the Pretest Estimator

Because the outcome of the pretest is random, the MSE matrix of the pretest estimator is not simply a weighted average of the variance of the unbiased estimator $\hat{\boldsymbol{\beta}}$ and the MSE matrix for $\tilde{\boldsymbol{\beta}}$. The problem was first analyzed by Magnus and Durbin (1999) and then by Danilov and Magnus (2004) under the assumptions of the classical normal linear model, according to which the disturbances are normally distributed.

Let the regression (5.76) represent the DGP, which is thereby assumed to have true parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and σ^2. Then we have

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\gamma} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}, \tag{5.79}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where the $k \times r$ matrix $\boldsymbol{Q}$ is equal to $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}$. From the estimating equations for the restricted and unrestricted regressions, we see that

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) = \boldsymbol{0} \quad \text{and}$$
$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}) = \boldsymbol{0},$$

where $\hat{\boldsymbol{\gamma}}$ is the OLS estimator from the unrestricted regression. By subtracting one of these equations from the other, we find that

$$\boldsymbol{X}^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\gamma}}) = \boldsymbol{0},$$

and, on premultiplying by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ and using (5.79), we conclude that

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} - \boldsymbol{Q}\hat{\boldsymbol{\gamma}} = \boldsymbol{\beta} - \boldsymbol{Q}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}. \tag{5.80}$$

Then, from (5.78), (5.79), and (5.80),

$$\acute{\boldsymbol{\beta}} - \boldsymbol{\beta} = -\boldsymbol{Q}(\hat{\lambda}\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}. \tag{5.81}$$

The MSE matrix for the pretest estimator $\acute{\boldsymbol{\beta}}$ is therefore the second-moment matrix of the right-hand side of equation (5.81), that is,

$$\begin{aligned}
&\mathrm{E}\big((\acute{\boldsymbol{\beta}} - \boldsymbol{\beta})(\acute{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\big) \\
&= \mathrm{E}\big[\big(-\boldsymbol{Q}(\hat{\lambda}\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}\big)\big(\boldsymbol{u}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} - (\hat{\lambda}\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top\boldsymbol{Q}^\top\big)\big] \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \boldsymbol{Q}\mathrm{E}\big[(\hat{\lambda}\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\hat{\lambda}\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top\big]\boldsymbol{Q}^\top.
\end{aligned} \tag{5.82}$$

The last step here follows exactly when, as we supposed, the disturbances are normal. In that case, $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{u}$ is independent of $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, since the fact that these vectors are uncorrelated implies that they are independent. It follows that $\boldsymbol{X}^\top\boldsymbol{u} = \boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{u}$ is independent of $\hat{\boldsymbol{\gamma}}$, since $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = (\boldsymbol{Z}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, and also of the variance estimator s^2 from the unrestricted model, which is the only other random element in the F statistic, and so in $\hat{\lambda}$. Thus, since $\mathrm{E}(\boldsymbol{X}^\top\boldsymbol{u}) = \boldsymbol{0}$, the expectations of the cross terms in the middle line of (5.82) vanish.

If we regard $\hat{\lambda}\hat{\boldsymbol{\gamma}}$ as a biased estimator of $\boldsymbol{\gamma}$, then (5.82) can be interpreted as

$$\mathrm{MSE}(\acute{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \boldsymbol{Q}\,\mathrm{MSE}(\hat{\lambda}\hat{\boldsymbol{\gamma}})\,\boldsymbol{Q}^\top. \tag{5.83}$$

Properties of Pretest Estimators

The result (5.83) allows us to compare the MSE of the pretest estimator $\acute{\boldsymbol{\beta}}$ with the MSEs of the restricted and unrestricted estimators. This comparison

Copyright © 2021, Russell Davidson and James G. MacKinnon

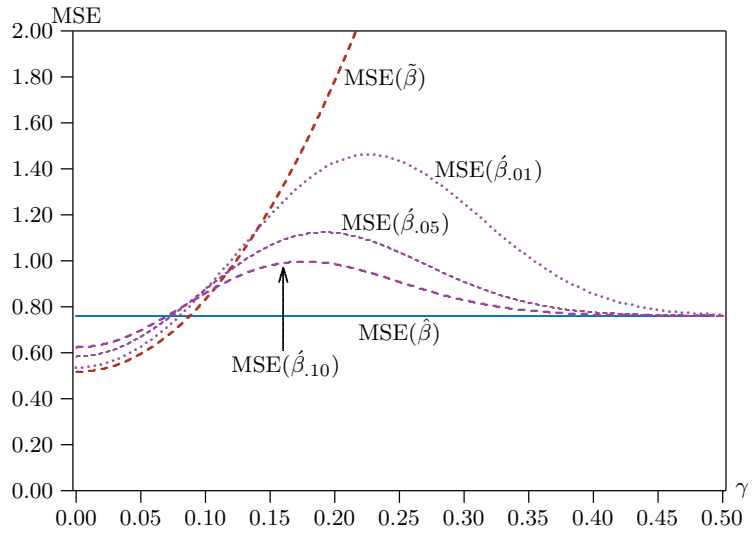


Figure 5.10 MSEs of Several Estimators

turns out to be quite illuminating. For simplicity, we confine our attention to the model

$$\boldsymbol{y} = \beta\boldsymbol{x} + \gamma\boldsymbol{z} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{NID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{5.84}$$

in which there is just one parameter of interest and one restriction, and MSE is therefore a scalar rather than a matrix. We would get very similar results if there were several parameters of interest and/or several restrictions. We assume that the two regressors are bivariate normal with correlation $\rho = 0.5$. The potential reduction in variance from using the restricted estimator $\tilde{\beta}$ or the pretest estimator $\acute{\beta}$ rather than the unrestricted estimator $\hat{\beta}$ is evidently increasing in $|\rho|$, but so is the potential bias.

Figure 5.10 shows the MSE for five different estimators of β as functions of γ in the model (5.84). The horizontal line is the MSE of the unrestricted OLS estimator, $\hat{\beta}$. It is the only unbiased estimator here, and therefore it is the only one for which the MSE does not depend on γ.

The MSE of the restricted estimator $\tilde{\beta}$ is lower than $\mathrm{MSE}(\hat{\beta})$ when γ is sufficiently small. However, as γ increases, $\mathrm{MSE}(\tilde{\beta})$ increases in proportion to γ^2 (in other words, in proportion to the NCP), rapidly becoming so large that it is impossible to show it on the figure. If ρ had been larger, $\mathrm{MSE}(\tilde{\beta})$ would have increased even more rapidly.

The other three estimators are all pretest estimators. They differ only in the level of the pretest, which is .10, .05, or .01, and they are therefore denoted $\acute{\beta}_{.10}$, $\acute{\beta}_{.05}$, and $\acute{\beta}_{.01}$. The three MSE functions have similar shapes, but they become substantially more extreme as the level of the pretest becomes smaller.

Copyright © 2021, Russell Davidson and James G. MacKinnon

For small values of γ, the pretest estimators are more efficient than $\hat{\beta}$ but less efficient than $\tilde{\beta}$. For very large values of γ, the pretest estimators perform essentially the same as $\hat{\beta}$, presumably because the pretests always reject.

There is a large region in the middle of the figure where the pretest estimators perform better than $\tilde{\beta}$ but less well than $\hat{\beta}$. The increase in MSE, especially for $\acute{\beta}_{.01}$, is very substantial over a wide range of values of γ. For each of the pretest estimators, there is also a fairly small region near the point where $\text{MSE}(\tilde{\beta})$ crosses $\text{MSE}(\hat{\beta})$ for which that pretest estimator performs worse than either $\tilde{\beta}$ or $\hat{\beta}$.

Figure 5.10 makes it clear that the level of the pretest is important. When the level is relatively high, the potential gain in efficiency for small values of γ is smaller, but the potential increase in MSE due to bias for intermediate values is very much smaller. Thus there is absolutely no reason to use a "conventional" significance level like .05 when pretesting, and it is probably safer to use a higher level.

5.10 Final Remarks

This chapter has introduced a number of important concepts. Later, we will encounter many types of hypothesis test, sometimes exact but more commonly asymptotic. Some of the asymptotic tests work well in finite samples, but others emphatically do not. In Chapter 7, we will introduce the concept of bootstrap tests, which often work very much better than asymptotic tests when exact tests are not available.

Although hypothesis testing plays a central role in classical econometrics, it is not the only method by which econometricians attempt to make inferences from parameter estimates about the true values of parameters. In the next chapter, we turn our attention to the other principal method, namely, the construction of confidence intervals and confidence regions.

5.11 Appendix: Linear Combinations of Normal Variables

An important property of the normal distribution, used in our discussion in Section 5.2 and essential to the derivation of the multivariate normal distribution in Section 5.3, is that any linear combination of independent normally distributed random variables is itself normally distributed. To see this, it is enough to show it for independent standard normal variables, because, by (5.10), all normal variables can be generated as linear combinations of standard normal ones plus constants.

The proof given here of this property makes use of the concept of the **moment-generating function** or **MGF** of a distribution. We make no further use of the

Copyright © 2021, Russell Davidson and James G. MacKinnon

concept in this book, and so readers may safely skip this Appendix if they are willing to take the result on trust, or else consult the rather clumsy, but more elementary, proof in ETM. If X is a scalar random variable, the moment-generating function of its distribution is defined to be $m_X(t) \equiv \text{E}(\text{e}^{tX})$. It can be shown – see for instance Billingsley (1995) – that the MGF of a r.v. X uniquely determines its distribution. The MGF of the standard normal distribution is calculated as follows: Let $Z \sim \text{N}(0,1)$, then

$$\text{E}(\text{e}^{tZ}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \text{e}^{tz} \text{e}^{-z^2/2} \, \text{d}z.$$

The exponent of the exponential in the integrand can be written as

$$-\tfrac{1}{2}z^2 + tz = -\tfrac{1}{2}(z^2 - 2tz) = -\tfrac{1}{2}\big((z-t)^2 - t^2\big),$$

and so, on changing the integration variable to $y = z - t$, w see that the MGF is

$$\text{e}^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \text{e}^{-y^2/2} \, \text{d}y = \text{e}^{t^2/2}.$$

A r.v. X that has the $\text{N}(\mu, \sigma^2)$ distribution can be represented as $\mu + \sigma Z$, with $Z \sim \text{N}(0,1)$. It is clear that the MGF of X is

$$\text{E}(\text{e}^{tX}) = \text{e}^{t\mu} \text{E}(\text{e}^{\sigma t Z}) = \exp(\mu t + \sigma^2 t^2/2).$$

Consider now the linear combination $W \equiv \sum_{i=1}^m a_i Z_i$ of independent standard normal variables Z_i, $i = 1, \ldots, m$. It is immediate that $\text{E}W = 0$ and $\text{Var}(W) \equiv \sigma_W^2 = \sum_{i=1}^m a_i^2$. The MGF of W is

$$m_W(t) = \text{E}(\text{e}^{tW}) = \text{E}\big(\exp(\sum_{i=1}^m t a_i Z_i)\big) = \text{E}\big(\prod_{i=1}^m \exp(t a_i Z_i)\big).$$

Since the Z_i are mutually independent, the expectation of the product is the product of the expectations:

$$m_W(t) = \prod_{i=1}^m \text{E}\big(\exp(t a_i Z_i)\big) = \prod_{i=1}^m \exp \tfrac{1}{2} t^2 a_i^2 = \exp \tfrac{1}{2}(t^2 \sum_{i=1}^m a_i^2).$$

The rightmost expression above is the MGF of the normal distribution with expectation zero and variance σ_W^2, from which it follows that W is normally distributed. It is easy to show that if an expectation μ is added to W, then W has the $\text{N}(\mu, \sigma_W^2)$ distribution.

Copyright © 2021, Russell Davidson and James G. MacKinnon

5.12 Exercises

5.1 Suppose that the random variable z has the $\mathrm{N}(0,1)$ density. If z is a test statistic used in an equal-tail test, the corresponding P value, according to (5.08), is $p(z) \equiv 2(1-\Phi(|z|))$. Show that $F_p(\cdot)$, the CDF of $p(z)$, is the CDF of the uniform distribution on $[0,1]$. In other words, show that

$$F_p(x) = x \quad \text{for all } x \in [0,1].$$

5.2 Extend Exercise 2.6 to show that the third and fourth moments of the standard normal distribution are 0 and 3, respectively. Use these results in order to calculate the centered and uncentered third and fourth moments of the $\mathrm{N}(\mu, \sigma^2)$ distribution.

5.3 Let the density of the random variable x be $f(x)$. Show that the density of the random variable $w \equiv tx$, where $t > 0$, is $(1/t)f(w/t)$. Next let the joint density of the set of random variables x_i, $i = 1, \ldots, m$, be $f(x_1, \ldots, x_m)$. For $i = 1, \ldots, m$, let $w_i = t_i x_i$, $t_i > 0$. Show that the joint density of the w_i is

$$f(w_1, \ldots, w_m) = \frac{1}{\prod_{i=1}^m t_i} f\Big(\frac{w_1}{t_1}, \ldots, \frac{w_m}{t_m}\Big).$$

***5.4** Consider the random variables x_1 and x_2, which are bivariate normal with $x_1 \sim \mathrm{N}(0, \sigma_1^2)$, $x_2 \sim \mathrm{N}(0, \sigma_2^2)$, and correlation ρ. Show that the expectation of x_1 conditional on x_2 is $\rho(\sigma_1/\sigma_2)x_2$ and that the variance of x_1 conditional on x_2 is $\sigma_1^2(1-\rho^2)$. How are these results modified if the expectations of x_1 and x_2 are μ_1 and μ_2, respectively?

5.5 Suppose that, as in the previous question, the random variables x_1 and x_2 are bivariate normal, with expectations 0, variances σ_1^2 and σ_2^2, and correlation ρ. Show that $f(x_1, x_2)$, the joint density of x_1 and x_2, is given by

$$\frac{1}{2\pi} \frac{1}{(1-\rho^2)^{1/2}\sigma_1\sigma_2} \exp\left(\frac{-1}{2(1-\rho^2)}\left(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)\right). \tag{5.85}$$

Then use this result to show that x_1 and x_2 are statistically independent if $\rho = 0$.

***5.6** Let the random variables x_1 and x_2 be distributed as bivariate normal, with expectations μ_1 and μ_2, variances σ_1^2 and σ_2^2, and covariance σ_{12}. Using the result of Exercise 5.5, write down the joint density of x_1 and x_2 in terms of the parameters just specified. Then find the marginal density of x_1.

What is the density of x_2 conditional on x_1? Show that the expectation of x_2 conditional on x_1 can be written as $\mathrm{E}(x_2 \,|\, x_1) = \beta_1 + \beta_2 x_1$, and solve for the parameters β_1 and β_2 as functions of the parameters of the bivariate distribution. How are these parameters related to the least-squares estimates that would be obtained if we regressed realizations of x_2 on a constant and realizations of x_1?

5.7 (For readers comfortable with moment-generating functions) The multivariate moment-generating function of an m-vector $\boldsymbol{x}$ of random variables is defined as a function of the m-vector $\boldsymbol{t}$ as follows:

$$m_{\boldsymbol{W}}(\boldsymbol{t}) = \mathrm{E}(\exp \boldsymbol{t}^\top \boldsymbol{x}).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Show that the MGF of the multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, where $\boldsymbol{\mu}$ is an m-vector and $\boldsymbol{\Omega}$ an $m \times m$ matrix, is $\exp \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Omega}\, \boldsymbol{t}$.

5.8 Consider the linear regression model

$$y_t = \beta_1 + \beta_2 x_{t1} + \beta_3 x_{t2} + u_t.$$

Rewrite this model so that the restriction $\beta_2 - \beta_3 = 1$ becomes a single zero restriction.

***5.9** Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, where there are n observations and k regressors. Suppose that this model is potentially subject to r restrictions which can be written as $\boldsymbol{R\beta} = \boldsymbol{r}$, where $\boldsymbol{R}$ is an $r \times k$ matrix and $\boldsymbol{r}$ is an r–vector. Rewrite the model so that the restrictions become r zero restrictions.

***5.10** Show that the t statistic (5.22) is $(n-k)^{1/2}$ times the cotangent of the angle between the n–vectors $\boldsymbol{M}_1\boldsymbol{y}$ and $\boldsymbol{M}_1\boldsymbol{x}_2$.

Now consider the regressions

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}, \text{ and} \\ \boldsymbol{x}_2 &= \boldsymbol{X}_1\boldsymbol{\gamma}_1 + \gamma_2\boldsymbol{y} + \boldsymbol{v}. \end{aligned} \tag{5.86}$$

What is the relationship between the t statistic for $\beta_2 = 0$ in the first of these regressions and the t statistic for $\gamma_2 = 0$ in the second?

5.11 Show that the OLS estimates $\tilde{\boldsymbol{\beta}}_1$ from the restricted model (5.26) can be obtained from those of the unrestricted model (5.25) by the formula

$$\tilde{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1 + (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2.$$

Hint: Equation (5.35) is useful for this exercise.

5.12 Consider regressions (5.40) and (5.39), which are numerically equivalent. Drop the normality assumption and assume that the disturbances are merely IID. Show that the SSR from these regressions is equal to the sum of the SSRs from the two subsample regressions:

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}_1, \quad \boldsymbol{u}_1 \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \text{ and} \\ \boldsymbol{y}_2 &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}_2, \quad \boldsymbol{u}_2 \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}). \end{aligned}$$

5.13 When performing a Chow test, one may find that one of the subsamples is smaller than k, the number of regressors. Without loss of generality, assume that $n_2 < k$. Show that, in this case, the F statistic becomes

$$\frac{(\mathrm{RSSR} - \mathrm{SSR}_1)/n_2}{\mathrm{SSR}_1/(n_1 - k)},$$

and that the numerator and denominator really have the degrees of freedom used in this formula.

5.14 Prove that the F statistic (5.30) for $\boldsymbol{\beta}_2 = \boldsymbol{0}$ in equation (5.25) is independent, under the null hypothesis, of the vector of restricted estimates $\tilde{\boldsymbol{\beta}}_1$ when the disturbances are normally, identically, and independently distributed.

Copyright © 2021, Russell Davidson and James G. MacKinnon

5.15 Show that the square of the t statistic (5.22) is a special case of the Wald statistic (5.66). Recall that this statistic is testing the hypothesis that $\beta_2 = 0$ in the linear regression model (5.18).

5.16 Write the restrictions that are being tested by the F statistic (5.30) in the form of equations (5.65), and show that r times the F statistic is a special case of the Wald statistic (5.66).

5.17 The file **house-price-data.txt** contains 546 observations. Regress the logarithm of the house price on a constant, the logarithm of lot size, and the other ten explanatory variables, as in Exercise 2.23.

One of the explanatory variables is the number of storeys, which can take on the values 1, 2, 3, and 4. A more general specification would allow the effect on log price of each number of storeys to be different. Test the original model against this more general one using an F test. Report the test statistic, the degrees of freedom, and the P value.

Test the same hypothesis again, this time using a Wald test. Write out the vectors of restrictions and their covariance matrix, both of which appear in equation (5.66), explicitly. Once again, report the test statistic, the degrees of freedom, and the P value.

5.18 Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathbf{I}), \quad \mathrm{E}(\boldsymbol{u} \,|\, \boldsymbol{X}) = \boldsymbol{0},$$

where $\boldsymbol{X}$ is an $n \times k$ matrix. If σ_0 denotes the true value of σ, how is the quantity $\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}/\sigma_0^2$ distributed? Use this result to derive a test of the null hypothesis that $\sigma = \sigma_0$. Is this a one-tailed test or a two-tailed test?

⋆**5.19** P values for two-tailed tests based on statistics that have asymmetric distributions are not calculated as in Section 5.2. Let the CDF of the statistic τ be denoted as F, where $F(-x) \neq 1 - F(x)$ for general x. Suppose that, for any level α, the critical values c_α^- and c_α^+ are defined, analogously to (5.06), by the equations

$$F(c_\alpha^-) = \alpha/2 \quad \text{and} \quad F(c_\alpha^+) = 1 - \alpha/2.$$

Show that the marginal significance level, or P value, associated with a realized statistic $\hat{\tau}$ is $2\min(F(\hat{\tau}), 1 - F(\hat{\tau}))$.

5.20 The file **house-price-data.txt** contains 546 observations. Regress the logarithm of the house price on a constant, the logarithm of lot size, and the other ten explanatory variables, as in Exercise 5.17. Obtain a sensible estimate of σ, the standard deviation of the disturbances. Then test the hypothesis that $\sigma = 0.20$ at the .05 level. Report a P value for the test. **Hint:** See Exercises 5.18 and 4.19.

Now test the hypothesis that $\sigma \leq 0.20$ at the .05 level. Report a P value for the test. Comment on the results of the two tests.

5.21 Suppose that z is a test statistic distributed as $\mathrm{N}(0, 1)$ under the null hypothesis, and as $\mathrm{N}(\lambda, 1)$ under the alternative, where λ depends on the DGP that generates the data. If c_α is defined by (5.07), show that the power of the two-tailed test at level α based on z is equal to

$$\Phi(\lambda - c_\alpha) + \Phi(-c_\alpha - \lambda).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Plot this power function for λ in the interval $[-5, 5]$ for $\alpha = .05$ and $\alpha = .01$.

5.22 Show that, if the m–vector $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$, the expectation of the noncentral chi-squared variable $\boldsymbol{z}^\top \boldsymbol{z}$ is $m + \boldsymbol{\mu}^\top \boldsymbol{\mu}$.

5.23 Consider the linear regression model with n observations,

$$\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{d} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{NID}(\boldsymbol{0}, \sigma^2\mathbf{I}). \tag{5.87}$$

The only regressor here is a dummy variable, with each element equal to 1 for n_1 observations and equal to 0 for the remaining $n - n_1$ observations.

First, find the standard error of $\hat{\beta}_2$ as a function of n, n_1, and σ. Then find the probability that a test at the .05 level will reject the null hypothesis that $\beta_2 = 0$ as a function of the standard error and β_2. Using these results, what is the smallest sample size for which you could reject the null hypothesis that $\beta_2 = 0$ with probability at least 0.9 when $\sigma = 1$ and the true value of β_2 is 0.5? Assume that n_1 is chosen optimally given n; see Exercise 4.11.

Again assuming that n_1 is chosen optimally, graph the smallest sample size for which you could reject the null hypothesis that $\beta_2 = 0$ in equation (5.87) with probability at least 0.9 when $\sigma = 1$ against the true value of β_2 for $\beta_2 = 0.1, 0.2, \ldots, 1.0$. Use a logarithmic scale for the vertical axis.

5.24 Consider the exact test for $\sigma = \sigma_0$ in the classical normal linear model (5.17) that was derived in Exercise 5.18, and suppose that $k = 3$. Plot the power function for this test at the .05 level for the null hypothesis that $\sigma = 1$ over the interval $0.5 < \sigma < 2.0$ for three values of the sample size, namely, $n = 13$, $n = 23$, and $n = 43$. **Hint:** This exercise does not require any simulations, but it does require you to calculate the cumulative χ^2 distribution function many times and its inverse a few times.

5.25 Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}),$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ denote, respectively, $n \times k_1$ and $n \times k_2$ matrices of predetermined regressors, with $k = k_1 + k_2$. Write down the asymptotic results for $\hat{\boldsymbol{\beta}}_1$ that are analogous to equations (5.55) and (5.56).

5.26 The file **earnings-data.txt** contains 46,302 observations on 32 variables taken from the Current Population Survey. Each observation is for a woman who lived and worked in California in the specified year. The variables are:

`earn` = reported weekly earnings in current dollars

`age` = age in years

`ed1` = did not finish high school

`ed2` = high school graduate

`ed3` = some college or associate's degree

`ed4` = four-year college degree

`ed5` = at least one postgraduate degree

`year` = calendar year (1992 to 2015)

`y92` ... `y15` = year dummy variables constructed from `year`

Copyright © 2021, Russell Davidson and James G. MacKinnon

Regress the log of earnings on a constant, `age`, `age`$^2/100$, the four highest education dummies, and `year`. What was the effect of dividing `age`^{2} by 100? Why was this a sensible thing to do?

Add as many of the year dummy variables to this regression as you can. How many can you add? Does it matter which ones you add? Do the coefficient and standard errors suggest that these regressors are needed?

Test the hypothesis that the coefficients on these year dummy variables are all equal to zero.

5.27 Using the data in the file **earnings-data.txt**, run a regression that has exactly the same explanatory power as the second regression in Exercise 5.26 but which does not contain a constant term and does not include the `year` variable. Explain the relationship between the coefficients on the variable `y15` in the two regressions.

5.28 Reformulate the regression of Exercise 5.27 so that one of the coefficients measures the change in log earnings from obtaining a postgraduate degree. Test the hypothesis that this difference is 0.20.

5.29 Create two dummy variables, **young** and **old**. The first of these is 1 if **age** ≤ 35, and the second is 1 if **age** ≥ 60. Add the two dummies to the regression of Exercise 5.27, and perform both an F test and a Wald test of the hypothesis that neither of them actually belongs in the regression. Report P values for both tests. How do you interpret the results of these tests?

5.30 The regression of Exercise 5.27 implies that the expectation of log earnings first increases and then decreases with age. At what age is this expectation maximized? Test the hypothesis that the age at which it is maximized is actually 50. **Hint:** See Section 4.5.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 6

Confidence Sets and Sandwich Covariance Matrices

6.1 Introduction

Hypothesis testing, which was the subject of the previous chapter, is the foundation for all inference in classical econometrics. It can be used to find out whether restrictions imposed by economic theory are compatible with the data, and whether various aspects of the specification of a model appear to be correct. However, once we are confident that a model is correctly specified and incorporates whatever restrictions are appropriate, we often want to make inferences about the values of some of the parameters that appear in the model. Although this can be done by performing a battery of hypothesis tests, it is usually more convenient to construct **confidence sets** for the individual parameters of specific interest.

In order to construct a confidence set, we need a suitable **family of tests** for a set of point null hypotheses. A different test statistic must be calculated for each different null hypothesis that we consider, but usually there is just one *type* of statistic that can be used to test all the different null hypotheses. For instance, if we wish to test the hypothesis that a scalar parameter θ in a regression model equals 0, we can use a t test. But we can also use a t test for the hypothesis that $\theta = \theta_0$ for any specified real number θ_0. Thus, in this case, we have a family of t statistics indexed by θ_0.

Given a family of tests capable of testing a set of hypotheses about a (scalar) parameter θ, all with the same level α, we can use these tests to construct a confidence set for the parameter. By definition, a confidence set is a subset of the real line that contains all values θ_0 for which the hypothesis that $\theta = \theta_0$ is not rejected by the appropriate test in the family. For level α, a confidence set so obtained is said to be a $1-\alpha$ confidence set, or to be at **confidence level** $1-\alpha$. For a scalar parameter, a confidence set is normally an interval of the real line, hence the term **confidence interval**. In applied work, .95 confidence intervals are particularly popular, followed by .99 and .90 ones.

Unlike the parameters we are trying to make inferences about, confidence sets are random. Every different sample that we draw from the same DGP yields a different confidence set. The probability that the random set includes, or

covers, the true value of the parameter is called the **coverage probability**, or just the **coverage**, of the set. Suppose that all the tests in the family have exactly level α, that is, they reject their corresponding null hypotheses with probability exactly equal to α when the hypothesis is true. Let the value of θ under the true DGP be θ_0. Then θ_0 is contained in the confidence set if and only if the hypothesis that $\theta = \theta_0$ is not rejected. The probability of this event, and so the coverage probability, is exactly $1 - \alpha$.

Confidence intervals may be either **exact** or **approximate**. When the exact distribution of the test statistics used to construct a confidence interval is known, the coverage is equal to the confidence level, and the interval is exact. Otherwise, we have to be content with approximate confidence intervals, which may be based either on asymptotic theory or on the bootstrap. In the next section, we discuss both exact confidence intervals and approximate ones based on asymptotic theory. In Chapter 7, after we have introduced bootstrap hypothesis tests, we will discuss bootstrap confidence intervals.

When we are interested in two or more parameters jointly, it can be more informative to construct a **confidence region** instead of, or in addition to, several confidence intervals. The confidence region for a set of k model parameters, such as the components of a k–vector $\boldsymbol{\theta}$, is a k–dimensional subset of E^k, often the k–dimensional analog of an ellipse. The region is constructed in such a way that, for every point represented by the k–vector $\boldsymbol{\theta}_0$ in the confidence region, the joint hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is not rejected by the appropriate member of a family of tests at level α. Confidence regions constructed in this way cover the true values of the parameter vector $100(1 - \alpha)\%$ of the time, either exactly or approximately. In Section 6.3, we show how to construct confidence regions and explain the relationship between confidence regions and confidence intervals.

In earlier chapters, we assumed that the disturbances in linear regression models are independently and identically distributed. This assumption yields a simple form for the true covariance matrix of a vector of OLS parameter estimates, expression (4.38), and a simple way of estimating this matrix. However, it is excessively restrictive in many cases. In Sections 6.4, 6.5, and 6.6, we relax the IID assumption. In Section 6.4, we develop methods for obtaining **heteroskedasticity-robust standard errors** that can be used even when the form of the heteroskedasticity is unknown. These simple and widely-used methods are based on "sandwich" covariance matrix estimators.

In Section 6.5, we further weaken the assumptions needed to estimate the covariance matrix of a vector of OLS estimates by allowing for autocorrelation of the disturbances, thereby obtaining **heteroskedasticity and autocorrelation consistent**, or **HAC**, covariance matrix estimators. Then, in Section 6.6, we allow the disturbances to be dependent within "clusters" of observations. This yields methods for obtaining **cluster-robust standard errors**. In Section 6.7, we then discuss a widely-used class of empirical regression models, called **difference in differences**, for which it is very commonly assumed that the disturbances are clustered. Finally, in Section 6.8, we discuss the **delta method**, a procedure for obtaining standard errors, estimated covariance matrices, and approximate confidence intervals for nonlinear functions of estimated parameters.

Copyright © 2021, Russell Davidson and James G. MacKinnon

6.2 Exact and Asymptotic Confidence Intervals

A confidence interval for some scalar parameter θ consists of all values θ_0 for which the hypothesis $\theta = \theta_0$ cannot be rejected at some specified level α. Thus, as we will see in a moment, we can construct a confidence interval by "inverting" a test statistic. If the finite-sample distribution of the test statistic is known, we obtain an **exact confidence interval**. If, as is more commonly the case, only the asymptotic distribution of the test statistic is known, we obtain an **asymptotic confidence interval**, which may or may not be reasonably accurate in finite samples. Whenever a test statistic based on asymptotic theory has poor finite-sample properties, a confidence interval based on that statistic has poor coverage: In other words, the interval does not cover the true parameter value with the specified probability. In such cases, it may well be worthwhile to seek other test statistics that yield different confidence intervals with better coverage.

To begin with, suppose that we wish to base a confidence interval for the parameter θ on a family of test statistics that have a distribution or asymptotic distribution like the F or the χ^2 distribution under their respective nulls. Statistics of this type are always positive, and tests based on them reject their null hypotheses when the statistics are sufficiently large. Such tests are often equivalent to two-tailed tests based on statistics distributed as standard normal or Student's t. Let us denote the test statistic for the hypothesis that $\theta = \theta_0$ by the random variable $\tau(\boldsymbol{y}, \theta_0)$. Here $\boldsymbol{y}$ denotes the sample used to compute the particular realization of the statistic. It is the random element in the statistic, since $\tau(\cdot)$ is just a deterministic function of its arguments.

For each θ_0, the test consists of comparing the realized $\tau(\boldsymbol{y}, \theta_0)$ with the level-α critical value of the distribution of the statistic under the null. If the critical value is c_α, then any value θ_0 belongs to the confidence set if and only if

$$\tau(\boldsymbol{y}, \theta_0) \leq c_\alpha. \tag{6.01}$$

If θ_0 happens to be the parameter for the true DGP, then by the definition of c_α we have that

$$\Pr\big(\tau(\boldsymbol{y}, \theta_0) \leq c_\alpha\big) = 1 - \alpha. \tag{6.02}$$

Thus the true θ_0 is included in the (random) confidence set with probability $1 - \alpha$, so that the confidence level is equal to the coverage. But if c_α is a critical value for the asymptotic distribution of $\tau(\boldsymbol{y}, \theta_0)$, rather than for the exact distribution, then (6.02) is only approximately true.

Copyright © 2021, Russell Davidson and James G. MacKinnon

For concreteness, let us suppose that

$$\tau(\boldsymbol{y}, \theta_0) \equiv \left(\frac{\hat{\theta} - \theta_0}{s_\theta}\right)^2, \tag{6.03}$$

where $\hat{\theta}$ is an estimate of θ, and s_θ is the corresponding standard error, that is, an estimate of the standard deviation of $\hat{\theta}$. If $\hat{\theta}$ is an OLS estimate of a regression coefficient, then $\tau(\boldsymbol{y}, \theta_0)$ is the square of the t statistic for the hypothesis that $\theta = \theta_0$. Under the conditions of the classical normal linear model, discussed in Section 5.4, the test statistic (6.03) would be distributed as $F(1, n-k)$ under the null hypothesis. Therefore, the critical value c_α would be the level-α critical value of the $F(1, n-k)$ distribution. More generally, we will refer to any statistic of the form $(\hat{\theta} - \theta)/s_\theta$ as an **asymptotic t statistic**.

The values of θ on the boundary of the confidence set are the solutions of the equation

$$\tau(\boldsymbol{y}, \theta) = c_\alpha \tag{6.04}$$

for θ. Since the statistic (6.03) is a quadratic function of θ_0, this equation has two solutions. One of these solutions is the upper limit, θ_u, and the other is the lower limit, θ_l, of an interval, which is in fact the confidence interval that we are trying to construct, because it is clear that, for values of θ inside the interval, the inequality (6.01) is satisfied.

By using the formula (6.03) for the left-hand side of equation (6.04), taking the square root of both sides, and multiplying by s_θ, we obtain

$$|\hat{\theta} - \theta| = s_\theta c_\alpha^{1/2}. \tag{6.05}$$

As expected, there are two solutions to equation (6.05). These are

$$\theta_l = \hat{\theta} - s_\theta c_\alpha^{1/2} \quad \text{and} \quad \theta_u = \hat{\theta} + s_\theta c_\alpha^{1/2},$$

and so the $1 - \alpha$ confidence interval for θ is

$$\left[\hat{\theta} - s_\theta c_\alpha^{1/2},\ \ \hat{\theta} + s_\theta c_\alpha^{1/2}\right]. \tag{6.06}$$

Quantiles

When we speak of critical values, we are implicitly making use of the concept of a **quantile** of the distribution that the test statistic follows under the null hypothesis. If $F(x)$ denotes the CDF of a random variable X, and if the density $f(x) \equiv F'(x)$ exists and is strictly positive on the entire range of possible values for X, then q_α, the **α quantile** of F, for $0 \leq \alpha \leq 1$, satisfies the equation $F(q_\alpha) = \alpha$. The assumption of a strictly positive density means that F is strictly increasing over its range. Therefore, the inverse function F^{-1} exists, and $q_\alpha = F^{-1}(\alpha)$. For this reason, F^{-1} is sometimes called the

Copyright © 2021, Russell Davidson and James G. MacKinnon

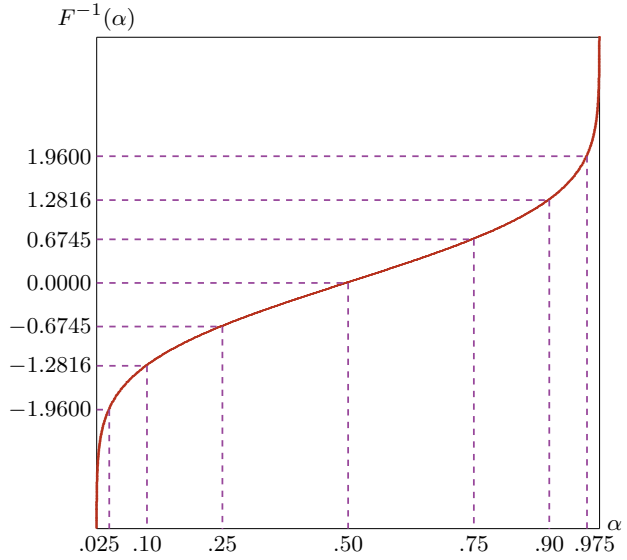


Figure 6.1 The quantile function of the standard normal distribution

quantile function. If F is not strictly increasing, or if the density does not exist, which, as we saw in Section 2.2, is the case for a discrete distribution, the α quantile does not necessarily exist, and is not necessarily uniquely defined, for all values of α.

The .5 quantile of a distribution is often called the **median**. For $\alpha = .25$, .5, and .75, the corresponding quantiles are called **quartiles**; for $\alpha = .2$, .4, .6, and .8, they are called **quintiles**; for $\alpha = i/10$ with i an integer between 1 and 9, they are called **deciles**; for $\alpha = i/20$ with $1 \leq i \leq 19$, they are called **vigintiles**; and, for $\alpha = i/100$ with $1 \leq i \leq 99$, they are called **centiles**, or, more frequently, **percentiles**. The quantile function of the standard normal distribution is shown in Figure 6.1. All three quartiles, the first and ninth deciles, and the .025 and .975 quantiles are shown in the figure.

Asymptotic Confidence Intervals

The interval (6.06) is exact only in the very restrictive circumstances of the classical normal linear model. If it is just an asymptotic interval, investigators may prefer to use the critical value given by the $\chi^2(1)$ distribution. For $\alpha = .05$, the critical value for the $\chi^2(1)$ distribution is the 0.95 quantile of the distribution, which is 3.8415, the square root of which is 1.9600. Thus the confidence interval given by (6.06) becomes

$$\left[\hat{\theta} - 1.96\, s_\theta,\ \ \hat{\theta} + 1.96\, s_\theta\right]. \tag{6.07}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

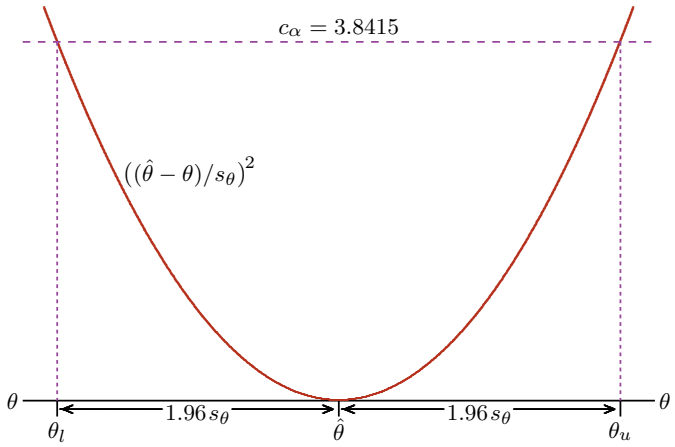


Figure 6.2 A symmetric confidence interval

This interval is shown in Figure 6.2, which illustrates the manner in which it is constructed. The value of the test statistic is on the vertical axis of the figure. The upper and lower limits of the interval occur at the values of θ where the test statistic (6.03) is equal to c_α, which in this case is 3.8415.

We would have obtained exactly the same confidence interval as (6.06) if we had started with the asymptotic t statistic $(\hat\theta - \theta_0)/s_\theta$ and used the $\mathrm{N}(0,1)$ distribution to perform a two-tailed test. For such a test, there are two critical values. If we wish to have the same probability mass in each tail, these are the $\alpha/2$ and the $1-\alpha/2$ quantiles of the distribution. It is conventional to denote these quantiles of the standard normal distribution by $z_{\alpha/2}$ and $z_{1-\alpha/2}$, respectively. Note that $z_{\alpha/2}$ is negative, since $\alpha/2 < 1/2$, and the median of the $\mathrm{N}(0,1)$ distribution is 0. By the symmetry of the normal distribution, it is easy to see that $z_{\alpha/2} = -z_{1-\alpha/2}$.

Equation (6.04), which has two solutions for a χ^2 test or an F test, is replaced by two equations, each with just one solution, as follows:

$$\tau(\boldsymbol{y}, \theta) = \pm c.$$

Here $\tau(\boldsymbol{y}, \theta)$ denotes the (signed) t statistic rather than the $\chi^2(1)$ statistic used in (6.04), and the positive number c can be defined either as $z_{1-\alpha/2}$ or as $-z_{\alpha/2}$. The resulting confidence interval $[\theta_l,\ \theta_u]$ can thus be written in two different ways:

$$\big[\hat\theta + s_\theta z_{\alpha/2},\ \hat\theta - s_\theta z_{\alpha/2}\big] \quad \text{and} \quad \big[\hat\theta - s_\theta z_{1-\alpha/2},\ \hat\theta + s_\theta z_{1-\alpha/2}\big]. \tag{6.08}$$

When $\alpha = .05$, we once again obtain the interval (6.07), since $z_{.025} = -1.96$ and $z_{.975} = 1.96$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Pivots

In order to explain why a confidence interval based on exact critical values has correct coverage, we need to introduce an important concept. A random variable with the property that its distribution is the same for all DGPs in a model $\mathbb{M}$ is said to be **pivotal**, or to be a **pivot**, for the model $\mathbb{M}$. The distribution is allowed to depend on the sample size, and perhaps on the observed values of exogenous variables. However, for any given sample size and set of exogenous variables, it must be invariant across all DGPs in $\mathbb{M}$.

A random function $\tau(\boldsymbol{y}, \theta)$ is said to be a **pivotal function** for $\mathbb{M}$, or just pivotal, if, when it is evaluated at the true value θ_0 corresponding to some DGP in $\mathbb{M}$, the result is a random variable whose distribution does not depend on what that DGP is. Pivotal functions of more than one model parameter are defined in exactly the same way. The function would merely be **asymptotically pivotal** if the asymptotic distribution were invariant to the choice of DGP but not the finite-sample distribution.

It is possible to construct an exact confidence interval based on a function $\tau(\boldsymbol{y}, \theta)$ only if this function is pivotal for the model $\mathbb{M}$ under consideration. Suppose that $\tau(\boldsymbol{y}, \theta)$ is an exact pivot. Then, the true θ_0 belongs to the confidence interval if and only if the inequality (6.01) holds, which, by (6.02), is an event of probability $1-\alpha$.

Even if it is not an exact pivot, the function $\tau(\boldsymbol{y}, \theta)$ must be asymptotically pivotal, since otherwise the critical value c_α would depend asymptotically on the unknown DGP in $\mathbb{M}$, and we could not construct a confidence interval with the correct coverage, even asymptotically. Of course, if c_α is only approximate, then the coverage of the interval will differ from $1-\alpha$ to a greater or lesser extent, in a manner that, in general, depends on the unknown true DGP.

The t and F statistics considered in Section 5.4 are pivots for the classical normal linear model subject to the null hypothesis under test. This is because, under that null model, they have their namesake distributions independently of the values of other regression parameters or the variance of the disturbances. Similarly, if $\hat\theta$ is an OLS estimator of a parameter of the classical normal linear model, the test statistic (6.03), or its signed squared root, which is just the conventional t statistic for the null hypothesis that $\theta = \theta_0$, is a pivotal function for the classical normal linear model. There are, however, very few exact pivots or pivotal functions outside the context of the classical normal linear model.

Asymmetric Confidence Intervals

The confidence interval (6.06), which is the same as the interval (6.08), is a **symmetric** one, because θ_l is as far below $\hat\theta$ as θ_u is above it. Although many confidence intervals are symmetric, not all of them share this property. The symmetry of (6.06) is a consequence of the symmetry of the standard normal distribution and of the form of the test statistic (6.03).

Copyright © 2021, Russell Davidson and James G. MacKinnon

It is possible to construct confidence intervals based on two-tailed tests even when the distribution of the test statistic is not symmetric. This evidently leads to an **asymmetric confidence interval**. For a chosen level α, we wish to reject whenever the statistic is too far into either the right-hand or the left-hand tail of the distribution. Unfortunately, there are many ways to interpret "too far" in this context. The simplest is probably to define the rejection region in such a way that there is a probability mass of $\alpha/2$ in each tail. This is called an **equal-tail confidence interval**. Two critical values are needed for each level, a lower one, c_α^-, which is the $\alpha/2$ quantile of the distribution, and an upper one, c_α^+, which is the $1-\alpha/2$ quantile. A realized statistic $\hat{\tau}$ leads to rejection at level α if either $\hat{\tau} < c_\alpha^-$ or $\hat{\tau} > c_\alpha^+$. Readers are asked to construct such an interval in Exercise 6.13.

It is also possible to construct confidence intervals based on one-tailed tests. Such an interval is open all the way out to infinity in one direction. Suppose that, for each θ_0, the null $\theta \leq \theta_0$ is tested against the alternative $\theta > \theta_0$. If the true parameter value is finite, we never want to reject the null for any θ_0 that substantially exceeds the true value. Consequently, the confidence interval is open out to plus infinity. Formally, the null is rejected only if the signed t statistic is algebraically greater than the appropriate critical value. For the $\mathrm{N}(0,1)$ distribution, this is $z_{1-\alpha}$ for level α. The null hypothesis $\theta \leq \theta_0$ is not rejected if $\tau(\boldsymbol{y}, \theta_0) \leq z_{1-\alpha}$, that is, if $\hat{\theta} - \theta_0 \leq s_\theta z_{1-\alpha}$, or, equivalently, $\theta_0 \geq \hat{\theta} - s_\theta z_{1-\alpha}$. The interval over which θ_0 satisfies this inequality is just

$$\big[\hat{\theta} - s_\theta z_{1-\alpha},\ \ +\infty\big] \tag{6.09}$$

Similarly, if the null hypothesis were $\theta \geq \theta_0$, the one-tailed interval would be

$$\big[-\infty,\ \ \hat{\theta} - s_\theta z_\alpha\big]. \tag{6.10}$$

Confidence Intervals for Regression Coefficients

In Section 5.4, we saw that, for the classical normal linear model, exact tests of linear restrictions on the parameters of the regression function are available, based on the t and F distributions. This implies that we can construct exact confidence intervals. Consider the classical normal linear model (5.18), in which the parameter vector $\boldsymbol{\beta}$ has been partitioned as $[\boldsymbol{\beta}_1 \vdots \beta_2]$, where $\boldsymbol{\beta}_1$ is a $(k-1)$–vector and β_2 is a scalar. The t statistic for the hypothesis that $\beta_2 = \beta_{20}$ for any particular value β_{20} can be written as

$$\frac{\hat{\beta}_2 - \beta_{20}}{s_2}, \tag{6.11}$$

where s_2 is the usual OLS standard error for $\hat{\beta}_2$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Any DGP in the model (5.18) satisfies $\beta_2 = \beta_{20}$ for some β_{20}. With the correct value of β_{20}, the t statistic (6.11) has the $t(n-k)$ distribution, and so

$$\Pr\left(t_{\alpha/2} \leq \frac{\hat{\beta}_2 - \beta_{20}}{s_2} \leq t_{1-\alpha/2}\right) = 1 - \alpha, \tag{6.12}$$

where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ denote the $\alpha/2$ and $1-\alpha/2$ quantiles of the $t(n-k)$ distribution, respectively. We can use equation (6.12) to find a $1-\alpha$ confidence interval for β_2. The left-hand side of the equation is equal to

$$\begin{aligned}
&\Pr\big(s_2 t_{\alpha/2} \leq \hat{\beta}_2 - \beta_{20} \leq s_2 t_{1-\alpha/2}\big) \\
&\quad = \Pr\big(-s_2 t_{\alpha/2} \geq \beta_{20} - \hat{\beta}_2 \geq -s_2 t_{1-\alpha/2}\big) \\
&\quad = \Pr\big(\hat{\beta}_2 - s_2 t_{\alpha/2} \geq \beta_{20} \geq \hat{\beta}_2 - s_2 t_{1-\alpha/2}\big).
\end{aligned}$$

Therefore, the confidence interval we are seeking is

$$\big[\hat{\beta}_2 - s_2 t_{1-\alpha/2},\ \ \hat{\beta}_2 - s_2 t_{\alpha/2}\big]. \tag{6.13}$$

At first glance, this interval may look a bit odd, because the upper limit is obtained by subtracting something from $\hat{\beta}_2$. What is subtracted is negative, however, because $t_{\alpha/2} < 0$, since it is in the lower tail of the t distribution. Thus the interval does in fact contain the point estimate $\hat{\beta}_2$.

It may still seem strange that the lower and upper limits of (6.13) depend, respectively, on the upper-tail and lower-tail quantiles of the $t(n-k)$ distribution. This actually makes perfect sense, however, as can be seen by looking at the infinite confidence interval (6.09) based on a one-tailed test. There, since the null is that $\theta \leq \theta_0$, the confidence interval must be open out to $+\infty$, and so only the lower limit of the confidence interval is finite. But the null is rejected when the test statistic is in the upper tail of its distribution, and so it must be the upper-tail quantile that determines the only finite limit of the confidence interval, namely, the lower limit. Readers are strongly advised to take some time to think this point through, since most people find it strongly counter-intuitive when they first encounter it, and they can accept it only after a period of reflection. The phenomenon is perfectly general for linear regression models, and is by no means a particular property of an exact confidence interval for the classical normal linear model.

It is easy to rewrite the confidence interval (6.13) so that it depends only on the positive, upper-tail, quantile, $t_{1-\alpha/2}$. Because Student's t distribution is symmetric, the interval (6.13) is the same as the interval

$$\big[\hat{\beta}_2 - s_2 t_{1-\alpha/2},\ \ \hat{\beta}_2 + s_2 t_{1-\alpha/2}\big]; \tag{6.14}$$

compare the two ways of writing the confidence interval (6.08). For concreteness, suppose that $\alpha = .05$ and $n - k = 32$. In this special case,

Copyright © 2021, Russell Davidson and James G. MacKinnon

$t_{1-\alpha/2} = t_{.975} = 2.037$. Thus the .95 confidence interval based on (6.14) extends from 2.037 standard errors below $\hat{\beta}_2$ to 2.037 standard errors above it. This interval is slightly wider than the interval (6.07), which is based on asymptotic theory.

We obtained the interval (6.14) by starting from the t statistic (6.11) and using Student's t distribution. As readers are asked to demonstrate in Exercise 6.2, we would have obtained precisely the same interval if we had started instead from the square of (6.11) and used the F distribution.

6.3 Confidence Regions

When we are interested in making inferences about the values of two or more parameters, it can be quite misleading to look at the confidence intervals for each of the parameters individually. By using confidence intervals, we are implicitly basing our inferences on the *marginal* distributions of the parameter estimates. However, if the estimates are not independent, the product of the marginal distributions may be very different from the joint distribution. In such cases, it makes sense to construct a confidence region.

The confidence intervals we discussed in the preceding section are all obtained by inverting t tests, whether exact or asymptotic, based on families of statistics of the form $(\hat{\theta} - \theta_0)/s_\theta$, possibly squared. If we wish instead to construct a confidence region, we must invert joint tests for several parameters. These are usually tests based on statistics that follow the F or χ^2 distributions, at least asymptotically.

A t statistic depends explicitly on a parameter estimate and its standard error. Similarly, many tests for several parameters depend on a vector of parameter estimates and an estimate of their covariance matrix. Even many statistics that appear not to do so, such as F statistics, actually do implicitly, as we will see shortly. Suppose that we are interested in $\boldsymbol{\theta}_2$, a subvector of the parameter vector $\boldsymbol{\theta}$. We have a k_2–vector of parameter estimates $\hat{\boldsymbol{\theta}}_2$, of which the covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\theta}}_2)$ can be estimated by $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_2)$. Then, in many circumstances, the Wald statistic

$$(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20})^\top \big(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_2)\big)^{-1} (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20}) \tag{6.15}$$

can be used to test the joint null hypothesis that $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$. The test statistic (6.15) is evidently a special case of the Wald statistic (5.66).

Exact Confidence Regions for Regression Parameters

Suppose that we want to construct a confidence region for the elements of the vector $\boldsymbol{\beta}_2$ in the classical normal linear model (5.25), which we rewrite here for ease of exposition:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{6.16}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are a k_1–vector and a k_2–vector, respectively. The F statistic that can be used to test the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is given in (5.30). If we wish instead to test $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$, then we can write (6.16) as

$$\boldsymbol{y} - \boldsymbol{X}_2\boldsymbol{\beta}_{20} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{6.17}$$

where $\boldsymbol{\gamma} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_{20}$, and test the hypothesis that $\boldsymbol{\gamma} = \boldsymbol{0}$. It is not hard to show that the F statistic for this hypothesis takes the form

$$\frac{(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})^\top \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2 (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})/k_2}{\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}/(n-k)}, \tag{6.18}$$

where $k = k_1 + k_2$; see Exercise 6.6.

It is easy to see that, when multiplied by k_2, the F statistic (6.18) is in the form of the Wald statistic (6.15). For the purposes of inference on $\boldsymbol{\beta}_2$, regression (6.16) is equivalent to the FWL regression

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals},$$

from which it follows that $\mathrm{Var}(\hat{\boldsymbol{\beta}}_2) = \sigma^2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}$. The denominator of expression (6.18) is just s^2, the OLS estimate of σ^2 from regression (6.16). Thus we see that k_2 times the F statistic (6.18) can be written in the form of the Wald statistic (6.15), with

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2) = s^2\big(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\big)^{-1}$$

yielding a consistent estimator of the variance of $\hat{\boldsymbol{\beta}}_2$; compare equation (4.64).

Under the assumptions of the classical normal linear model, the F statistic (6.18) follows the $F(k_2, n-k)$ distribution when the null hypothesis is true. Therefore, we can use it to construct an **exact confidence region**. If c_α denotes the $1-\alpha$ quantile of the $F(k_2, n-k)$ distribution, then the $1-\alpha$ confidence region is the set of all $\boldsymbol{\beta}_{20}$ for which

$$(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})^\top \boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2 (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20}) \leq c_\alpha\, k_2\, s^2. \tag{6.19}$$

Because the left-hand side of this inequality is quadratic in $\boldsymbol{\beta}_{20}$ and the matrix $\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2$ is positive definite, the confidence region is the interior of an ellipse for $k_2 = 2$ and the interior of a k_2–dimensional ellipsoid for $k_2 > 2$.

Confidence Ellipses and Confidence Intervals

Figure 6.3 illustrates what a **confidence ellipse** can look like when there are just two components in the vector $\boldsymbol{\beta}_2$, which we denote by β_1 and β_2, and the parameter estimates are negatively correlated. The ellipse, which defines a .95 confidence region, is centered at the parameter estimates $(\hat{\beta}_1, \hat{\beta}_2)$, with its

Copyright © 2021, Russell Davidson and James G. MacKinnon

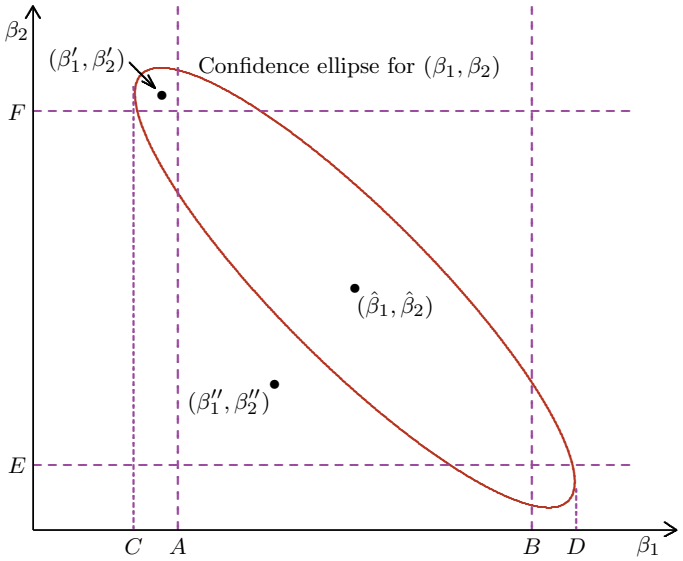


Figure 6.3 Confidence ellipses and confidence intervals

major axis oriented from upper left to lower right. Confidence intervals for β_1 and β_2 are also shown. The .95 confidence interval for β_1 is the line segment AB, and the .95 confidence interval for β_2 is the line segment EF. We would make quite different inferences if we considered AB and EF, and the rectangle they define, demarcated in Figure 6.3 by the lines drawn with long dashes, rather than the confidence ellipse. There are many points, such as (β_1'', β_2''), that lie outside the confidence ellipse but inside the two confidence intervals. At the same time, there are some points, like (β_1', β_2'), that are contained in the ellipse but lie outside one or both of the confidence intervals.

In the framework of the classical normal linear model, the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are bivariate normal. The t statistics used to test hypotheses about just one of β_1 or β_2 are based on the *marginal* univariate normal distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively, but the F statistics used to test hypotheses about both parameters at once are based on the *joint* bivariate normal distribution of the two estimators. If $\hat{\beta}_1$ and $\hat{\beta}_2$ are not independent, as is the case in Figure 6.3, then information about one of the parameters also provides information about the other. Only the confidence region, based on the joint distribution, allows this to be taken into account.

An example may be helpful at this point. Suppose that we are trying to model daily electricity demand during the summer months in an area where air conditioning is prevalent. Since the use of air conditioners, and hence electricity demand, is related to both temperature and humidity, we might

Copyright © 2021, Russell Davidson and James G. MacKinnon

want to use measures of both of them as explanatory variables. In many parts of the world, summer temperatures and humidity are strongly positively correlated. Therefore, if we include both variables in a regression, they may be quite collinear. If so, as we saw in Section 4.5, the OLS estimates must be relatively imprecise. This lack of precision implies that confidence intervals for the coefficients of both temperature and humidity are relatively long, and that confidence regions for both parameters jointly are long and narrow. However, it does not necessarily imply that the area of a confidence region is particularly large. This is precisely the situation that is illustrated in Figure 6.3. Think of β_1 as the coefficient on temperature and β_2 as the coefficient on humidity.

In Exercise 6.7, readers are asked to show that, when there are two explanatory variables in a linear regression model, the correlation between the OLS estimates of the parameters associated with these variables is the negative of the correlation between the variables themselves. Thus, in the example we have been discussing, a positive correlation between temperature and humidity leads to a negative correlation between the estimates of the temperature and humidity parameters, as shown in Figure 6.3. A point like (β_1'', β_2'') is excluded from the confidence region because the variation in electricity demand cannot be accounted for if *both* coefficients are small. But β_1'' cannot be excluded from the confidence interval for β_1 alone, because β_1'', which assigns a small effect to the temperature, is perfectly compatible with the data if a large effect is assigned to the humidity, that is, if β_2 is substantially greater than β_2''. At the same time, even though β_1' is outside the confidence interval for β_1, the point (β_1', β_2') is inside the confidence region, because the very high value of β_2' is enough to compensate for the very low value of β_1'.

The relation between a confidence region for two parameters and confidence intervals for each of the parameters individually is a subtle one. It is tempting to think that the ends of the intervals should be given by the extreme points of the confidence ellipse. This would imply, for example, that the confidence interval for β_1 in the figure is given by the line segment CD. Even without the insight afforded by the temperature-humidity example, however, we can see that this must be incorrect.

The inequality (6.19) defines the confidence region, for given parameter estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, as a set of values in the space of the vector $\boldsymbol{\beta}_{20}$. If instead we think of (6.19) as defining a region in the space of $\hat{\boldsymbol{\beta}}_2$ with $\boldsymbol{\beta}_{20}$ the true parameter vector, then we obtain a region of exactly the same size and shape as the confidence region, because (6.19) is symmetric in $\boldsymbol{\beta}_{20}$ and $\hat{\boldsymbol{\beta}}_2$. We can assign a probability of $1 - \alpha$ to the event that $\hat{\boldsymbol{\beta}}_2$ belongs to the new region, because the inequality (6.19) states that the F statistic is less than its $1 - \alpha$ quantile, an event of which the probability is $1 - \alpha$, by definition.

An exactly similar argument can be made for the confidence interval for β_1. In the two-dimensional framework of Figure 6.3, the entire infinitely high rectangle bounded by the vertical lines through the points A and B has the same size and shape as an area with probability $1 - \alpha$, since we are willing

Copyright © 2021, Russell Davidson and James G. MacKinnon

to allow β_2 to take on any real value. Because the infinite rectangle and the confidence ellipse must contain the *same* probability mass, neither can contain the other. Therefore, the ellipse must protrude outside the region defined by the one-dimensional confidence interval.

It can be seen from the inequality (6.19) that the orientation of a confidence ellipse and the relative lengths of its axes are determined by $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2)$. When the two parameter estimates are positively correlated, the ellipse is oriented from lower left to upper right. When they are negatively correlated, it is oriented from upper left to lower right, as in Figure 6.3. When the correlation is zero, the axes of the ellipse are parallel to the coordinate axes. The variances of the two parameter estimates determine the height and width of the ellipse. If the variances are equal and the correlation is zero, then the confidence ellipse is a circle.

Asymptotic Confidence Regions

When test statistics like (6.18), with known finite-sample distributions, are not available, the easiest way to construct an approximate confidence region is to base it on a Wald statistic like (6.15), which can be used with any k_2–vector of parameter estimates $\hat{\boldsymbol{\theta}}_2$ that is root-n consistent and asymptotically normal and has a covariance matrix that can be consistently estimated. If c_α denotes the $1-\alpha$ quantile of the $\chi^2(k_2)$ distribution and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_2)$ denotes the estimated covariance matrix, then an approximate $1-\alpha$ confidence region is the set of all $\boldsymbol{\theta}_{20}$ such that

$$(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20})^\top \big(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_2)\big)^{-1} (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20}) \leq c_\alpha. \tag{6.20}$$

Like the exact confidence region defined by (6.19), this **asymptotic confidence region** is elliptical or ellipsoidal.

6.4 Heteroskedasticity-Robust Inference

All the testing procedures used in this chapter and the preceding one make use of estimated covariance matrices or standard errors derived from them. If we are to make reliable inferences about the values of parameters, these estimates need to be reliable. In our discussion of how to estimate the covariance matrix of the OLS parameter vector $\hat{\boldsymbol{\beta}}$ in Sections 4.4 and 4.7, we made the rather strong assumption that the disturbances of the regression model are IID. This assumption is needed to show that $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, the usual estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$, is consistent in the sense of equation (5.57). However, even without the IID assumption, it is possible to obtain a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$.

In this section, we relax the IID assumption by allowing the disturbances to be independent but not identically distributed. We focus on the linear regression

Copyright © 2021, Russell Davidson and James G. MacKinnon

model with exogenous regressors,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{6.21}$$

where $\boldsymbol{\Omega}$, the disturbance covariance matrix, is an $n \times n$ matrix with t^{th} diagonal element equal to ω_t^2 and all the off-diagonal elements equal to 0. Since $\boldsymbol{X}$ is assumed to be exogenous, the expectations in (6.21) can be treated as conditional on $\boldsymbol{X}$. Conditional on $\boldsymbol{X}$, then, the disturbances in (6.21) are uncorrelated and have mean 0, but they do not have the same variance for all observations. These disturbances are said to be **heteroskedastic**, or to exhibit **heteroskedasticity**, a subject of which we spoke briefly in Section 2.3. If, instead, all the disturbances do have the same variance, then, as one might expect, they are said to be **homoskedastic**, or to exhibit **homoskedasticity**. Here we assume that the investigator knows nothing about the ω_t^2. In other words, the form of the heteroskedasticity is completely unknown.

The assumption in (6.21) that $\boldsymbol{X}$ is exogenous is fairly strong, but it is often reasonable for cross-section data, as we discussed in Section 4.2. We make it largely for simplicity, since we would obtain essentially the same asymptotic results if we replaced it with the weaker assumption that $\mathrm{E}(u_t \,|\, \boldsymbol{X}_t) = 0$, so that $\boldsymbol{X}$ is predetermined. When the data are generated by a DGP that belongs to (6.21) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the exogeneity assumption implies that $\hat{\boldsymbol{\beta}}$ is unbiased; recall equation (4.12), which in no way depends on assumptions about the covariance matrix of the disturbances.

Whatever the form of the disturbance covariance matrix $\boldsymbol{\Omega}$, the covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}}$ is equal to

$$\begin{aligned} \mathrm{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big) &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \end{aligned} \tag{6.22}$$

This form of covariance matrix is often called a **sandwich covariance matrix**, for the obvious reason that the matrix $\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$ is sandwiched between the two instances of the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. The covariance matrix of an inefficient estimator very often takes this sandwich form.

It is easy to see intuitively why the OLS estimator is inefficient whenever there is heteroskedasticity. Observations with low variance presumably convey more information about the parameters than observations with high variance, and so the former should be given greater weight in an efficient estimator. Instead, OLS gives every observation the same weight.

If we knew the ω_t^2, we could easily evaluate the sandwich covariance matrix (6.22). In fact, as we will see in Chapter 9, we could do even better and actually obtain efficient estimates of $\boldsymbol{\beta}$. But it is assumed that we do not know the ω_t^2. Moreover, since there are n of them, one for each observation, we cannot hope to estimate the ω_t^2 consistently without making additional assumptions. Thus, at first glance, the situation appears hopeless. However,

Copyright © 2021, Russell Davidson and James G. MacKinnon

even though we cannot evaluate the matrix (6.22), we can *estimate* it without having to attempt the impossible task of estimating $\boldsymbol{\Omega}$ consistently.

For the purposes of asymptotic theory, we wish to consider the covariance matrix, not of $\hat{\boldsymbol{\beta}}$, but rather of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. This is just the limit of n times the matrix (6.22). By distributing factors of n in such a way that we can take limits of each of the factors in (6.22), we find that the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is

$$\Big(\lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X}\Big)^{-1} \Big(\lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}\Big) \Big(\lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X}\Big)^{-1}. \tag{6.23}$$

Under assumption (5.47), the factor $(\lim n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, which appears twice in (6.23) as the bread in the sandwich,[1] tends to a finite, deterministic, positive definite matrix $\boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}}$. To estimate the limit, we can simply use the matrix $(n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ itself.

What is not so trivial is to estimate the middle factor, $\lim(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})$, which is the filling in the sandwich. In a very famous paper, White (1980) showed that, under certain conditions, including the existence of the limit, this matrix can be estimated consistently by

$$\frac{1}{n} \boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X}, \tag{6.24}$$

where $\hat{\boldsymbol{\Omega}}$ is an *inconsistent* estimator of $\boldsymbol{\Omega}$. As we will see, there are several alternative versions of $\hat{\boldsymbol{\Omega}}$. The simplest version, and the one suggested in White (1980), is

$$\hat{\boldsymbol{\Omega}} = \operatorname{diag}(\hat{u}_t^2),$$

that is, a diagonal matrix with t^{th} diagonal element equal to $\hat{u}_t^2$, the t^{th} squared OLS residual.

The matrix $\lim(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})$, which is the middle factor of (6.23), is a $k \times k$ symmetric matrix. Therefore, it has exactly $\frac{1}{2}(k^2+k)$ distinct elements. Since this number is independent of the sample size, the matrix can be estimated consistently. Its ij^{th} element is

$$\lim_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^{n} \omega_t^2 x_{ti} x_{tj}\right). \tag{6.25}$$

This is to be estimated by the ij^{th} element of (6.24), which, for the simplest version of $\hat{\boldsymbol{\Omega}}$, is

$$\frac{1}{n} \sum_{t=1}^{n} \hat{u}_t^2 x_{ti} x_{tj}. \tag{6.26}$$

[1] It is a moot point whether to call this factor an ordinary limit, as we do here, or a probability limit, as we do in Section 5.5. The difference reflects the fact that, there, $\boldsymbol{X}$ is generated by some sort of DGP, usually stochastic, while here, we do everything conditional on $\boldsymbol{X}$. We would, of course, need probability limits if $\boldsymbol{X}$ were merely predetermined rather than exogenous.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Because $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$, $\hat{u}_t$ is consistent for u_t, and $\hat{u}_t^2$ is therefore consistent for u_t^2. Thus, asymptotically, expression (6.26) is equal to

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^{n} u_t^2 x_{ti} x_{tj} &= \frac{1}{n} \sum_{t=1}^{n} (\omega_t^2 + v_t) x_{ti} x_{tj} \\ &= \frac{1}{n} \sum_{t=1}^{n} \omega_t^2 x_{ti} x_{tj} + \frac{1}{n} \sum_{t=1}^{n} v_t x_{ti} x_{tj}, \end{aligned} \tag{6.27}$$

where v_t is defined to equal u_t^2 minus its mean of ω_t^2. Under suitable assumptions about the x_{ti} and the ω_t^2, we can apply a law of large numbers to the second term in the second line of (6.27); see White (1980, 2000) for details. Since v_t has mean 0 by construction, this term converges to 0, while the first term converges to expression (6.25).

The above argument shows that the left-hand side of equation (6.27) tends in probability to the limit (6.25). Because the former is asymptotically equivalent to expression (6.26), that expression also tends in probability to (6.25). Consequently, we can use the matrix (6.24), of which a typical element is (6.26), to estimate $\lim(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})$ consistently, and the matrix

$$(n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1} n^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X} (n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1} \tag{6.28}$$

to estimate expression (6.23) consistently. Of course, in practice, we ignore the factors of n^{-1} and use the matrix

$$\widehat{\operatorname{Var}}_{\text{h}}(\hat{\boldsymbol{\beta}}) \equiv (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \tag{6.29}$$

directly to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$. Expression (6.29) depends on $\hat{\boldsymbol{\Omega}}$ only through $\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}$, which is a symmetric $k \times k$ matrix. Notice that we can compute the latter directly by calculating $k(k+1)/2$ quantities like (6.26) without the factor of n^{-1}.

It is not very difficult to modify the arguments of Section 5.5 so that they apply to the model (6.21). Equations (5.55) and (5.56) of Theorem 5.3 would then be replaced by

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \text{N}\Big(\mathbf{0}, \boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}} \Big(\lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}\Big) \boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}}\Big) \tag{6.30}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \widehat{\operatorname{Var}}_{\text{h}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}} \Big(\lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}\Big) \boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}}. \tag{6.31}$$

We conclude that the OLS estimator $\hat{\boldsymbol{\beta}}$ for the model (6.21) is root-n consistent and asymptotically normal, with (6.29) providing a consistent estimator of its covariance matrix.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The sandwich estimator (6.29) that we have just derived is an example of a **heteroskedasticity-consistent covariance matrix estimator**, or **HCCME** for short. It was introduced to econometrics by White (1980), although there were some precursors in the statistics literature, notably Eicker (1963, 1967), Huber (1967), and Hinkley (1977). By taking square roots of the diagonal elements of (6.29), we can obtain standard errors that are asymptotically valid in the presence of heteroskedasticity of unknown form. These **heteroskedasticity-robust standard errors** are often enormously useful.

Alternative Forms of HCCME

The original HCCME (6.29) of White (1980), which is often called HC_0, simply uses squared residuals to estimate the diagonal elements of the matrix $\boldsymbol{\Omega}$. However, it is not a very good covariance matrix estimator. The reason is that, as we saw in Section 4.7, least-squares residuals tend to be too small. There are several better estimators that inflate the squared residuals slightly so as to offset this tendency. Three well-known methods for obtaining better estimates of the ω_t^2 are:

HC_1: Use $\hat{u}_t^2$ in $\hat{\boldsymbol{\Omega}}$ and then multiply the entire matrix (6.29) by the scalar $n/(n-k)$, thus incorporating a standard degrees-of-freedom correction.

HC_2: Use $\hat{u}_t^2/(1-h_t)$ in $\hat{\boldsymbol{\Omega}}$, where $h_t \equiv \boldsymbol{X}_t(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_t^\top$ is the t^{th} diagonal element of the "hat" matrix $\boldsymbol{P}_{\boldsymbol{X}}$ that projects orthogonally on to the space spanned by the columns of $\boldsymbol{X}$. Recall the result (4.58) that, when the variance of all the u_t is σ^2, the expectation of $\hat{u}_t^2$ is $\sigma^2(1-h_t)$. Therefore, the ratio of $\hat{u}_t^2$ to $1-h_t$ would have expectation σ^2 if the disturbances were homoskedastic.

HC_3: Use $\hat{u}_t^2/(1-h_t)^2$ in $\hat{\boldsymbol{\Omega}}$. This is a slightly simplified version of what one gets by employing a statistical technique called the **jackknife**. Dividing by $(1-h_t)^2$ may seem to be overcorrecting the residuals. However, when the disturbances are heteroskedastic, observations with large variances tend to influence the estimates a lot, and they therefore tend to have residuals that are very much too small. Thus, this estimator may be attractive if large variances are associated with large values of h_t.

With minor modifications, the argument used in the preceding subsection for HC_0 shows that all of these procedures give the correct answer asymptotically, but none of them can be expected to do so in finite samples. In fact, inferences based on any HCCME, especially HC_0 and HC_1, may be seriously inaccurate even when the sample size is moderately large if some observations have much higher leverage than others.

The HC_1 and HC_2 covariance matrices, and the original jackknife version of HC_3, were discussed in MacKinnon and White (1985), which found limited evidence that the jackknife seemed to work best. Later simulations in Long and Ervin (2000) also support the use of HC_3. However, theoretical work in Chesher (1989) and Chesher and Austin (1991) gives more ambiguous results and suggests that HC_2 may sometimes outperform HC_3. Simulations in MacKinnon (2012), which also reviews a number of other procedures for heteroskedasticity-robust inference, confirm this prediction. It appears that the best procedure to use depends on the $\boldsymbol{X}$ matrix and on the form of the heteroskedasticity.

Copyright © 2021, Russell Davidson and James G. MacKinnon

In Chapter 7, we will introduce bootstrap methods that are suitable for use with the model (6.21). By combining these with appropriate covariance matrix estimators, we can often obtain substantially more accurate inferences than simply using an HCCME, especially for samples that are small or involve a few observations with high leverage; see MacKinnon (2012).

When Does Heteroskedasticity Matter?

Even when the disturbances are heteroskedastic, there are cases in which we do not necessarily have to use an HCCME. Consider the ij^{th} element of $n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$, which is

$$\frac{1}{n}\sum_{t=1}^{n}\omega_t^2 x_{ti}x_{tj}. \tag{6.32}$$

If the limit as $n\to\infty$ of the average of the ω_t^2, $t=1,\dots,n$, exists and is denoted σ^2, then expression (6.32) can be rewritten as

$$\sigma^2\frac{1}{n}\sum_{t=1}^{n}x_{ti}x_{tj} + \frac{1}{n}\sum_{t=1}^{n}(\omega_t^2-\sigma^2)x_{ti}x_{tj}.$$

The first term here is just the ij^{th} element of $\sigma^2 n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$. Should it be the case that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}(\omega_t^2-\sigma^2)x_{ti}x_{tj} = 0 \tag{6.33}$$

for $i,j=1,\dots,k$, then we find that

$$\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X} = \sigma^2\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}. \tag{6.34}$$

In this special case, we can replace the middle term of (6.23) by the right-hand side of (6.34), and we find that the asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ is just

$$\Big(\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1}\sigma^2\Big(\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\Big)\Big(\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1} = \sigma^2\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1}.$$

The usual OLS estimate of σ^2 is $s^2 = \big(1/(n-k)\big)\sum_{t=1}^{n}\hat{u}_t^2$ and, if we assume that we can apply a law of large numbers, the probability limit of this is

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\omega_t^2 = \sigma^2, \tag{6.35}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

by definition. Thus we see that, in this special case, the usual OLS covariance matrix estimator (4.64) is valid asymptotically. This important result was originally shown in White (1980).

Equation (6.33) always holds when we are estimating the expectation of the dependent variable by a sample mean. In that case, $\boldsymbol{X} = \boldsymbol{\iota}$, a vector with typical element $\iota_t = 1$, and

$$\frac{1}{n}\sum_{t=1}^{n}\omega_t^2 x_{ti}x_{tj} = \frac{1}{n}\sum_{t=1}^{n}\omega_t^2 \iota_t^2 = \frac{1}{n}\sum_{t=1}^{n}\omega_t^2 \to \sigma^2 \text{ as } n \to \infty.$$

This shows that we do not have to worry about heteroskedasticity when calculating the standard error of a sample mean. Of course, equation (6.33) also holds when the disturbances are homoskedastic. In that case, the σ^2 given by (6.35) is just the variance of each of the disturbances.

Although equation (6.33) holds only in certain special cases, it does make one thing clear. Any form of heteroskedasticity affects the efficiency of the OLS parameter estimator, but only heteroskedasticity that is related to the squares and cross-products of the x_{ti} affects the validity of the usual OLS covariance matrix estimator.

6.5 HAC Covariance Matrix Estimators

The assumption that the matrix $\boldsymbol{\Omega}$ is diagonal is what makes it possible to estimate the matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$ consistently and obtain an HCCME, even though $\boldsymbol{\Omega}$ itself cannot be estimated consistently. However, valid covariance matrix estimators can sometimes be obtained under weaker assumptions about $\boldsymbol{\Omega}$. In this and the next section, we investigate some possibilities.

The matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$ can sometimes be estimated consistently when the disturbances of a regression model using time-series data are correlated across time periods. As we mentioned in Section 2.3, such disturbances are said to display serial correlation or **autocorrelation**. Serial correlation is frequently encountered in models estimated using time-series data. Often, observations that are close to each other are strongly correlated, but observations that are far apart are uncorrelated or nearly so. In this situation, only the elements of $\boldsymbol{\Omega}$ that are on or close to the principal diagonal are large. When this is the case, we may be able to obtain an estimate of the covariance matrix of the parameter estimates that is **heteroskedasticity and autocorrelation consistent**, or **HAC**. Computing a HAC covariance matrix estimator is essentially similar to computing an HCCME, but it is somewhat more complicated[2].

[2] For no good reason, terminology has developed in such a way that HCCME functions as a noun, while HAC functions as an adjective.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The asymptotic covariance matrix of the vector $n^{-1/2}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ of (scaled) estimating functions, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, is defined as follows:

$$\boldsymbol{\Sigma} \equiv \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0)^\top\boldsymbol{X} = \operatorname*{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}. \tag{6.36}$$

A HAC estimator of $\boldsymbol{\Sigma}$ is a matrix $\hat{\boldsymbol{\Sigma}}$ constructed so that $\hat{\boldsymbol{\Sigma}}$ consistently estimates $\boldsymbol{\Sigma}$ when the disturbances u_t display any pattern of heteroskedasticity and/or autocorrelation that satisfies certain, generally quite weak, conditions. In order to derive such an estimator, we begin by rewriting the definition of $\boldsymbol{\Sigma}$ in an alternative way:

$$\boldsymbol{\Sigma} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{E}\big(u_t u_s \boldsymbol{X}_t^\top\boldsymbol{X}_s\big), \tag{6.37}$$

in which we assume that a law of large numbers can be used to justify replacing the probability limit in (6.36) by the expectations in (6.37).

For regression models with heteroskedasticity but no autocorrelation, only the terms with $t = s$ contribute to (6.37). Therefore, for such models, we can estimate $\boldsymbol{\Sigma}$ consistently by simply ignoring the expectation operator and replacing the disturbances u_t by least-squares residuals $\hat{u}_t$, possibly with a modification designed to offset the tendency for such residuals to be too small. The obvious way to estimate (6.37) when there may be serial correlation is again simply to drop the expectations operator and replace $u_t u_s$ by $\hat{u}_t\hat{u}_s$, where $\hat{u}_t$ denotes the t^{th} OLS residual. Unfortunately, this approach does not work. To see why not, we need to rewrite (6.37) in yet another way. Let us define the **autocovariance matrices** of the $\boldsymbol{X}_t^\top u_t$ as follows:

$$\boldsymbol{\Gamma}(j) \equiv \begin{cases} \dfrac{1}{n}\displaystyle\sum_{t=j+1}^{n}\mathrm{E}(u_t u_{t-j}\boldsymbol{X}_t^\top\boldsymbol{X}_{t-j}) & \text{for } j \geq 0, \\ \dfrac{1}{n}\displaystyle\sum_{t=-j+1}^{n}\mathrm{E}(u_{t+j}u_t\boldsymbol{X}_{t+j}^\top\boldsymbol{X}_t) & \text{for } j < 0. \end{cases} \tag{6.38}$$

Because there are k estimating functions, these are $k\times k$ matrices. It is easy to check that $\boldsymbol{\Gamma}(j) = \boldsymbol{\Gamma}^\top(-j)$. Then, in terms of the matrices $\boldsymbol{\Gamma}(j)$, expression (6.37) becomes

$$\boldsymbol{\Sigma} = \lim_{n\to\infty}\sum_{j=-n+1}^{n-1}\boldsymbol{\Gamma}(j) = \lim_{n\to\infty}\Big(\boldsymbol{\Gamma}(0) + \sum_{j=1}^{n-1}\big(\boldsymbol{\Gamma}(j)+\boldsymbol{\Gamma}^\top(j)\big)\Big). \tag{6.39}$$

Therefore, in order to estimate $\boldsymbol{\Sigma}$, we apparently need to estimate all of the autocovariance matrices for $j = 0, \ldots, n-1$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

If $\hat{u}_t$ denotes a typical OLS residual, the **sample autocovariance matrix** of order j, $\hat{\boldsymbol{\Gamma}}(j)$, is just the appropriate expression in (6.38), without the expectation operator, and with the random variables u_t and u_{t-j} replaced by $\hat{u}_t$ and $\hat{u}_{t-j}$, respectively. For any $j \geq 0$, this is

$$\hat{\boldsymbol{\Gamma}}(j) = \frac{1}{n} \sum_{t=j+1}^{n} \hat{u}_t \hat{u}_{t-j} \boldsymbol{X}_t^\top \boldsymbol{X}_{t-j}. \tag{6.40}$$

Unfortunately, the sample autocovariance matrix $\hat{\boldsymbol{\Gamma}}(j)$ of order j is not a consistent estimator of the true autocovariance matrix for arbitrary j. Suppose, for instance, that $j = n-2$. Then, from (6.40), we see that $\hat{\boldsymbol{\Gamma}}(j)$ has only two terms, and no conceivable law of large numbers can apply to only two terms. In fact, $\hat{\boldsymbol{\Gamma}}(n-2)$ must tend to zero as $n \to \infty$ because of the factor of n^{-1} in its definition.

The solution to this problem is to restrict our attention to models for which the actual autocovariances mimic the behavior of the sample autocovariances, and for which therefore the actual autocovariance of order j tends to zero as $j \to \infty$. A great many stochastic processes generate disturbances for which the $\boldsymbol{\Gamma}(j)$ do have this property. In such cases, we can drop most of the sample autocovariance matrices that appear in the sample analog of (6.39) by eliminating ones for which $|j|$ is greater than some chosen threshold, say p. This yields the following estimator for $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\Sigma}}_{\text{HW}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \big(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}^\top(j)\big), \tag{6.41}$$

We refer to this estimator as the **Hansen-White estimator**, because it was originally proposed by Hansen (1982) and White and Domowitz (1984); see also White (2000).

For the purposes of asymptotic theory, it is necessary to let the parameter p, which is called the **lag truncation parameter**, go to infinity in (6.41) at some suitable rate as the sample size goes to infinity. A typical rate would be $n^{1/4}$. This ensures that, for large enough n, all the nonzero $\boldsymbol{\Gamma}(j)$ are estimated consistently. Unfortunately, this type of result does not say how large p should be in practice. In most cases, we have a given, finite, sample size, and we need to choose a specific value of p.

The Hansen-White estimator (6.41) suffers from one very serious deficiency: In finite samples, it need not be positive definite or even positive semidefinite. If one happens to encounter a data set that yields a nondefinite $\hat{\boldsymbol{\Sigma}}_{\text{HW}}$, then, since a covariance matrix must be positive definite, (6.41) is unusable. Luckily, there are numerous ways out of this difficulty. The one that is most widely used was suggested by Newey and West (1987). The **Newey-West estimator** they propose is

$$\hat{\boldsymbol{\Sigma}}_{\text{NW}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^{p} \left(1 - \frac{j}{p+1}\right) \big(\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}^\top(j)\big), \tag{6.42}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

in which each sample autocovariance matrix $\hat{\boldsymbol{\Gamma}}(j)$ is multiplied by a weight $1 - j/(p+1)$ that decreases linearly as j increases. The weight is $p/(p+1)$ for $j = 1$, and it then decreases by steps of $1/(p+1)$ down to a value of $1/(p+1)$ for $j = p$. This estimator evidently tends to underestimate the autocovariance matrices, especially for larger values of j. Therefore, p should almost certainly be larger for (6.42) than for (6.41). As with the Hansen-White estimator, p must increase as n does, and the appropriate rate is $n^{1/3}$. A procedure for selecting p automatically was proposed by Newey and West (1994), but it is too complicated to discuss here.

Both the Hansen-White and the Newey-West HAC estimators of $\boldsymbol{\Sigma}$ can be written in the form

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X} \tag{6.43}$$

for an appropriate choice of $\hat{\boldsymbol{\Omega}}$. This fact follows from the observation that there exist $n \times n$ matrices $\boldsymbol{U}(j)$ such that the $\hat{\boldsymbol{\Gamma}}(j)$ can be expressed in the form $n^{-1} \boldsymbol{X}^\top \boldsymbol{U}(j) \boldsymbol{X}$, as readers are asked to check in Exercise 6.14.

The Newey-West estimator is by no means the only HAC estimator that is guaranteed to be positive definite. Andrews (1991) provides a detailed treatment of HAC estimation, suggests some alternatives to the Newey-West estimator, and shows that, in some circumstances, they may perform better than it does in finite samples.

6.6 Cluster-Robust Inference

In many areas of applied econometrics, data are collected at the individual level, but each observation is associated with a higher-level entity, such as a city, state, province, or country, a classroom or school, a hospital, or perhaps a time period. We can think of all the observations associated with each of these higher-level entities as forming a **cluster**. In many cases, it seems plausible that the disturbances for a regression model using data of this type may be correlated within the clusters.

One way to deal with clustering is to write the linear regression model as

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_G \end{bmatrix}, \tag{6.44}$$

where the data are divided into G clusters, indexed by g. The g^{th} cluster has n_g observations, and the entire sample has $n = \sum_{g=1}^{G} n_g$ observations. The matrix $\boldsymbol{X}$ and the vectors $\boldsymbol{y}$ and $\boldsymbol{u}$ have n rows, the matrix $\boldsymbol{X}$ has k columns, and the parameter vector $\boldsymbol{\beta}$ has k elements. For this model, it is customary to

Copyright © 2021, Russell Davidson and James G. MacKinnon

assume that the disturbances are uncorrelated across clusters but potentially correlated and heteroskedastic within clusters, so that

$$\mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g^\top) = \boldsymbol{\Omega}_g, \quad g = 1, \ldots, G. \tag{6.45}$$

The $n_g \times n_g$ covariance matrices $\boldsymbol{\Omega}_g$ are assumed to be unknown. Thus the covariance matrix $\boldsymbol{\Omega}$ of the entire vector $\boldsymbol{u}$ is assumed to be block diagonal, with the matrices $\boldsymbol{\Omega}_g$ forming the diagonal blocks.

Ordinary least squares estimation of equation (6.44) yields OLS estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{u}}$. The covariance matrix of $\hat{\boldsymbol{\beta}}$ is, of course, a sandwich:

$$\begin{aligned}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{\Omega}_g \boldsymbol{X}_g \right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.\end{aligned} \tag{6.46}$$

Notice that the matrix in the middle of the sandwich is actually the sum of G matrices, one for each cluster.

Why Clustering Matters

Before we discuss how to estimate the covariance matrix (6.46), it is important to appreciate the fact that this matrix can be very different from the classical covariance matrix $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and the sandwich covariance matrix (6.22) when $\boldsymbol{\Omega}$ is block-diagonal. Just how different it is depends on the regressors, the cluster sizes, and the intra-cluster correlations. When some or all of the clusters contain many observations, the diagonal elements of (6.46) can be very much larger than those of conventional covariance matrices, even when the intra-cluster correlations are very small.

The simplest and most popular way to model intra-cluster correlation is to use the **error components model**

$$u_{gi} = v_g + \varepsilon_{gi}, \quad v_g \sim \mathrm{IID}(0, \sigma_v^2), \quad \varepsilon_{gi} \sim \mathrm{IID}(0, \sigma_\varepsilon^2), \tag{6.47}$$

for $i = 1, \ldots, n_g$, $g = 1, \ldots, G$. Here v_g is a random variable that affects every observation in cluster g and no observation in any other cluster, while ε_{gi} is an **idiosyncratic shock** that affects only the single observation gi. This model implies that

$$\mathrm{Var}(u_{gi}) = \sigma_v^2 + \sigma_\varepsilon^2 \quad \text{and} \quad \mathrm{Cov}(u_{gi}, u_{gj}) = \sigma_v^2,$$

so that

$$\rho \equiv \frac{\mathrm{Cov}(u_{gi}, u_{gj})}{\mathrm{Var}(u_{gi})} = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2} \text{ for all } g \text{ and } i \neq j.$$

Thus all the intra-cluster correlations are the same and equal to ρ.

Copyright © 2021, Russell Davidson and James G. MacKinnon

There has been a good deal of analysis of this special case; see Kloek (1981) and Moulton (1986, 1990). Suppose for simplicity that the model contains only a constant and one regressor, with coefficient β_2, where the value of the regressor is fixed within each cluster. If n_g is the same for every cluster, then it can be shown that

$$\frac{\mathrm{Var}(\hat{\beta}_2)}{\mathrm{Var_c}(\hat{\beta}_2)} = 1 + (n_g - 1)\rho, \tag{6.48}$$

where $\mathrm{Var}(\hat{\beta}_2)$ is the true variance of $\hat{\beta}_2$ based on the matrix (6.46), and $\mathrm{Var_c}(\hat{\beta}_2)$ is the incorrect variance based on the conventional OLS covariance matrix $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$; see Angrist and Pischke (2008, Chapter 8). The square root of the right-hand side of equation (6.48) is sometimes called the **Moulton factor**. A very simple, but not very accurate, way to "correct" conventional standard errors is to multiply them by an estimate of the Moulton factor.

More generally, the ratio of the correct variance to the conventional one looks like (6.48), but with $n_g - 1$ replaced by a function of the cluster sizes and the intra-cluster correlations of the regressors. For a given sample size and set of regressors, the ratio is greatest when all the n_g are the same.

It is clear from (6.48) that the true variance of $\hat{\beta}_2$ can be very much greater than the incorrect, conventional one when n_g is large, even if ρ is quite small. For example, if $\rho = 0.05$, the true variance will be twice the conventional one when $n_g = 21$, four times the conventional one when $n_g = 61$, and 25 times the conventional one when $n_g = 481$. In practice, clusters often have thousands or even tens of thousands of observations, so that conventional standard errors may easily understate the true ones by factors of ten or more.

One obvious way to solve this problem is to include group fixed effects. This will explain the v_g, leaving only the ε_{gi}. However, group fixed effects cannot be included if any of the regressors of interest does not vary within clusters. In that case, the fixed effects will explain all the variation in those regressors, so that we cannot identify the coefficient(s) we are interested in. Unfortunately, this is a very common situation. It occurs whenever certain regressors, such as labor market conditions, tax rates, local prices or wages, or measures of local amenities, are measured at the group level. It also occurs whenever we are interested in the effects of laws or policies that affect entire groups. This was precisely the situation that motivated the analysis of Kloek (1981).

Even when it is possible to include group fixed effects, they very often do not solve the problem. In most applied problems, there is no reason to believe that intra-cluster correlations arise solely from an error components model like (6.47). They probably arise for a variety of reasons, including misspecification of the regression function and features of the way the data are collected, and they are almost certainly far more complicated than a model like (6.47) implies. Strong evidence that fixed effects do not fully account for intra-cluster correlations in large samples of individual data has been provided by Bertrand, Duflo, and Mullanaithan (2004) and MacKinnon and Webb (2016). Thus it

Copyright © 2021, Russell Davidson and James G. MacKinnon

appears that we should estimate the covariance matrix (6.46) whenever data appear to be clustered, whether or not the regression includes fixed effects. A very general analysis of the consequences of correlated disturbances is provided by Andrews (2005).

Cluster-Robust Covariance Matrix Estimation

It is natural to estimate the covariance matrix (6.46) by generalizing the concept of an HCCME. This idea seems to have been proposed first in Liang and Zeger (1986). For historical reasons, such an estimator is generally called a **cluster-robust variance estimator**, or **CRVE**, rather than a CCCME, as might seem more logical.

The simplest and most widely-used CRVE is

$$\mathrm{CV}_1: \quad \frac{G(n-1)}{(G-1)(n-k)}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g^\top\boldsymbol{X}_g\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{6.49}$$

where $\hat{\boldsymbol{u}}_g$ is the vector of OLS residuals for cluster g. Notice that each of the $k \times k$ matrices within the summation has rank one, because it is equal to the column vector $\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g$ times its transpose. This implies that the rank of the CRVE itself cannot exceed G, making it impossible to test more than G restrictions at once. When $n_g = 1$ for all g, so that $G = n$, expression (6.49) reduces to the familiar HC_1 matrix.

The degrees-of-freedom adjustment in (6.49) may seem odd. That is because it is really the product of two adjustments, one based on the sample size and the number of regressors, and one based on the number of clusters. It is customary to base inferences on the t distribution with $G-1$ degrees of freedom, because it turns out what matters for a reasonable asymptotic analysis is the number of clusters, not the number of observations. Intuitively, each of the terms in the summation in (6.49) contributes one degree of freedom, and a t test uses up one degree of freedom, leaving us with $G-1$ of them.[3] When G is small and n is large, critical values based on the $t(G-1)$ distribution can be substantially larger than ones based on the $t(n-k)$ distribution, potentially leading to different inferences.

The fact that the summation in expression (6.49) has only G terms also suggests that, if CV_1 is to estimate the true covariance matrix (6.46) consistently, the asymptotic construction should be such that the number of clusters G tends to infinity with the sample size. That is indeed the case; see Carter, Schnepel, and Steigerwald (2015). In fact, if the sample size goes to infinity while G remains fixed, $\hat{\boldsymbol{\beta}}$ is not even consistent; see Andrews (2005).

[3] For more formal analyses of this issue, see Donald and Lang (2007) and Bester, Conley, and Hansen (2011).

Copyright © 2021, Russell Davidson and James G. MacKinnon

As its name implies, the estimator CV_1 is analogous to HC_1. There are also CRVEs that are analogous to HC_2 and HC_3. Recall that the latter involve transforming the OLS residuals before using them to compute the filling in the sandwich. In the CRVE case, the transformations involve the residual vectors for each of the clusters. First, define the $n_g \times n_g$ matrices

$$\boldsymbol{M}_{gg} \equiv \mathbf{I}_{n_g} - \boldsymbol{X}_g(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_g^\top, \quad g = 1, \ldots, G. \tag{6.50}$$

These are the diagonal blocks of the $\boldsymbol{M}_{\boldsymbol{X}}$ matrix that correspond to each of the G clusters. They are *not* themselves orthogonal projection matrices, and are in fact positive definite provided that the columns of $\boldsymbol{X}_g$ are linearly independent. The CV_2 matrix uses the transformed residuals

$$\dot{\boldsymbol{u}}_g \equiv \boldsymbol{M}_{gg}^{-1/2}\hat{\boldsymbol{u}}_g,$$

where $\boldsymbol{M}_{gg}^{-1/2}$ denotes the inverse of the symmetric square root of the matrix $\boldsymbol{M}_{gg}$, and the CV_3 matrix uses the transformed residuals

$$\ddot{\boldsymbol{u}}_g \equiv \boldsymbol{M}_{gg}^{-1}\hat{\boldsymbol{u}}_g.$$

The CV_2 and CV_3 matrices have essentially the same form as (6.49), with $\dot{\boldsymbol{u}}_g$ and $\ddot{\boldsymbol{u}}_g$ replacing $\hat{\boldsymbol{u}}_g$, except that the scalar factor at the beginning is omitted because transforming the residuals has already scaled them up.[4] These estimators evidently reduce to HC_2 and HC_3 when each cluster has just one element. There is a good deal of evidence that confidence intervals based on CV_2 and CV_3 have better coverage properties than ones based on CV_1, although ones based on CV_3 are sometimes prone to overcover.

Inference based on the CV_1 covariance matrix (6.49) seems to work well if several key conditions are satisfied. First, clustering must be performed at the appropriate level. This is not always easy to achieve; see Cameron and Miller (2015). Second, the number of clusters must be reasonably large. For clusters of roughly equal sizes, 50 or so is probably sufficient, but a much larger number may be needed if cluster sizes vary a lot; see MacKinnon and Webb (2016). Third, the disturbances must be approximately homoskedastic across clusters.

When at least one of the conditions for reliable inference just given is violated, it is dangerous to use t statistics and confidence intervals based on the CV_1 covariance matrix (6.49). Using CV_2 or CV_3, if that is feasible, will probably work at least somewhat better than using CV_1, but often not well enough. Conceptually, the simplest way to make better inferences is probably to use bootstrap methods, which will be discussed in Chapter 7.

The literature on cluster-robust inference has grown rapidly since the turn of the century. For a much more detailed treatment of this literature than we have space for, see Cameron and Miller (2015).

[4] CV_2 was first proposed in Bell and McCaffrey (2002) and has been studied by Imbens and Kolesár (2016). Both it and CV_3 have been investigated by MacKinnon (2015) and Young (2015).

Copyright © 2021, Russell Davidson and James G. MacKinnon

6.7 Difference in Differences

Suppose that a new policy comes into effect in one or more jurisdictions (such as countries, states, provinces, or cities) at one or more points in time. Economists may be interested in seeing what effect, if any, the policy had on some variable of interest. The problem is to disentangle the effects of the policy change from other changes across time or across jurisdictions. One method that is commonly used is a type of linear regression called **difference in differences**, or **DiD** for short.[5]

Let us index jurisdictions by g and time periods by t, so that y_{gti} denotes the dependent variable for the i^{th} unit (for example, an individual, a household, or a firm) within jurisdiction g at time t. If $\mathrm{E}(y_{gti})$ could vary arbitrarily across both jurisdictions and time periods, there would be no way to identify the effects of the policy. Therefore, we assume that, in the absence of the policy,

$$y_{gti} = \eta_g + \lambda_t + u_{gti}, \tag{6.51}$$

where η_g is a jurisdiction fixed effect, λ_t is a time fixed effect, and u_{gti} is an idiosyncratic shock. This assumption is by no means innocuous, since it imposes a common jurisdiction fixed effect η_g on all time periods and a common time fixed effect, or **common trend**, λ_t on all jurisdictions. Suppose further that the only effect of the policy is to shift $\mathrm{E}(y_{gti})$ by a constant δ, so that equation (6.51) would include an additional term for any observation where the policy is active.

Initially, suppose there are only two jurisdictions, denoted a and b, and two time periods, denoted 1 and 2. If the policy is imposed in jurisdiction b in period 2 only, then we have four equations, one for each jurisdiction in each time period:

$$\begin{aligned} y_{a1i} &= \eta_a + \lambda_1 + u_{a1i}, \quad & y_{a2i} &= \eta_a + \lambda_2 + u_{a2i}, \\ y_{b1i} &= \eta_b + \lambda_1 + u_{b1i}, \quad & y_{b2i} &= \eta_b + \lambda_2 + \delta + u_{b2i}. \end{aligned} \tag{6.52}$$

Let $\bar{y}_{gt}$ and $\bar{u}_{gt}$ denote the average values of the y_{gti} and the u_{gti}, respectively, for $g = a, b$ and $t = 1, 2$. Then equations (6.52) and our assumption about the effect of the policy imply that

$$\bar{y}_{a2} - \bar{y}_{a1} = \lambda_2 - \lambda_1 + (\bar{u}_{a2} - \bar{u}_{a1}),$$

and

$$\bar{y}_{b2} - \bar{y}_{b1} = \delta + \lambda_2 - \lambda_1 + (\bar{u}_{b2} - \bar{u}_{b1}).$$

Therefore,

$$(\bar{y}_{b2} - \bar{y}_{b1}) - (\bar{y}_{a2} - \bar{y}_{a1}) = \delta + (\bar{u}_{b2} - \bar{u}_{b1}) - (\bar{u}_{a2} - \bar{u}_{a1}). \tag{6.53}$$

[5] For a more detailed discussion of the DiD methodology, see Angrist and Pischke (2008, Chapter 5).

Copyright © 2021, Russell Davidson and James G. MacKinnon

The quantity on the left of this equation is the difference between two first differences, $\bar{y}_{b2} - \bar{y}_{b1}$ and $\bar{y}_{a2} - \bar{y}_{a1}$. The quantity on the right is the parameter we want to estimate, δ, plus a linear combination of the disturbances. Notice that the parameters λ_1 and λ_2 have vanished. The difference in differences on the left of equation (6.53) is something that we can calculate. If we have a large enough sample, the variance of the disturbance term on the right of equation (6.53) should be small enough that the quantity on the left provides a reasonable estimate of δ.

Instead of actually computing the difference in differences on the left-hand side of equation (6.53), we can simply estimate a regression model. Define D^b_{gti} as a dummy variable that equals 1 if $g = b$ and 0 otherwise, and D^2_{gti} as a dummy variable that equals 1 if $t = 2$ and 0 otherwise. Then equations (6.52) can be combined into just one equation:

$$y_{gti} = \beta_1 + \beta_2 D^b_{gti} + \beta_3 D^2_{gti} + \delta D^b_{gti} D^2_{gti} + u_{gti}. \tag{6.54}$$

The first three coefficients here are related to the coefficients in equations (6.52) as follows:

$$\beta_1 = \eta_a + \lambda_1, \quad \beta_2 = \eta_b - \eta_a, \quad \beta_3 = \lambda_2 - \lambda_1.$$

The coefficient of interest is, of course, δ, which measures the effect of the treatment on jurisdiction b in period 2.

Clustering by Jurisdiction

Although there are studies that use the difference-in-differences methodology with just two jurisdictions and two time periods (Card and Krueger (1994) is a pioneering one), it is impossible to allow for clustered disturbances in that case. If we are to make valid inferences that allow for clustering at the jurisdiction level, it is essential to have at least a moderate number of jurisdictions, in a reasonable fraction of which the policy is imposed. We say that the jurisdictions in which the policy is imposed are **treated**.

In the general case, there are $G \geq 2$ jurisdictions that we wish to consider as clusters, of which G_1 are treated in at least some of the T time periods and G_0 are never treated. Instead of regression (6.54), we can then estimate a regression of the form

$$y_{gti} = \beta_1 + \sum_{j=2}^{G} \beta_j \mathrm{DJ}^j_{gti} + \sum_{k=1}^{T-1} \beta_{G+k} \mathrm{DT}^k_{gti} + \delta \mathrm{TR}_{gti} + u_{gti}, \tag{6.55}$$

where the DJ^j_{gti} are jurisdiction dummies that equal 1 when $g = j$, the DT^k_{gti} are time dummies that equal 1 when $t = k$, and TR_{gti} is a treatment dummy that equals 1 when jurisdiction g is treated in time period t. Thus the treatment dummy TR_{gti} equals 1 for the treated observations in the G_1 treated

Copyright © 2021, Russell Davidson and James G. MacKinnon

clusters. It equals 0 for the remaining observations in those clusters and for all observations in the G_0 untreated clusters. Of course, equation (6.55) could also include additional explanatory variables, provided they vary at the individual level.

It would be impossible to estimate equation (6.55) if any jurisdiction were treated in every period, because there would be perfect collinearity between at least one of the jurisdiction dummies and the treatment dummy. Thus if every jurisdiction were either treated in every period or not treated in every period, all of the jurisdiction dummies would have to be dropped. We could still estimate an equation like (6.55) with fewer parameters, but we could not identify the parameter δ separately from the the jurisdiction fixed effects η_g. Thus, even though equation (6.55) is not explicitly written in terms of a difference in differences, the basic idea of DiD is still embodied in it, because we can identify the parameter δ only if some jurisdictions are treated in some periods and not treated in others.

When computing test statistics or confidence intervals based on equations like (6.55), it is obligatory to use a cluster-robust covariance matrix. In principle, one could cluster either by jurisdictions or by jurisdiction-period pairs. In most cases, it seems to be best to cluster at the jurisdiction level; see Bertrand, Duflo, and Mullanaithan (2004) and Cameron and Miller (2015). However, if this leads to the number of treated clusters being small, there is a serious risk of severe errors of inference; see MacKinnon and Webb (2016).

6.8 The Delta Method

Econometricians often want to perform inference on nonlinear functions of model parameters. This requires them to estimate the standard error of a nonlinear function of parameter estimates or, more generally, the covariance matrix of a vector of such functions. One popular way to do so is called the **delta method**. It is based on an asymptotic approximation.

For simplicity, let us start with the case of a single parameter. Suppose that we have estimated a scalar parameter θ, which might be one of the coefficients of a linear regression model, and that we are interested in the parameter $\gamma \equiv g(\theta)$, where $g(\cdot)$ is a monotonic function that is continuously differentiable. In this situation, the obvious way to estimate γ is to use $\hat{\gamma} = g(\hat{\theta})$. Since $\hat{\theta}$ is a random variable, so is $\hat{\gamma}$. The problem is to estimate the variance of $\hat{\gamma}$.

Since $\hat{\gamma}$ is a function of $\hat{\theta}$, it seems logical that $\mathrm{Var}(\hat{\gamma})$ should be a function of $\mathrm{Var}(\hat{\theta})$. If $g(\theta)$ is an affine function, then we already know how to calculate $\mathrm{Var}(\hat{\gamma})$; recall the result (4.44). The idea of the delta method is simply to find a linear approximation to $g(\theta)$ and then apply (4.44) to this approximation.

Copyright © 2021, Russell Davidson and James G. MacKinnon

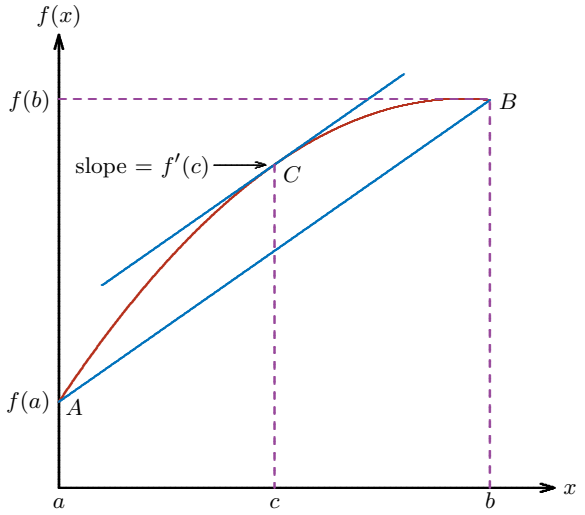


Figure 6.4 Taylor's Theorem

Taylor's Theorem

It is frequently necessary in econometrics to obtain linear approximations to nonlinear functions. The mathematical tool most commonly used for this purpose is **Taylor's Theorem**. In its simplest form, Taylor's Theorem applies to functions of a scalar argument that are differentiable at least once on some real interval $[a, b]$, with the derivative a continuous function on $[a, b]$. Figure 6.4 shows the graph of such a function, $f(x)$, for $x \in [a, b]$.

The coordinates of A are $(a, f(a))$, and those of B are $(b, f(b))$. Thus the slope of the line AB is $\big(f(b) - f(a)\big)/(b - a)$. What drives the theorem is the observation that there must always be a value between a and b, like c in the figure, at which the derivative $f'(c)$ is equal to the slope of AB. This is a consequence of the continuity of the derivative. If it were not continuous, and the graph of $f(x)$ had a corner, the slope might always be greater than $f'(c)$ on one side of the corner, and always be smaller on the other. But if $f'(x)$ is continuous on $[a, b]$, then there must exist c such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

This can be rewritten as $f(b) = f(a) + (b - a)f'(c)$. If we let $h = b - a$, then, since c lies between a and b, it must be the case that $c = a + \lambda h$, for some λ between 0 and 1. Thus we obtain

$$f(a + h) = f(a) + hf'(a + \lambda h). \tag{6.56}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Equation (6.56), which is the simplest expression of Taylor's Theorem, is also known as the **Mean Value Theorem**.

Although equation (6.56) is an exact relationship, it involves the quantity λ, which is unknown. It is more usual just to set $\lambda = 0$, so as to obtain a linear approximation to the function $f(x)$ for x in the neighborhood of a. This approximation, called a **first-order Taylor expansion** around a, is

$$f(a+h) \cong f(a) + hf'(a), \tag{6.57}$$

where the symbol "$\cong$" means "is approximately equal to." The right-hand side of this equation is an affine function of h.

Taylor's Theorem can be extended in order to provide approximations that are quadratic or cubic functions, or polynomials of any desired order. The exact statement of the theorem, with terms proportional to powers of h up to h^p, is

$$f(a+h) = f(a) + \sum_{i=1}^{p-1} \frac{h^i}{i!} f^{(i)}(a) + \frac{h^p}{p!} f^{(p)}(a+\lambda h).$$

Here $f^{(i)}$ is the i^{th} derivative of f, and once more $0 < \lambda < 1$. The approximate version of the theorem sets $\lambda = 0$ and gives rise to a **p^{th}-order Taylor expansion** around a. A commonly-encountered example of the latter is the **second-order Taylor expansion**

$$f(a+h) \cong f(a) + hf'(a) + \frac{1}{2}h^2 f''(a).$$

Both versions of Taylor's Theorem require as a regularity condition that $f(x)$ should have a p^{th} derivative that is continuous on $[a, a+h]$.

There are also multivariate versions of Taylor's Theorem, and we will need them from time to time. If $f(\boldsymbol{x})$ is now a scalar-valued function of the m–vector $\boldsymbol{x}$, then, for $p = 1$, the Mean Value Theorem states that, if $\boldsymbol{h}$ is also an m–vector,

$$f(\boldsymbol{x}+\boldsymbol{h}) \cong f(\boldsymbol{x}) + \sum_{j=1}^{m} h_j f_j(\boldsymbol{x}+\lambda\boldsymbol{h}), \tag{6.58}$$

where h_j is the j^{th} component of $\boldsymbol{h}$, f_j is the partial derivative of f with respect to its j^{th} argument, and, as before, $0 < \lambda < 1$.

The Delta Method for a Scalar Parameter

If we assume that the estimator $\hat{\theta}$ is root-n consistent and asymptotically normal, then

$$n^{1/2}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} \mathrm{N}\big(0, V^\infty(\hat{\theta})\big), \tag{6.59}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where θ_0 denotes the true value of θ. Here we use $V^\infty(\hat{\theta})$ as a shorthand way of writing the asymptotic variance of $n^{1/2}(\hat{\theta} - \theta_0)$.

In order to find the asymptotic distribution of $\hat{\gamma} = g(\hat{\theta})$, we perform a first-order Taylor expansion of $g(\hat{\theta})$ around θ_0. Using (6.57), we obtain

$$\hat{\gamma} \cong g(\theta_0) + g'(\theta_0)(\hat{\theta} - \theta_0), \tag{6.60}$$

where $g'(\theta_0)$ is the first derivative of $g(\theta)$, evaluated at θ_0. Given the root-n consistency of $\hat{\theta}$, (6.60) can be rearranged into an **asymptotic equality**. Two deterministic quantities are said to be **asymptotically equal** if they tend to the same limits as $n \to \infty$. Similarly, two random quantities are said to be asymptotically equal if the difference between them tends to zero in probability. As usual, we need a power of n to make things work correctly. Here, we multiply both sides of (6.60) by $n^{1/2}$. If we denote $g(\theta_0)$, which is the true value of γ, by γ_0, then (6.60) becomes

$$n^{1/2}(\hat{\gamma} - \gamma_0) \stackrel{a}{=} g_0' n^{1/2}(\hat{\theta} - \theta_0), \tag{6.61}$$

where the symbol $\stackrel{a}{=}$ is used for asymptotic equality, and $g_0' \equiv g'(\theta_0)$. In Exercise 6.19, readers are asked to check that, if we perform a second-order Taylor expansion instead of a first-order one, the last term of the expansion vanishes asymptotically. This justifies (6.61) as an asymptotic equality.

Equation (6.61) shows that $n^{1/2}(\hat{\gamma} - \gamma_0)$ is asymptotically normal with mean 0, since the right-hand side of (6.61) is just g_0' times a quantity that is asymptotically normal with mean 0; recall (6.59). The variance of $n^{1/2}(\hat{\gamma} - \gamma_0)$ is clearly $(g_0')^2 V^\infty(\hat{\theta})$, and so we conclude that

$$n^{1/2}(\hat{\gamma} - \gamma_0) \stackrel{a}{\sim} \mathrm{N}\big(0, (g_0')^2 V^\infty(\hat{\theta})\big). \tag{6.62}$$

This shows that $\hat{\gamma}$ is root-n consistent and asymptotically normal when $\hat{\theta}$ is.

The result (6.62) leads immediately to a practical procedure for estimating the standard error of $\hat{\gamma}$. If the standard error of $\hat{\theta}$ is s_θ, then the standard error of $\hat{\gamma}$ is

$$s_\gamma \equiv \big|g'(\hat{\theta})\big|\, s_\theta. \tag{6.63}$$

This procedure can be based on any asymptotically valid estimator of the standard deviation of $\hat{\theta}$. For example, if θ were one of the coefficients of a linear regression model, then s_θ could be the square root of the corresponding diagonal element of the usual estimated OLS covariance matrix, or it could be the square root of the corresponding diagonal element of an estimated HCCME, CRVE, or HAC estimator.

In practice, the delta method is usually very easy to use. For example, consider the case in which $\gamma = \theta^2$. Then $g'(\theta) = 2\theta$, and the formula (6.63) tells us that $s_\gamma = 2|\hat{\theta}|s_\theta$. Notice that s_γ depends on $\hat{\theta}$, something that is not true for any of the standard errors we have discussed previously.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Confidence Intervals and the Delta Method

Although the result (6.63) is simple and practical, it reveals some of the limitations of asymptotic theory. Whenever the relationship between $\hat{\theta}$ and $\hat{\gamma}$ is nonlinear, it is impossible that both estimators should be normally distributed in finite samples. Suppose that $\hat{\theta}$ really did happen to be normally distributed. Then, unless $g(\cdot)$ were linear, $\hat{\gamma}$ could not possibly be normally, or even symmetrically, distributed. Similarly, if $\hat{\gamma}$ were normally distributed, $\hat{\theta}$ could not be. Moreover, as the example at the end of the last subsection showed, s_γ generally depends on $\hat{\theta}$. This implies that the numerator of a t statistic for γ is not independent of the denominator. However, independence was essential to the result, in Section 5.4, that the t statistic actually follows the Student's t distribution.

The preceding arguments suggest that confidence intervals and test statistics based on asymptotic theory may often not be reliable in finite samples. Asymptotic normality of the parameter estimates is an essential underpinning of all asymptotic tests and confidence intervals or regions. When the finite-sample distributions of estimates are far from the limiting normal distribution, one cannot expect any asymptotic procedure to perform well.

Despite these caveats, we may still wish to construct an asymptotic confidence interval for γ based on the second interval in (6.08). The result is

$$\big[\hat{\gamma} - s_\gamma z_{1-\alpha/2},\ \ \hat{\gamma} + s_\gamma z_{1-\alpha/2}\big], \tag{6.64}$$

where s_γ is the delta method estimate (6.63), and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. This confidence interval can be expected to work well whenever the finite-sample distribution of $\hat{\gamma}$ is well approximated by the normal distribution and s_γ is a reliable estimator of its standard deviation.

Using (6.64) is not the only way to obtain an asymptotic confidence interval for γ, however. Another approach, which usually leads to an asymmetric interval, is to transform the asymptotic confidence interval for the underlying parameter θ. The latter interval, which is similar to the second interval in (6.08), is

$$\big[\hat{\theta} - s_\theta z_{1-\alpha/2},\ \ \hat{\theta} + s_\theta z_{1-\alpha/2}\big].$$

Transforming the endpoints of this interval by the function $g(\cdot)$ gives the following interval for γ:

$$\big[g(\hat{\theta} - s_\theta z_{1-\alpha/2}),\ \ g(\hat{\theta} + s_\theta z_{1-\alpha/2})\big]. \tag{6.65}$$

This formula assumes that $g'(\theta) > 0$. If $g'(\theta) < 0$, the two ends of the interval would have to be interchanged.

Whenever $g(\theta)$ is a nonlinear function, the confidence interval (6.65) must be asymmetric. This is illustrated in Figure 6.5. The lower horizontal line shows

Copyright © 2021, Russell Davidson and James G. MacKinnon

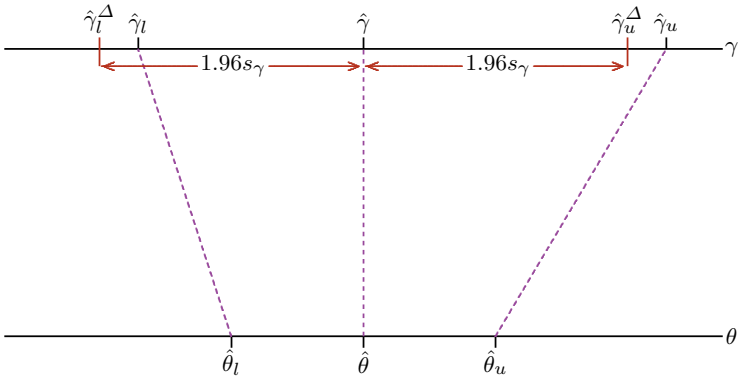


Figure 6.5 Two confidence intervals for $\gamma \equiv \theta^2$

$\hat{\theta}$ and a symmetric confidence interval for θ based on the standard error s_θ. The upper horizontal line shows $\hat{\gamma} = \hat{\theta}^2$ and two different confidence intervals. The one with limits $\hat{\gamma}_l$ and $\hat{\gamma}_u$ is obtained by transforming the ends of the interval for θ using (6.65), as the dashed lines in the figure show. The one with limits $\hat{\gamma}_l^\Delta$ and $\hat{\gamma}_u^\Delta$ is based on the delta method using the result that, in this case, $s_\gamma = 2\hat{\theta}s_\theta$. The interval based on (6.65) can be expected to work better than the delta-method interval if the finite-sample distribution of $\hat{\theta}$ is well approximated by the normal distribution and s_θ is a reliable estimator of the standard deviation of $\hat{\theta}$.

The Vector Case

The result (6.62) can easily be extended to the case in which both $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are vectors. Suppose that the former is a k–vector and the latter is an l–vector, with $l \leq k$. The relation between $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is $\boldsymbol{\gamma} \equiv \boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}(\boldsymbol{\theta})$ is an l–vector of monotonic functions that are continuously differentiable. The vector version of (6.59) is

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} \mathrm{N}\big(\mathbf{0}, \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}})\big), \tag{6.66}$$

where $\boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}})$ is the asymptotic covariance matrix of the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Using the result (6.66) and a first-order Taylor expansion of $\boldsymbol{g}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$, it can be shown that the vector analog of (6.62) is

$$n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \overset{a}{\sim} \mathrm{N}\big(\mathbf{0}, \boldsymbol{G}_0 \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}}) \boldsymbol{G}_0^\top\big), \tag{6.67}$$

where $\boldsymbol{G}_0$ is an $l \times k$ matrix with typical element $\partial g_i(\theta)/\partial \theta_j$, called the **Jacobian matrix**, evaluated at $\boldsymbol{\theta}_0$; see Exercise 6.21. The asymptotic covariance

Copyright © 2021, Russell Davidson and James G. MacKinnon

matrix that appears in (6.67) is an $l \times l$ matrix. It has full rank l if $\boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}})$ is nonsingular and the matrix of derivatives $\boldsymbol{G}_0$ has full rank l.

In practice, the covariance matrix of $\hat{\boldsymbol{\gamma}}$ may be estimated by the matrix

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\gamma}}) \equiv \hat{\boldsymbol{G}}\, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\, \hat{\boldsymbol{G}}^{\top}, \tag{6.68}$$

where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{G}} \equiv \boldsymbol{G}(\hat{\boldsymbol{\theta}})$. This result, which is similar to (4.44), can be very useful. However, like all results based on asymptotic theory, it should be used with caution. As in the scalar case discussed above, the vector $\hat{\boldsymbol{\gamma}}$ cannot possibly be normally distributed if the vector $\hat{\boldsymbol{\theta}}$ is normally distributed when $\boldsymbol{g}(\cdot)$ is not an affine function.

6.9 Final Remarks

In this chapter, we have introduced the key concepts of confidence intervals. The idea is first to construct a family of test statistics for the null hypotheses that the parameter of interest is equal to a particular value. The limits of the confidence interval are then obtained by solving the equation that sets the statistic equal to the critical values given by some appropriate distribution. The critical values may be quantiles of a finite-sample distribution, such as Student's t distribution, quantiles of an asymptotic distribution, such as the standard normal distribution, or (as we will see in Chapter 7) quantiles of a bootstrap EDF. We also briefly discussed some procedures for constructing confidence regions.

All of the methods for constructing confidence intervals and regions that we have discussed require standard errors or, more generally, estimated covariance matrices. The second half of the chapter therefore deals with ways to estimate these under weaker assumptions than were made in Chapter 4. Much of this material is applicable to estimation methods other than OLS. Procedures for the estimation of covariance matrices in the presence of heteroskedasticity of unknown form, similar to those discussed in Section 6.4, are useful in the context of many different methods of estimation. So are procedures for the estimation of covariance matrices in the presence of autocorrelation or clustering, similar to those discussed in sections 5 and 6. The delta method, which was discussed in Section 6.8, is even more general, since it can be used whenever one parameter, or vector of parameters, is a nonlinear function of another.

Copyright © 2021, Russell Davidson and James G. MacKinnon

6.10 Exercises

6.1 Find the .025, .05, .10, and .20 quantiles of the standard normal distribution using a statistics package or some other computer program. Use these to obtain whatever quantiles of the $\chi^2(1)$ distribution you can.

6.2 Starting from the square of the t statistic (6.11), and using the $F(1, n-k)$ distribution, obtain a .99 confidence interval for the parameter β_2 in the classical normal linear model (5.18). Then show that you would have obtained the same interval by using (6.11) itself and the $t(n-k)$ distribution.

6.3 The file **group-earnings-data.txt** contains sorted data on four variables for 4,266 individuals. One of the variables is income, y, and the other three are dummy variables, d_1, d_2, and d_3, which correspond to different age ranges. Regress y on all three dummy variables. Then use the regression output to construct a .95 asymptotic confidence interval for the mean income of individuals who belong to age group 3.

6.4 Using the same data as Exercise 6.3, regress y on a constant for individuals in age group 3 only. Use the regression output to construct a .95 asymptotic confidence interval for the mean income of group 3 individuals. Explain why this confidence interval is not the same as the one you constructed previously.

6.5 Generate 999 realizations of a random variable that follows the $\chi^2(2)$ distribution, and find the .95 and .99 "quantiles" of the EDF, that is the 950^{th} and 990^{th} entries in the sorted list of the realizations. Compare these with the .95 and .99 quantiles of the $\chi^2(2)$ distribution.

6.6 Show that the F statistic for the null hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$ in the model (6.16), or, equivalently, for the null hypothesis that $\boldsymbol{\gamma}_2 = \mathbf{0}$ in (6.17), can be written as (6.18). Interpret the numerator of expression (6.18) as a random variable constructed from the multivariate normal vector $\hat{\boldsymbol{\beta}}_2$.

***6.7** Consider a regression model with just two explanatory variables, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, both of which are centered:

$$\boldsymbol{y} = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \boldsymbol{u}. \tag{6.69}$$

Let $\hat{\rho}$ denote the **sample correlation** of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Since both regressors are centered, the sample correlation is

$$\hat{\rho} \equiv \frac{\sum_{t=1}^{n} x_{t1} x_{t2}}{\left((\sum_{t=1}^{n} x_{t1}^2)(\sum_{t=1}^{n} x_{t2}^2) \right)^{1/2}},$$

where x_{t1} and x_{t2} are typical elements of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. This can be interpreted as the correlation of the joint EDF of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

Show that, under the assumptions of the classical normal linear model, the correlation between the OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ is equal to $-\hat{\rho}$. Which, if any, of the assumptions of this model can be relaxed without changing this result?

6.8 Consider the .95 level confidence region for the parameters β_1 and β_2 of the regression model (6.69). In the two-dimensional space $\mathcal{S}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ generated by

Copyright © 2021, Russell Davidson and James G. MacKinnon

the two regressors, consider the set of points of the form $\beta_{10}\boldsymbol{x}_1+\beta_{20}\boldsymbol{x}_2$, where (β_{10}, β_{20}) belongs to the confidence region. Show that this set is a circular disk with center at the OLS estimates $(\boldsymbol{x}_1\hat{\beta}_1 + \boldsymbol{x}_2\hat{\beta}_2)$. What is the radius of the disk?

6.9 Using the data in the file **group-earnings-data.txt**, regress y on all three dummy variables, and compute a heteroskedasticity-consistent standard error for the coefficient of d_3. Using these results, construct a .95 asymptotic confidence interval for the mean income of individuals that belong to age group 3. Compare this interval with the ones you constructed in Exercises 5.3 and 5.4.

⋆6.10 Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega},$$

where the number of observations, n, is equal to $3m$. The first three rows of the matrix $\boldsymbol{X}$ are

$$\begin{bmatrix} 1 & 4 \\ 1 & 8 \\ 1 & 15 \end{bmatrix},$$

and every subsequent group of three rows is identical to this first group. The covariance matrix $\boldsymbol{\Omega}$ is diagonal, with typical diagonal element equal to $\omega^2 x_{t2}^2$, where $\omega > 0$, and x_{t2} is the t^{th} element of the second column of $\boldsymbol{X}$.

What is the variance of $\hat{\beta}_2$, the OLS estimate of β_2? What is the probability limit, as $n \to \infty$, of the ratio of the conventional estimate of this variance, which incorrectly assumes homoskedasticity, to a heteroskedasticity-consistent estimate based on (6.29)?

6.11 Consider the linear regression model

$$\boldsymbol{y} = \delta_1\boldsymbol{d}_1 + \delta_2\boldsymbol{d}_2 + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{6.70}$$

which is similar to regression (4.83). The two regressors are dummy variables, with every element of $\boldsymbol{d}_2$ equal to 1 minus the corresponding element of $\boldsymbol{d}_1$. The vector $\boldsymbol{d}_1$ has n_1 elements equal to 1, and the vector $\boldsymbol{d}_2$ has $n_2 = n - n_1$ elements equal to 1. The covariance matrix $\boldsymbol{\Omega}$ is a diagonal matrix. The square root of the t^{th} diagonal element is ω_t, which can take on just two values that depend on the values of the regressors:

$$\omega_t = \sigma \text{ if } d_{t1} = 1 \text{ and } \omega_t = \lambda\sigma \text{ if } d_{t2} = 1.$$

As in Exercise 3.11, the parameter of interest is $\gamma \equiv \delta_2 - \delta_1$.

First, find the true standard error of $\hat{\gamma}$ as a function of n_1, n_2, σ, and λ. Then show that, when $n \to \infty$ and n_1/n tends to a constant ϕ as that happens, the square root of the appropriate element of an HC_0 covariance matrix provides a valid standard error for $\hat{\gamma}$.

6.12 Generate N simulated data sets, where N is between 1000 and 1,000,000, depending on the capacity of your computer, from each of the following two data generating processes:

$$\text{DGP 1:} \quad y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t, \quad u_t \sim \mathrm{N}(0,1)$$

$$\text{DGP 2:} \quad y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t, \quad u_t \sim \mathrm{N}(0,\sigma_t^2), \quad \sigma_t^2 = \mathrm{E}(y_t)^2.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

There are 50 observations, $\boldsymbol{\beta} = [1 \vdots 1 \vdots 1]$, and the data on the exogenous variables may be found in the file **mw-data.txt**. These data were originally used by MacKinnon and White (1985).

For each of the two DGPs and each of the N simulated data sets, construct .95 confidence intervals for β_1 and β_2 using the usual OLS covariance matrix and the HCCMEs HC_0, HC_1, HC_2, and HC_3. The OLS interval should be based on the Student's t distribution with 47 degrees of freedom, and the others should be based on the $\mathrm{N}(0,1)$ distribution. Report the proportion of the time that each of these confidence intervals included the true values of the parameters.

On the basis of these very limited results, which covariance matrix estimator would you recommend using in practice?

6.13 The file **house-price-data.txt** contains 546 observations. Regress the logarithm of the house price on a constant, the logarithm of lot size, and the other ten explanatory variables, as in Exercise 1.23. Then obtain .95 confidence intervals for σ and σ^2. Which of these intervals is closer to being symmetric? **Hint:** See Exercise 4.18.

Another way to form a confidence interval for σ is to make use of the fact that, under normality, the variance of s is approximately equal to $s^2/2n$. Form a second confidence interval for σ based on this result. How are the two intervals related?

⋆6.14 Give the explicit form of the $n \times n$ matrix $\boldsymbol{U}(j)$ for which $\hat{\boldsymbol{\Gamma}}(j)$, defined in (6.40), takes the form $n^{-1}\boldsymbol{W}^\top\boldsymbol{U}(j)\boldsymbol{W}$.

6.15 This question uses data from the file **house-price-data.txt**, which contains 546 observations. Regress the logarithm of the house price on a constant, the logarithm of lot size, and the other ten explanatory variables, as in Exercise 4.17. One of the explanatory variables is the number of storeys, which can take on the values 1, 2, 3, and 4. Construct a heteroskedasticity-robust .99 confidence interval for the difference in the expectation of the log price between a 3-storey house and a 2-storey house.

Now estimate a more general model, as in Exercise 4.17, in which the effect on log price of each number of storeys is allowed to differ. Using this more general model, construct a heteroskedasticity-robust .99 confidence interval for the difference in the expectation of the log price between a 3-storey house and a 2-storey house. Comment on the differences between the two intervals.

6.16 The file **earnings-data.txt** contains 46,302 observations on 32 variables taken from the Current Population Survey. Each observation is for a woman who lived and worked in California in the specified year. For a list of variables, see Exercise 4.26.

Regress the log of earnings on `age`, `age`$^2/100$, the four highest education dummy variables, and all of the year dummies, with no constant term. There should be 30 regressors. Report the coefficients on the age and education variables together with three estimated standard errors. One of the standard errors should be based on the assumption of IID disturbances, one should be based on the assumption that the disturbances are uncorrelated but possibly heteroskedastic, and one should be based on the assumption that they are clustered by age. What do you conclude about the validity of the three sets of standard errors?

Copyright © 2021, Russell Davidson and James G. MacKinnon

6.17 According to the estimates you obtained in Exercise 6.16, by what percentage (on average) do earnings for women with an advanced degree exceed those for women with a four-year university degree? Using cluster-robust standard errors, construct two .95 confidence intervals for this percentage increase. One of them should be based on the delta method, and the other should be obtained by transforming the ends of an interval based directly on one or more estimated coefficients.

6.18 This question also uses the data in **earnings-data.txt**. As in Exercise 4.29, create two dummy variables, **young** and **old**. The first of these is 1 if **age** ≤ 35, and the second is 1 if **age** ≥ 60. Add the two dummies to the regression of Exercise 6.16, and perform a Wald test, based on a CRVE, of the hypothesis that neither of them actually belongs in the regression. Report two P values, one based on the χ^2 distribution and one based on the F distribution.

6.19 Write down a second-order Taylor expansion of the nonlinear function $g(\hat{\theta})$ around θ_0, where $\hat{\theta}$ is an OLS estimator and θ_0 is the true value of the parameter θ. Explain why the last term is asymptotically negligible relative to the second term.

6.20 In Exercise 4.30, readers were asked to find the age at which the expectations of log earnings is maximized according to the regression model estimated in that exercise and also in Exercise 6.16. Using the delta method, construct two .95 confidence intervals for this age, one based on an HCCME and one based on a CRVE.

6.21 Using a multivariate first-order Taylor expansion, show that, if $\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta})$, the asymptotic covariance matrix of the l–vector $n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ is given by the $l \times l$ matrix $\boldsymbol{G}_0 \boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}}) \boldsymbol{G}_0^{\top}$. Here $\boldsymbol{\theta}$ is a k–vector with $k \geq l$, $\boldsymbol{G}_0$ is an $l \times k$ matrix with typical element $\partial g_i(\theta)/\partial \theta_j$, evaluated at $\boldsymbol{\theta}_0$, and $\boldsymbol{V}^{\infty}(\hat{\boldsymbol{\theta}})$ is the $k \times k$ asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

6.22 Suppose that $\gamma = \exp(\beta)$ and $\hat{\beta} = 1.324$, with a standard error of 0.2432. Calculate $\hat{\gamma} = \exp(\hat{\beta})$ and its standard error.

Construct two different .99 confidence intervals for γ. One should be based on (6.64), and the other should be based on (6.65).

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 7

The Bootstrap

7.1 Introduction

When we introduced the concept of a test statistic in Section 5.2, we specified that it should have a known distribution under the null hypothesis. This is a very strong requirement. In the context of linear regression models, it is generally valid only for the classical normal linear model, in which the regressors are exogenous and the disturbances are normally, independently, and identically distributed.

Traditionally, the way to make inferences for less restrictive models has been to rely on asymptotic theory. In Section 5.6, we relaxed certain assumptions and developed large-sample test statistics for which the distribution is known only approximately. Then, in Chapter 6, we discussed several ways to estimate standard errors under assumptions that are weaker than the usual assumption of IID disturbances. These included heteroskedasticity-robust standard errors in Section 6.4, HAC standard errors in Section 6.5, cluster-robust standard errors in Section 6.6, and standard errors for nonlinear functions of parameter estimates in Section 6.8. In all of these cases, t statistics and Wald statistics follow distributions that are known only asymptotically. This leads to tests and confidence intervals that are not exact in finite samples and may sometimes be very misleading.

With the remarkable increase in computing power over the past few decades, another way to make inferences when the finite-sample distribution of a test statistic is unknown has become very popular. Instead of comparing the test statistic with its asymptotic distribution, we compare it with the empirical distribution function, or EDF, of a large number of simulated test statistics; recall the definition of an EDF in equation (5.42). Such tests are usually referred to as **bootstrap tests**, and each of the simulated test statistics is computed using a randomly generated **bootstrap sample**. Ideally, these bootstrap samples closely resemble the actual sample. In many cases, bootstrap tests turn out to be more reliable in finite samples than asymptotic tests.

In this chapter, we discuss **bootstrap methods** in some detail. The term **bootstrap**, which was introduced in Efron (1979), is taken from the phrase "to pull oneself up by one's own bootstraps." Although the link between this improbable activity and simulated samples is tenuous at best, the term is by

now firmly established in statistics and econometrics. Some authors simply refer to **the bootstrap**, as if it were a single procedure. Such a terminology is extremely misleading, since there is actually an enormous number of ways to generate bootstrap samples and a great many ways to make inferences based on bootstrap samples.

Much of the chapter deals with bootstrap testing, but we also discuss bootstrap confidence intervals in some detail. In the next section, we introduce some basic concepts of computer simulation. In Section 7.3, we introduce the key ideas of Monte Carlo tests and bootstrap tests. In Section 7.4, we discuss ways to generate bootstrap data for regression models with IID disturbances. The so-called Golden Rules of Bootstrapping are presented in Section 7.5, along with a generic algorithm for performing bootstrap tests. In Section 7.6, we discuss ways to generate bootstrap data for regression models with heteroskedastic disturbances, and, in Section 7.7, we consider autocorrelation. In Section 7.8, we discuss bootstrap confidence intervals and bootstrap standard errors.

7.2 Basic Concepts of Computer Simulation

The idea of using a linear regression model to obtain simulated data was introduced in Section 2.3. First, it is necessary to choose a DGP contained in the model of interest. In keeping with the definition of a DGP as a unique recipe for simulation, this means that all parameters, all probability distributions, all exogenous variables, and the sample size must be uniquely specified in order to define a DGP. Once that is done, then, as we saw in Section 2.3 for a linear regression model, we have a simple algorithm for generating a simulated sample from the DGP. An essential element of the algorithm is generating random disturbances by use of a random number generator. Before going any further, some background on such things is in order.

Random Number Generators

A **random number generator**, or **RNG**, is a program for generating random numbers. Most such programs generate numbers that appear to be drawings from the uniform $\mathrm{U}(0,1)$ distribution, which can then be transformed into drawings from other distributions. There is a very large literature on RNGs. Useful references include Knuth (1998, Chapter 3), Gentle (1998), and L'Ecuyer (2012).

Random number generators have been a topic of active research for many decades. In the early days of computing, it was a scandal that an RNG called RANDU was used for much numerical computation, although Knuth described it as "truly horrible". Things are better now. Since computers are finite machines, any RNG has a **period**, that is, the number of seemingly independent random numbers it can generate before cycling back to the

Copyright © 2021, Russell Davidson and James G. MacKinnon

numbers it generated at first. The periods of RANDU and of many of its successors are much too small for the large-scale simulation experiments regularly performed nowadays. The sort of RNG most recommended at present is the **Mersenne twister** of Matsumoto and Nishimura (1998). The most commonly used Mersenne twister has a period of $2^{19937} - 1$, which is adequate for most purposes. The name comes from Marin Mersenne, a 17th-century French cleric, who studied the numbers now called **Mersenne primes**, which are prime numbers equal to an integer power of 2, minus 1. The thoroughly non-obvious principle that underlies this sort of RNG can, if needed, be used in order to create RNGs of still longer periods.

The raw output of an RNG is a sequence of numbers that have most of the properties of a genuinely IID sequence drawn from the $\mathrm{U}(0,1)$ distribution. We refer to the elements of the sequence as **random numbers**. There are several ways to use random numbers to generate drawings from a normal distribution. The simplest, but not the fastest, is to use the fact that, if Y is distributed as $\mathrm{U}(0,1)$, then $\Phi^{-1}(Y)$ is distributed as $\mathrm{N}(0,1)$; this follows from the result of Exercise 7.6. Most of the random number generators available in econometrics software packages use faster algorithms to generate drawings from the standard normal distribution, usually in a way entirely transparent to the user, who merely has to ask for so many independent drawings from $\mathrm{N}(0,1)$. Drawings from $\mathrm{N}(\mu, \sigma^2)$ can then be obtained by use of the formula (5.10).

In the valuable book by Devroye (1986), recipes are given for generating realizations from a wide class of distributions. For the distributions we encountered in Section 5.3, t, F, χ^2, *etc*, the definitions given in that section can be used directly to generate random variables from those distributions.

For many of the bootstrap methods to be discussed in later sections, we need to generate random positive integers between 1 and some upper limit, say M, rather than drawings from either the uniform or the standard normal distribution. If Y is once more a random number, all we need to do is compute $J = \lceil YM \rceil$, where the **ceiling function** $\lceil \cdot \rceil$ denotes the smallest integer no smaller than its argument, and then J is the random integer we require. When $YM \leq 1$, $J = 1$. When $1 < YM \leq 2$, $J = 2$, and so on. Thus J takes any particular integer value between 1 and M with probability $1/M$.

All good econometrics and statistics packages incorporate high-quality routines to generate random numbers. Econometricians who want to generate a number of simulated samples can simply call the RNG, and possibly provide a **seed**. The seed serves to initialize the sequence of random numbers. Most packages will pick a seed based on the system clock if one is not provided, so that different sequences of random numbers will be generated each time the RNG is called. In some cases, however, it is desirable (or even essential) to use the same sequence of random numbers repeatedly. This can be accomplished by providing the same seed every time the RNG is called.

Copyright © 2021, Russell Davidson and James G. MacKinnon

7.3 Bootstrap Testing

One of the most important uses of bootstrap methods is to perform hypothesis tests when exact tests are not available. There are generally good reasons to expect a **bootstrap test** to provide more reliable inferences than an asymptotic test based on the same test statistic. Moreover, in some cases, bootstrap tests are readily available when asymptotic tests are difficult or impossible to compute.

A hypothesis, null or alternative, can always be represented by a model $\mathbb{M}$, that is, the set of those DGPs that satisfy the requirements of the corresponding hypothesis. For instance, the null and alternative hypotheses (5.26) and (5.25) associated with an F test of several restrictions are both classical normal linear models. The most fundamental sort of null hypothesis that we can test is a **simple hypothesis**. Such a hypothesis is represented by a model that contains one and only one DGP. Unsurprisiingly, simple hypotheses are very rare in econometrics. The usual case is that of a **compound hypothesis**, which is represented by a model that contains more than one DGP. This can cause serious problems. Except in certain special cases, such as the exact tests in the classical normal linear model that we investigated in Section 5.4, a test statistic has different distributions under the different DGPs contained in the model. When this is so and we do not know just which DGP in the model generated our data, we cannot know the distribution of the test statistic.

In Section 6.2, we introduced the concept of a pivotal random variable. Such a random variable has the property that its distribution is the same for all DGPs in a model $\mathbb{M}$. When the distribution of a test statistic is known exactly under the null hypothesis, it must be pivotal with respect to the null-hypothesis model. But a test statistic can be pivotal without having a distribution that the investigator knows; we will discuss some examples below.

The principle of the bootstrap is that, when we want to use some function or functional of an unknown DGP, we use an estimate of that DGP, called a **bootstrap DGP**, in its place. Bootstrap tests attempt to get around the problem of statistics that are not pivotal by using the data to estimate the unknown DGP on which the distribution of the test statistic depends. How well this works depends on how sensitive the distribution is to the unknown parameters or other unknown features of the DGP and on how well the bootstrap DGP mimics the true DGP.

When we relaxed the assumptions of the classical normal linear model in Section 5.6, we obtained test statistics with unknown finite-sample distributions that depend on the distribution of the disturbances and perhaps on the parameters of the regression function. They are therefore not pivotal statistics. However, their asymptotic distributions are independent of such things, and are thus invariant across all the DGPs of the model that represents the null hypothesis. As we saw in Section 6.2, such statistics are said to be asymptotically pivotal.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Simulated P Values

The key idea of bootstrap testing is to use a bootstrap DGP to generate a (usually large) number of bootstrap samples, each of which is used to compute a bootstrap test statistic, say τ_b^*, for $b = 1, \ldots, B$. The τ_b^* are then used to calculate a **bootstrap P value** for the actual test statistic $\hat{\tau}$. A bootstrap P value is a particular type of **simulated P value** for which the simulated test statistics are obtained from bootstrap samples. In principle, and occasionally in practice, we may wish to obtain simulated test statistics in some other way.

The theoretical justification for using simulation to estimate P values is the Fundamental Theorem of Statistics, which we discussed in Section 5.5. It tells us that the empirical distribution of a set of independent drawings of a random variable generated by some DGP converges to the CDF of the random variable under that DGP. This is just as true of simulated drawings generated by the computer as for random variables generated by a natural random mechanism. Thus, if we knew that a certain test statistic was pivotal but did not know how it was distributed, we could select any DGP in the null model and generate simulated samples from it. For each of these, we could then compute the test statistic. If the simulated samples are mutually independent, the set of simulated test statistics thus generated constitutes a set of independent drawings from the distribution of the test statistic, and their EDF is a consistent estimate of the CDF of that distribution.

Suppose that we have computed a test statistic $\hat{\tau}$ which can be thought of as a realization of a random variable τ. We wish to test a null hypothesis represented by a model $\mathbb{M}$ for which τ is pivotal. In practice, $\hat{\tau}$ might be a t statistic, an F statistic, or some other type of test statistic. We want to reject the null whenever $\hat{\tau}$ is sufficiently large, as would be the case for an F statistic, a t statistic when the rejection region is in the upper tail, or a squared t statistic. If we denote by F the CDF of the distribution of τ under the null hypothesis, then the P value for a test based on $\hat{\tau}$ is

$$p(\hat{\tau}) \equiv 1 - F(\hat{\tau}). \tag{7.01}$$

Since $\hat{\tau}$ is computed directly from our original data, this P value can be estimated if we can estimate the CDF F evaluated at $\hat{\tau}$.

In order to estimate a P value by simulation, we choose some DGP in $\mathbb{M}$, and draw B samples of size n from it. How to choose B will be discussed in the next subsection; B is typically rather large, and $B = 999$ may often be a reasonable choice. We let $\boldsymbol{y}_b^*$, $b = 1, \ldots, B$, denote the simulated samples.

Using each of the $\boldsymbol{y}_b^*$, we compute a simulated test statistic τ_b^*, in exactly the same way that $\hat{\tau}$ was computed from the original data $\boldsymbol{y}$. We can then construct the EDF of the τ_b^* by the equivalent of equation (5.42):

$$\hat{F}^*(x) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(\tau_b^* \leq x). \tag{7.02}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Our estimate of the true P value (7.01) is therefore

$$\hat{p}^*(\hat{\tau}) = 1 - \hat{F}^*(\hat{\tau}) = 1 - \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\tau_b^* \leq \hat{\tau}) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \hat{\tau}). \tag{7.03}$$

The third equality in equations (7.03) can be understood by noting that the rightmost expression is the proportion of simulations for which τ_b^* is greater than $\hat{\tau}$, while the second expression from the right is one the minus the proportion for which τ_b^* is less than or equal to $\hat{\tau}$. These last two expressions are evidently equal.

We can see that $\hat{p}^*(\hat{\tau})$, like every P value, must lie between 0 and 1. For example, if $B = 999$, and 36 of the τ_b^* were greater than $\hat{\tau}$, we would have $\hat{p}^*(\hat{\tau}) = 36/999 = .036$. In this case, since $\hat{p}^*(\hat{\tau})$ is less than .05, we would reject the null hypothesis at the .05 level. Since the EDF converges to the CDF of the τ_b^*, it follows that, if B were infinitely large, this procedure would yield an exact test, and the outcome of the test would be the same as if we computed the P value analytically using the CDF of τ. In fact, as we will see shortly, this procedure yields an exact test even for finite values of B, provided B is chosen in a certain way.

The simulated P value (7.03) is **one-tailed**. It is appropriate for a test that rejects whenever the test statistic is sufficiently extreme in the upper tail, such as a Wald test. However, it is not appropriate for a test that rejects in both tails, such as a t test. There are two ways to compute simulated P values for such tests, and they can sometimes yield very different results.

If we are willing to assume that τ is symmetrically distributed around zero, then we can use the **symmetric simulated *P* value**

$$\hat{p}_{\mathrm{s}}^*(\hat{\tau}) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\big(|\tau_b^*| > |\hat{\tau}|\big), \tag{7.04}$$

which effectively converts a two-tailed test into a one-tailed test. If we are not willing to make this assumption, which can be seriously incorrect for a test statistic that is based on a biased parameter estimate, we can instead use the **equal-tail simulated *P* value**

$$\hat{p}_{\mathrm{et}}^*(\hat{\tau}) = 2\min\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\tau_b^* \leq \hat{\tau}),\ \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \hat{\tau})\right). \tag{7.05}$$

Here we actually perform two tests, one against values in the lower tail of the distribution and the other against values in the upper tail. We take the minimum because it corresponds to whichever tail of the EDF $\hat{\tau}$ actually lies in. The factor of 2 is necessary to take account of the fact that we did this. Without it, $\hat{p}_{\mathrm{et}}^*$ would lie between 0 and 0.5.

Copyright © 2021, Russell Davidson and James G. MacKinnon

If the mean of the τ_b^* is far from zero, the values of $\hat{p}_{\mathrm{s}}^*$ and $\hat{p}_{\mathrm{et}}^*$ may be very different. When there is no reason to believe that τ is not symmetrically distributed around zero, and the two P values are similar, it probably makes sense to rely on $\hat{p}_{\mathrm{s}}^*$. However, when they differ substantially, it is much better to rely on $\hat{p}_{\mathrm{et}}^*$; see Exercise 7.1.

Equation (7.04) is intended to be applied to test statistics that can take either sign, such as t statistics. For test statistics that are always positive, such as ones that are asymptotically χ^2, it usually makes no sense to use this equation. Instead, equation (7.03) is usually applicable. We would use an equal-tail P value only if we wanted to reject for small values of the test statistic as well as for large ones.

Equations (7.03), (7.04), and (7.05) imply that the results of a bootstrap test are invariant to monotonically increasing transformations of the test statistic. Applying the same transformation to all the test statistics does not affect the rank of $\hat{\tau}$ in the sorted list of $\hat{\tau}$ and the τ_b^*, and therefore it does not affect the bootstrap P value. For example, it is easy to see from equation (7.04) that we would obtain exactly the same results if we replaced $|\hat{\tau}|$ and $|\tau_b^*|$ by $\hat{\tau}^2$ and τ_b^{*2}.

Monte Carlo Tests

The sort of test we have just described, which is based on simulating a pivotal test statistic, is called a **Monte Carlo test**. This sort of test was first proposed by Dwass (1957); Dufour and Khalaf (2001) provides a more detailed introduction. Simulation experiments in general are often referred to as **Monte Carlo experiments**, because they involve generating random numbers, as do the games played in casinos. Around the time that computer simulations first became possible, the most famous casino was the one in Monte Carlo. If computers had been developed just a little later, we would probably be talking now of Las Vegas tests and Las Vegas experiments.

We have seen that, for a Monte Carlo test, the simulated P value $p^*(\hat{\tau})$ converges to the true P value $p(\hat{\tau})$ as $B \to \infty$, and a test based on $p(\hat{\tau})$ is exact. This is a consequence of the Fundamental Theorem of Statistics and the fact that τ is pivotal. Perhaps more surprisingly, a Monte Carlo test can always be made exact without B becoming large, provided B is chosen so that it satisfies a certain condition. This condition is simply that, if we wish to perform a test at level α, then B should be chosen to satisfy the condition that $\alpha(B+1)$ is an integer. If $\alpha = .05$, the values of B that satisfy this condition are 19, 39, 59, and so on. If $\alpha = .01$, they are 99, 199, 299, and so on.

It is illuminating to see why B should be chosen in this way. Imagine that we sort the original test statistic $\hat{\tau}$ and the B bootstrap statistics τ_b^*, $b = 1, \ldots, B$, from largest to smallest. If τ is pivotal, then, under the null hypothesis, these are all independent drawings from the same distribution. Thus the rank r of $\hat{\tau}$ in the sorted set can have $B + 1$ possible values, $r = 0, 1, \ldots, B$, all of

Copyright © 2021, Russell Davidson and James G. MacKinnon

them equally likely under the null hypothesis. Here, r is defined in such a way that there are exactly r simulations for which $\tau_b^* > \hat{\tau}$. Thus, if $r = 0$, $\hat{\tau}$ is the largest value in the set, and if $r = B$, it is the smallest. The estimated P value $\hat{p}^*(\hat{\tau})$ is just r/B.

The Monte Carlo test rejects if $r/B < \alpha$, that is, if $r < \alpha B$. Under the null, the probability that this inequality is satisfied is the proportion of the $B+1$ possible values of r that satisfy it. We may use the **floor function**, and denote by $\lfloor \alpha B \rfloor$ the largest integer that is no greater than αB. It is then easy to see that there are exactly $\lfloor \alpha B \rfloor + 1$ values of r such that $r < \alpha B$, namely, $0, 1, \ldots, \lfloor \alpha B \rfloor$. Thus the probability of rejection is $(\lfloor \alpha B \rfloor + 1)/(B+1)$. If we equate this probability to α, we find that

$$\alpha(B+1) = \lfloor \alpha B \rfloor + 1.$$

Since the right-hand side above is the sum of two integers, a necessary condition for this equality to hold is that $\alpha(B+1)$ is a positive integer. It is also a sufficient condition, as can see as follows. Let $\alpha(B+1) = k$, a positive integer. Then $\alpha B = k - \alpha$, and, since $0 \leq \alpha < 1$, $\lfloor \alpha B \rfloor = k - 1$. We saw that the rejection probability in this case is $(\lfloor \alpha B \rfloor + 1)/(B+1)$, and this is $k/(B+1) = \alpha$, by the definition of k. Therefore, the probability of Type I error is precisely α if and only if $\alpha(B+1)$ is a positive integer.

By a similar argument, it can be shown that $(\alpha/2)(B+1)$ must be an integer if we are to obtain an exact test based on an equal-tail P value computed with equation (7.05).

Although this reasoning is rigorous only if τ is an exact pivot, experience shows that bootstrap P values based on nonpivotal statistics are less misleading if $\alpha(B+1)$ is an integer.

As a concrete example, suppose that $\alpha = .05$ and $B = 99$. Then there are 5 out of 100 values of r, namely, $r = 0, 1, \ldots, 4$, that would lead us to reject the null hypothesis. Since these are equally likely if the test statistic is pivotal, we make a Type I error precisely 5% of the time, and the test is exact. But suppose instead that $B = 89$. Since the same 5 values of r would still lead us to reject the null, we would now do so with probability $5/90 = .0556$.

Bootstrap P Values

Although pivotal test statistics do arise from time to time, most test statistics in econometrics are not pivotal. The vast majority of them are, however, asymptotically pivotal. If a test statistic has a known asymptotic distribution that does not depend on anything unobservable, as do t and F statistics under the relatively weak assumptions of Section 5.5, then it is certainly asymptotically pivotal. Even if it does not follow a known asymptotic distribution, a test statistic may be asymptotically pivotal.

A statistic that is not an exact pivot cannot be used for a Monte Carlo test. However, approximate P values for statistics that are only asymptotically pivotal, or even nonpivotal, can be obtained by use of the bootstrap. This method

Copyright © 2021, Russell Davidson and James G. MacKinnon

can be a valuable alternative to the large sample tests based on asymptotic theory that we discussed in previous sections.

The difference between a Monte Carlo test and a bootstrap test is that for the former, the DGP is assumed to be known, whereas, for the latter, it is necessary to estimate a **bootstrap DGP** from which to draw the simulated samples. Unless the null hypothesis under test is a simple hypothesis, the DGP that generated the original data is unknown, and so it cannot be used to generate simulated data. The bootstrap DGP is an estimate of the unknown true DGP. The hope is that, if the bootstrap DGP is close, in some sense, to the true one, then data generated by the bootstrap DGP will be similar to data that would have been generated by the true DGP, if it were known. If so, then a simulated P value obtained by use of the bootstrap DGP is close enough to the true P value to allow accurate inference.

Even for models as simple as the linear regression model, there are many ways to specify the bootstrap DGP. The key requirement is that it should satisfy the restrictions of the null hypothesis. If this is assured, then how well a bootstrap test performs in finite samples depends on how good an estimate the bootstrap DGP is of the process that would have generated the test statistic if the null hypothesis were true. In the next subsection, we discuss bootstrap DGPs for regression models.

7.4 Bootstrap DGPs for Regression Models

If the null and alternative hypotheses are regression models, the simplest approach is to estimate the model that corresponds to the null hypothesis and then use the estimates thus obtained to generate the bootstrap samples, under the assumption that the disturbances are normally distributed. We considered examples of such procedures in Section 2.3 and in Exercise 2.26.

Since bootstrapping is quite unnecessary in the context of the classical normal linear model, we will take for our example a linear regression model with normal disturbances, and all but one of the regressors exogenous, the other being the lagged dependent variable:

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Z}_t\boldsymbol{\gamma} + \delta y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \tag{7.06}$$

where $\boldsymbol{X}_t$ and $\boldsymbol{\beta}$ each have $k_1 - 1$ elements, $\boldsymbol{Z}_t$ and $\boldsymbol{\gamma}$ each have k_2 elements, and the null hypothesis is that $\boldsymbol{\gamma} = \boldsymbol{0}$. Thus the model that represents the null is

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \delta y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \tag{7.07}$$

The observations are assumed to be indexed in such a way that y_0 is observed, along with n observations on y_t, $\boldsymbol{X}_t$, and $\boldsymbol{Z}_t$ for $t = 1, \ldots, n$. By estimating the models (7.06) and (7.07) by OLS, we can compute the F statistic for

Copyright © 2021, Russell Davidson and James G. MacKinnon

$\boldsymbol{\gamma} = \mathbf{0}$, which we will call $\hat{\tau}$. Because the regression function contains a lagged dependent variable, however, the F test based on $\hat{\tau}$ is not exact.

The model (7.07) is a fully specified parametric model, which means that each set of parameter values for $\boldsymbol{\beta}$, δ, and σ^2 defines just one DGP. The simplest type of bootstrap DGP for fully specified models is given by the **parametric bootstrap**. The first step in constructing a parametric bootstrap DGP is to estimate (7.07) by OLS, yielding the restricted estimates $\tilde{\boldsymbol{\beta}}$, $\tilde{\delta}$, and $\tilde{s}^2 \equiv \text{SSR}(\tilde{\boldsymbol{\beta}}, \tilde{\delta})/(n - k_1)$. Then the bootstrap DGP is given by

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + \tilde{\delta}y_{t-1}^* + u_t^*, \quad u_t^* \sim \text{NID}(0, \tilde{s}^2), \tag{7.08}$$

which is just the element of the model (7.07) characterized by the parameter estimates under the null, with stars to indicate that the data are simulated. Notice that y_{t-1}^* rather than y_{t-1} appears on the right-hand side of equation (7.08). This means that each bootstrap sample is constructed recursively, observation by observation:

$$\begin{aligned} y_1^* &= \boldsymbol{X}_1\tilde{\boldsymbol{\beta}} + \tilde{\delta}y_0 + u_1^* \\ y_2^* &= \boldsymbol{X}_2\tilde{\boldsymbol{\beta}} + \tilde{\delta}y_1^* + u_2^* \\ &\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ y_n^* &= \boldsymbol{X}_n\tilde{\boldsymbol{\beta}} + \tilde{\delta}y_{n-1}^* + u_n^*. \end{aligned} \tag{7.09}$$

Every bootstrap sample here is conditional on the observed value of y_0. There are other ways of dealing with pre-sample values of the dependent variable, but this is certainly the most convenient, and it may, in many circumstances, be the only method that is feasible.

Of course, the recursion in (7.09) will explode if $|\tilde{\delta}| > 1$, and the resulting bootstrap samples will probably not resemble the actual sample, especially when n is large. This should rarely be a problem if the model (7.07) is correctly specified and the true value of δ is substantially less than one in absolute value, in part because $\tilde{\delta}$ is biased towards zero; see Exercises 4.1 and 4.2. Thus, in most cases, we would expect the estimate $\tilde{\delta}$ to satisfy the stationarity condition that $|\tilde{\delta}| < 1$. If it does not, we can always replace $\tilde{\delta}$ by a number slightly below unity, such as 0.99.

The rest of the procedure for computing a bootstrap P value is identical to the one for computing a simulated P value for a Monte Carlo test. For each of the B bootstrap samples, $\boldsymbol{y}_b^*$, a bootstrap test statistic τ_b^* is computed from $\boldsymbol{y}_b^*$ in just the same way as $\hat{\tau}$ was computed from the original data, $\boldsymbol{y}$. The bootstrap P value $\hat{p}^*(\hat{\tau})$ is then computed by formula (7.03).

A Nonparametric Bootstrap DGP

The parametric bootstrap procedure that we have just described, based on the DGP (7.08), does not allow us to relax the strong assumption that the

Copyright © 2021, Russell Davidson and James G. MacKinnon

disturbances are normally distributed. How can we construct a satisfactory bootstrap DGP if we extend the models (7.06) and (7.07) to admit nonnormal disturbances? If we knew the true distribution of the disturbances, whether or not it was normal, we could always generate the $\boldsymbol{u}^*$ from it. Since we do not know it, we will have to find some way to estimate this distribution.

Under the null hypothesis, the OLS residual vector $\tilde{\boldsymbol{u}}$ for the restricted model is a consistent estimator of the disturbance vector $\boldsymbol{u}$. This is an immediate consequence of the consistency of the OLS estimator itself. In the particular case of model (7.07), we have for each t that

$$\operatorname*{plim}_{n\to\infty} \tilde{u}_t = \operatorname*{plim}_{n\to\infty} \big(y_t - \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} - \tilde{\delta}y_{t-1}\big) = y_t - \boldsymbol{X}_t\boldsymbol{\beta}_0 - \delta_0 y_{t-1} = u_t,$$

where $\boldsymbol{\beta}_0$ and δ_0 are the parameter values for the true DGP. This means that, if the u_t are mutually independent drawings from the disturbance distribution, then so are the residuals $\tilde{u}_t$, asymptotically.

From the Fundamental Theorem of Statistics, we know that the empirical distribution function of the disturbances is a consistent estimator of the unknown CDF of their distribution. Because the residuals consistently estimate the disturbances, it follows that the EDF of the residuals is also a consistent estimator of the CDF of the disturbance distribution. Thus, if we draw bootstrap disturbances from the empirical distribution of the residuals, we are drawing them from a distribution that tends to the true distribution of the disturbances as $n \to \infty$. This is completely analogous to using estimated parameters in the bootstrap DGP that tend to the true parameters as $n \to \infty$.

Drawing simulated disturbances from the empirical distribution of the residuals is called **resampling**. In order to **resample the residuals**, all n residuals are, metaphorically speaking, thrown into a hat and then randomly pulled out one at a time, with replacement. Thus each bootstrap sample contains some of the residuals exactly once, some of them more than once, and some of them not at all. The value of each drawing must be the value of one of the residuals, with equal probability for each residual. This is precisely what we mean by the empirical distribution of the residuals.

To resample concretely rather than metaphorically, we can proceed as follows. First, we draw a random number Y from the $\text{U}(0, 1)$ distribution. Then, as described in Section 7.2, we use Y to construct a positive integer J that takes on all the values $1, 2, \ldots, n$ with equal probability. The bootstrap disturbance is then the J^{th} residual. Repeating this procedure n times yields a single set of bootstrap disturbances drawn from the empirical distribution of the residuals.

As an example of how resampling works, suppose that $n = 10$, and the ten residuals are

$$6.45,\ 1.28,\ -3.48,\ 2.44,\ -5.17,\ -1.67,\ -2.03,\ 3.58,\ 0.74,\ -2.14.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Notice that these numbers sum to zero. Now suppose that, when forming one of the bootstrap samples, the ten drawings from the set $\{1, 2, \ldots, n\}$ are

$$7,\ 3,\ 8,\ 3,\ 10,\ 2,\ 9,\ 9,\ 2,\ 3.$$

The disturbances for this bootstrap sample are then

$$-2.03,\ -3.48,\ 3.58,\ -3.48,\ -2.14,\ 1.28,\ 0.74,\ 0.74,\ 1.28,\ -3.48.$$

Some of the residuals appear just once in this particular sample, some of them (numbers 2, 3, and 9) appear more than once, and some of them (numbers 1, 4, 5, and 6) do not appear at all. On average, however, each of the residuals appears once in each of the bootstrap samples.

If we adopt this resampling procedure, we can write the bootstrap DGP as

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + \tilde{\delta}y_{t-1}^* + u_t^*, \quad u_t^* \sim \text{EDF}(\tilde{\boldsymbol{u}}), \tag{7.10}$$

where $\text{EDF}(\tilde{\boldsymbol{u}})$ denotes the distribution that assigns probability $1/n$ to each of the elements of the residual vector $\tilde{\boldsymbol{u}}$. The DGP (7.10) is one form of what is usually called a **nonparametric bootstrap**, although, since it still uses the parameter estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\delta}$, it should really be called **semiparametric** rather than nonparametric. A more neutral term that we will favor is **resampling bootstrap**. Once bootstrap disturbances have been drawn by resampling, bootstrap samples can be created by the recursive procedure.

The empirical distribution of the residuals may fail to satisfy some of the properties that the null hypothesis imposes on the true distribution of the disturbances, and so the DGP (7.10) may fail to belong to the null hypothesis. One case in which this failure has grave consequences arises when the regression (7.07) does not contain a constant term, because then the sample mean of the residuals is not, in general, equal to 0. The expectation of the EDF of the residuals is simply their sample mean; recall Exercise 2.1. Thus, if the bootstrap disturbances are drawn from a distribution with nonzero expectation, the bootstrap DGP lies outside the null hypothesis. It is, of course, simple to correct this problem. We just need to *center* the residuals before throwing them into the hat, by subtracting their mean $\bar{u}$. When we do this, the bootstrap disturbances are drawn from $\text{EDF}(\tilde{\boldsymbol{u}} - \bar{u}\boldsymbol{\iota})$, a distribution that does indeed have an expectation of 0.

A somewhat similar argument gives rise to an improved bootstrap DGP. If the sample mean of the restricted residuals is 0, then the variance of their empirical distribution is the second moment $n^{-1}\sum_{t=1}^n \tilde{u}_t^2$. Thus, by using the definition (4.63) of $\tilde{s}^2$ in Section 4.6, we see that the variance of the empirical distribution of the residuals is $\tilde{s}^2(n-k_1)/n$. Since we do not know the value of σ_0^2, we cannot draw from a distribution with exactly that variance. However, as with the parametric bootstrap (7.08), we can at least draw from a distribution with variance $\tilde{s}^2$. This is easy to do by drawing from the EDF

Copyright © 2021, Russell Davidson and James G. MacKinnon

of the **rescaled residuals**, which are obtained by multiplying the OLS residuals by $(n/(n-k_1))^{1/2}$. If we resample these rescaled residuals, the distribution of the bootstrap disturbances is

$$\text{EDF}\left(\left(\frac{n}{n-k_1}\right)^{1/2}\tilde{\boldsymbol{u}}\right), \tag{7.11}$$

which has variance $\tilde{s}^2$. A somewhat more complicated approach, based on the result (4.58), is explored in Exercise 7.7.

Although they may seem strange, these resampling procedures often work astonishingly well, except perhaps when the sample size is very small or the distribution of the disturbances is very unusual; see Exercise 7.5. If the distribution of the disturbances displays substantial skewness (that is, a nonzero third moment) or excess kurtosis (that is, a fourth moment greater than $3\sigma_0^4$), then there is a good chance that the EDF of the recentered and rescaled residuals does so as well.

Other methods for bootstrapping regression models nonparametrically and semiparametrically are discussed by Efron and Tibshirani (1993), Davison and Hinkley (1997), and Horowitz (2001), which also discuss many other aspects of the bootstrap. A more advanced book, which deals primarily with the relationship between asymptotic theory and the bootstrap, is Hall (1992).

How Many Bootstraps?

It is important that B should be sufficiently large, since two problems can arise if it is not. The first problem is that the outcome of the test depends on the sequence of random numbers used to generate the bootstrap samples. Different investigators may therefore obtain different results, even though they are using the same data and testing the same hypothesis. The second problem, which we will discuss in the next section, is that the ability of a bootstrap test to reject a false null hypothesis declines as B becomes smaller. As a rule of thumb, we suggest choosing $B = 999$. If calculating the τ_b^* is inexpensive and the outcome of the test is at all ambiguous, it may be desirable to use a larger value, like 9,999. On the other hand, if calculating the τ_b^* is very expensive and the outcome of the test is unambiguous, because $\hat{p}^*$ is far from α, it may be safe to use a value as small as 99.

It is not actually necessary to choose B in advance. An alternative approach, which is a bit more complicated but can save a lot of computer time, has been proposed by Davidson and MacKinnon (2000). The idea is to calculate a sequence of estimated P values, based on increasing values of B, and to stop as soon as the estimate $\hat{p}^*$ allows us to be very confident that p^* is either greater or less than α. For example, we might start with $B = 99$, then perform an additional 100 simulations if we cannot be sure whether or not to reject the null hypothesis, then perform an additional 200 simulations if we still cannot be sure, and so on. Eventually, we either stop when we are confident that the

Copyright © 2021, Russell Davidson and James G. MacKinnon

null hypothesis should or should not be rejected, or when B has become so large that we cannot afford to continue.

7.5 The Golden Rules of Bootstrapping

Although bootstrap tests based on test statistics that are merely asymptotically pivotal are not exact, there are strong theoretical reasons to believe that they generally perform better than tests based on approximate asymptotic distributions. The errors committed by both asymptotic and bootstrap tests diminish as the sample size n increases, but those committed by bootstrap tests diminish more rapidly. The fundamental theoretical result on this point is due to Beran (1988). The results of a number of Monte Carlo experiments have provided strong support for this proposition. References include Horowitz (1994), Godfrey (1998), and Davidson and MacKinnon (1999a, 1999b, 2002a).

If a test statistic τ is asymptotically pivotal for a given model $\mathbb{M}$, then its finite-sample distribution should not vary too much as a function of the specific DGP, μ say, within that model. Under conventional asymptotic constructions, the distance between the distribution of τ under the DGP μ for sample size n and that for infinite n tends to zero like some negative power of n, commonly $n^{-1/2}$. The concept of "distance" between distributions can be realised in various ways, some ways being more relevant for bootstrap testing than others.

Heuristically speaking, if the distance between the finite-sample distribution for any DGP $\mu \in \mathbb{M}$ and the limiting distribution is of order $n^{-\delta}$ for some $\delta > 0$, then, since the limiting distribution is the same for all $\mu \in \mathbb{M}$, the distance between the finite-sample distributions for two DGPs μ_1 and μ_2 in $\mathbb{M}$ is also of order $n^{-\delta}$. If now the distance between μ_1 and μ_2 is also small, in some sense, say of order $n^{-\varepsilon}$, it should be the case that the distance between the distributions of τ under μ_1 and μ_2 should be of order $n^{-(\delta+\varepsilon)}$. In typical cases, $\delta = \varepsilon = 1/2$, so that the distance between the true and bootstrap DGPs is $O_p(n^{-1})$, rather than the distance between the true and limiting asymptotic DGPs. This is an instance of an **asymptotic refinement** for the bootstrap.

Bootstrap Versus Asymptotic Tests

We can illustrate this by means of an example. Consider the following simple special case of the linear regression model (7.06)

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \tag{7.12}$$

where the null hypothesis is that $\beta_3 = 0.9$. A Monte Carlo experiment to investigate the properties of tests of this hypothesis would work as follows. First, we fix a DGP in the model (7.12) by choosing values for the parameters.

Copyright © 2021, Russell Davidson and James G. MacKinnon

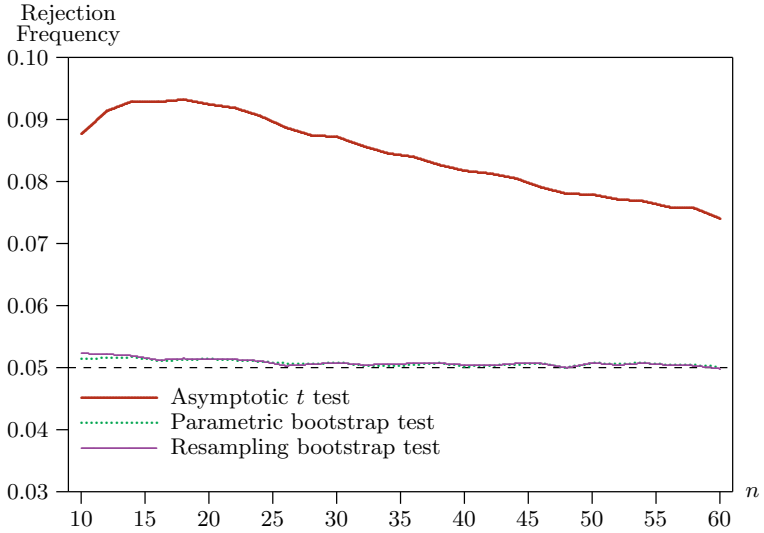


Figure 7.1 Rejection frequencies for bootstrap and asymptotic tests

Here $\beta_3 = 0.9$, and so we investigate only what happens under the null hypothesis. For each **replication**, we generate an artificial data set from our chosen DGP and compute the ordinary t statistic for $\beta_3 = 0.9$. We then compute three P values. The first of these, for the asymptotic test, is computed using Student's t distribution with $n - 3$ degrees of freedom, and the other two are bootstrap P values from the parametric and resampling bootstraps, with residuals rescaled using (7.11), for $B = 199$.[1] We perform many replications and record the frequencies with which tests based on the three P values reject at the .05 level. Figure 7.1 shows the rejection frequencies based on 500,000 replications for each of 31 sample sizes: $n = 10, 12, 14, \ldots, 60$.

The results of this experiment are striking. The asymptotic test overrejects quite noticeably, although it gradually improves as n increases. In contrast, the two bootstrap tests overreject only very slightly. Their rejection frequencies are always very close to the nominal level of .05, and they approach that level quite quickly as n increases. For the very smallest sample sizes, the parametric bootstrap seems to outperform the resampling one, but, for most sample sizes, there is nothing to choose between them.

[1] We used $B = 199$, a smaller value than we would ever recommend using in practice, in order to reduce the costs of doing the Monte Carlo experiments. Because experimental errors tend to cancel out across replications, this does not materially affect the results of the experiments.

Copyright © 2021, Russell Davidson and James G. MacKinnon

This example is, perhaps, misleading in one respect. For linear regression models, asymptotic t and F tests generally do not perform as badly as the asymptotic t test does here. For example, the t test for $\beta_3 = 0$ in (7.12) performs much better than the t test for $\beta_3 = 0.9$; it actually underrejects moderately in small samples. However, the example is not at all misleading in suggesting that bootstrap tests often perform extraordinarily well, even when the corresponding asymptotic test does not perform well at all.

The Golden Rules

Since in testing the bootstrap is used to estimate the distribution of a test statistic under the null hypothesis, the first golden rule of bootstrapping is:

Golden Rule 1:

> The bootstrap DGP must belong to the model $\mathbb{M}_0$ that represents the null hypothesis.

It is not always possible, or, even if it is, it may be difficult to obey this rule in some cases, as we will see with confidence intervals. In that case, we may use the common technique of changing the null hypothesis so that the bootstrap DGP that is to be used does satisfy it.

If, in violation of this rule, the null hypothesis tested by the bootstrap statistics is not satisfied by the bootstrap DGP, a bootstrap test can be wholly lacking in power. Test power springs from the fact that a statistic has different distributions under the null and the alternative. Bootstrapping under the alternative confuses these different distributions, and so leads to completely unreliable inference, even in the asymptotic limit.

Whereas Golden Rule 1 must be satisfied in order to have an asymptotically justified test, Golden Rule 2 is concerned rather with making the probability of rejecting a true null with a bootstrap test as close as possible to the significance level. It is motivated by the argument of Beran discussed earlier.

Golden Rule 2:

> Unless the test statistic is pivotal for the null model $\mathbb{M}_0$, the bootstrap DGP should be as good an estimate of the true DGP as possible, under the assumption that the true DGP belongs to $\mathbb{M}_0$.

How this second rule can be followed depends very much on the particular test being performed, but quite generally it means that we want the bootstrap DGP to be based on estimates that are *efficient* under the null hypothesis.

These rules are based on a similar pair of rules set out in Hall and Wilson (1991).

Copyright © 2021, Russell Davidson and James G. MacKinnon

The Algorithm

Once the sort of bootstrap DGP has been chosen, the procedure for conducting a bootstrap test based on simulated bootstrap samples follows the following algorithm

(i) Compute the test statistic from the original sample; call its realised value $\hat{\tau}$.

(ii) Determine the realisations of all other data-dependent things needed to set up the bootstrap DGP.

(iii) Generate B bootstrap samples, and for each one compute a realisation of the bootstrap statistic, τ_b^*, $b = 1, \dots B$. It is prudent to choose B so that $\alpha(B+1)$ is an integer for all interesting significance levels α, typically 1%, 5%, and 10%.

(iv) Compute the simulated bootstrap P value as the proportion of bootstrap statistics τ_b^* that are more extreme than $\hat{\tau}$. For a statistic that rejects for large values, for instance, we have

$$P_{\text{bs}} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(\tau_b^* > \hat{\tau}),$$

where $\mathbb{I}(\cdot)$ is an indicator function, with value 1 if its Boolean argument is true, and 0 if it is false.

The bootstrap test rejects the null hypothesis at significance level α if $P_{\text{bs}} < \alpha$.

The Power of Bootstrap Tests

The power of a bootstrap test depends on B, the number of bootstrap samples, and the reason for this fact is illuminating. If, to any test statistic, we add random noise independent of the statistic, we inevitably reduce the power of tests based on that statistic. The bootstrap P value $\hat{p}^*(\hat{\tau})$ defined in (7.03) is simply an estimate of the **ideal bootstrap P value**

$$p^*(\hat{\tau}) \equiv \Pr(\tau > \hat{\tau}) = \operatorname*{plim}_{B \to \infty} \hat{p}^*(\hat{\tau}),$$

where $\Pr(\tau > \hat{\tau})$ is evaluated under the bootstrap DGP, conditional on the realized $\hat{\tau}$. When B is finite, $\hat{p}^*$ differs from p^* because of random variation in the bootstrap samples. This random variation is generated in the computer, and is therefore completely independent of the random variable τ. The bootstrap testing procedure presented at the end of the preceding subsection incorporates this random variation, and in so doing it reduces the power of the test.

Another example of how randomness affects test power is provided by the tests z_{β_2} and t_{β_2}, which were discussed in Section 5.4. Recall that z_{β_2} follows the $\mathrm{N}(0,1)$ distribution, because σ is known, and t_{β_2} follows the $t(n-k)$

Copyright © 2021, Russell Davidson and James G. MacKinnon

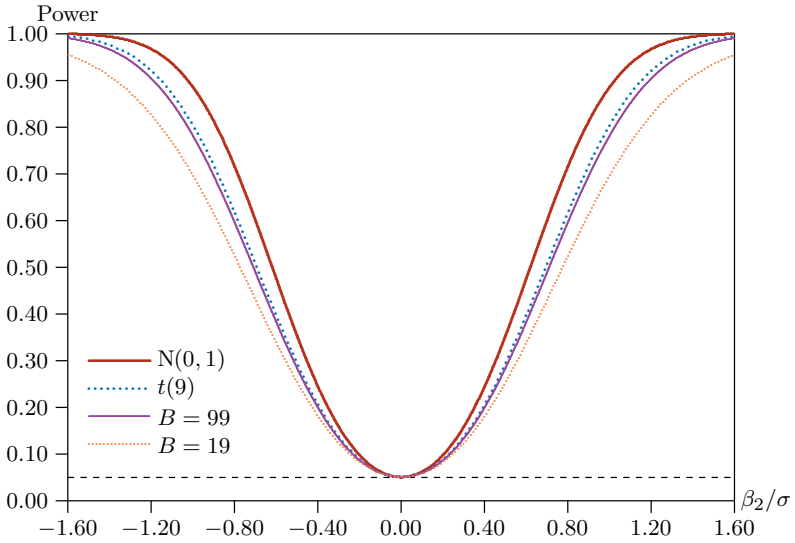


Figure 7.2 Power functions for tests at the .05 level

distribution, because σ has to be estimated. As equation (5.23) shows, t_{β_2} is equal to z_{β_2} times the random variable σ/s, which has the same distribution under the null and alternative hypotheses, and is independent of z_{β_2}. Therefore, multiplying z_{β_2} by σ/s simply adds independent random noise to the test statistic. This additional randomness requires us to use a larger critical value, and that in turn causes the test based on t_{β_2} to be less powerful than the test based on z_{β_2}.

Both types of power loss are illustrated in Figure 7.2. It shows power functions for four tests at the .05 level of the null hypothesis that $\beta_2 = 0$ in the simple model used to generate Figure 5.7, but with only 10 observations. All four tests are exact, as can be seen from the fact that, in all cases, power equals .05 when $\beta_2 = 0$. For all values of $\beta_2 \neq 0$, there is a clear ordering of the four curves in Figure 7.2. The highest curve is for the test based on z_{β_2}, which uses the $\mathrm{N}(0,1)$ distribution and is available only when σ is known. The next three curves are for tests based on t_{β_2}. The loss of power from using t_{β_2} with the $t(9)$ distribution, instead of z_{β_2} with the $\mathrm{N}(0,1)$ distribution, is quite noticeable. Of course, 10 is a very small sample size; the loss of power from not knowing σ would be very much less for more reasonable sample sizes. There is a further loss of power from using a bootstrap test with finite B. This further loss is quite modest when $B = 99$, but it is substantial when $B = 19$.

Figure 7.2 suggests that the loss of power from using bootstrap tests is generally modest, except when B is very small. However, readers should be warned that the loss can be more substantial in other cases. A reasonable

Copyright © 2021, Russell Davidson and James G. MacKinnon

rule of thumb is that power loss is very rarely a problem when $B = 999$, and that it is never a problem when $B = 9{,}999$.

7.6 Heteroskedasticity

All the bootstrap DGPs that we have looked at so far are based on models where either the observations are IID, or else some set of quantities that can be estimated from the data, like the disturbances of a regression model, are IID, or at least white noise. Situations in which disturbances are not white noise were discussed in Chapter 6, in the context of the various sorts of covariance matrix estimators supposedly robust to the phenomena of heteroskedasticity, in Section 6.4, autocorrelation, in Section 6.5, and clustering, in Section 6.6. In any of these circumstances, if the covariance matrix of the disturbances is known, or can be consistently estimated, then bootstrap disturbances can be generated so as to be jointly distributed with that covariance matrix. But this is often not the case. In this section, we will see how the bootstrap can be used effectively in the presence of heteroskedasticity; in the next, we will consider autocorrelation. Dealing with clustering will be postponed until the next chapter.

If the disturbances of a regression are heteroskedastic, with an unknown pattern of heteroskedasticity, there is nothing that is even approximately white noise. There exist of course test statistics robust to heteroskedasticity of unknown form, based on one of the numerous variants of the Eicker-White Heteroskedasticity Consistent Covariance Matrix Estimator (HCCME) discussed in Section 6.4. Use of an HCCME gives rise to statistics that are approximately pivotal for models that admit heteroskedasticity of unknown form.

For bootstrapping, it is very easy to satisfy Golden Rule 1, since either a parametric bootstrap or a resampling bootstrap of the sort we have described in Section 7.4 belongs to a null hypothesis model that, since it allows heteroskedasticity, must also allow the special case of homoskedasticity. But Golden Rule 2 poses a more severe challenge.

Pairs Bootstrap

The first suggestion for bootstrapping models with heteroskedasticity bears a variety of names: among them the (y, X) bootstrap or the **pairs bootstrap**. The approach was proposed in Freedman (1981). Instead of resampling the dependent variable, or residuals, possibly centred or rescaled, one resamples **pairs** consisting of an observation of the dependent variable along with the set of explanatory variables for that same observation. One selects an index s at random from the set $1, \dots, n$, and then an observation of a bootstrap sample is the pair $(y_s, \boldsymbol{X}_s)$, where $\boldsymbol{X}_s$ is a row vector of all the explanatory variables for observation s.

Copyright © 2021, Russell Davidson and James G. MacKinnon

This bootstrap implicitly assumes that the pairs $(y_t, \boldsymbol{X}_t)$ are IID under the null hypothesis. Although this is still a restrictive assumption, ruling out any form of dependence among observations, it does allow for any sort of heteroskedasticity of y_t conditional of $\boldsymbol{X}_t$. The objects resampled are IID drawings from the *joint* distribution of y_t and $\boldsymbol{X}_t$. However, there is no obvious way by which the pairs bootstrap can be made to satisfy Golden Rule 1, let alone Rule 2.

In Flachaire (1999), this is partially rectified. It now resamples pairs $(\hat{u}_t, \boldsymbol{X}_t)$, where the $\hat{u}_t$ are the OLS residuals from estimation of the *unrestricted* model, possibly rescaled in various ways. Then, if s is an integer drawn at random from the set $1, \dots, n$, y_t^* is generated by

$$y_t^* = \boldsymbol{X}_{s1}\tilde{\boldsymbol{\beta}}_1 + \hat{u}_s,$$

where $\boldsymbol{\beta}_1$ contains the elements of $\boldsymbol{\beta}$ that are not in $\boldsymbol{\beta}_2$, and $\tilde{\boldsymbol{\beta}}_1$ is the *restricted* OLS estimate. Similarly, $\boldsymbol{X}_{s1}$ contains the elements of $\boldsymbol{X}_s$ of which the coefficients are elements of $\boldsymbol{\beta}_1$. By construction, the vector of the $\hat{u}_t$ is orthogonal to all of the vectors containing the observations of the explanatory variables. Thus in the empirical joint distribution of the pairs $(\hat{u}_t, \boldsymbol{X}_t)$, the first element, $\hat{u}$, is uncorrelated with the second element, $\boldsymbol{X}$. However any relation between the variance of $\hat{u}$ and the explanatory variables is preserved, as with Freedman's pairs bootstrap. In addition, the new bootstrap DGP now satisfies the null hypothesis as originally formulated.

Wild Bootstrap

The null model on which any form of pairs bootstrap is based posits the joint distribution of the dependent variable y and the explanatory variables. If it is assumed that the explanatory variables are exogenous, conventional practice is to compute statistics, and their distributions, conditional on them. One way in which this can be done is to use the so-called **wild bootstrap**; see Wu (1986), Liu (1988), Mammen (1993), and Davidson and Flachaire (2008). The rather odd name of this bootstrap procedure is due to Mammen, who says "We call this bootstrap procedure wild bootstrap because n different distributions are estimated by only n observations."

For a regression model, the wild bootstrap DGP takes the form

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + s_t^*\tilde{u}_t$$

where $\tilde{\boldsymbol{\beta}}$ is as usual the restricted least-squares estimate of the regression parameters, and the $\tilde{u}_t$ are the restricted least-squares residuals. Notice that no resampling takes place here; both the explanatory variables and the residual for bootstrap observation t come from observation t of the original sample. The new random elements introduced are the s_t^*, which are IID drawings from a distribution with expectation 0 and variance 1.

Copyright © 2021, Russell Davidson and James G. MacKinnon

This bootstrap DGP satisfies Golden Rule 1 easily: since s_t^* and $\tilde{u}_t$ are independent, the latter having been generated by the real DGP and the former by the random number generator, the expectation of the bootstrap disturbance $s_t^*\tilde{u}_t$ is 0. Conditional on the residual $\tilde{u}_t$, the variance of $s_t^*\tilde{u}_t$ is $\tilde{u}_t^2$. If the residual is accepted as a proxy for the unobserved disturbance u_t, then the unconditional expectation of $\tilde{u}_t^2$ is the true variance of u_t, and this fact goes a long way towards satisfying Golden Rule 2. The simplest HCCME, HC_0 uses exactly the same strategy to estimate the latent variances.

For a long time, the most commonly used distribution for the s_t^* was the following two-point distribution,

$$s_t^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}), \end{cases} \tag{7.13}$$

which was suggested by Mammen because, with it, $\mathrm{E}((s_t^*)^3) = 1$. If the true disturbances, and also the explanatory variables, are skewed, Mammen gives arguments designed to show that this is a desirable property for the accuracy of bootstrap inference. A simpler two-point distribution, proposed by Davidson and Flachaire, is the **Rademacher distribution**

$$s_t^* = \begin{cases} -1 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}. \end{cases} \tag{7.14}$$

Use of the Rademacher distribution leaves the absolute value of each residual unchanged in the bootstrap DGP, while assigning it an arbitrary sign. They show by a theoretical argument that this procedure is exact, up to the discreteness of the distribution for the special case in which the null hypothesis involves every one of the parameters in the regression, and show, by means of simulation experiments, that it yields inference in many cases better than other choices.

There is a good deal of evidence that the wild bootstrap works reasonably well for univariate regression models, even when there is quite severe heteroskedasticity. See, among others, Gonçalves and Kilian (2004) and MacKinnon (2006). Even when the disturbances are actually homoskedastic, the wild bootstrap often appears to perform as well as a comparable residual bootstrap method. The cost of insuring against heteroskedasticity generally seems to be very small. There seems to be no reason to use the pairs bootstrap when the only issue is potential heteroskedasticity.

Copyright © 2021, Russell Davidson and James G. MacKinnon

7.7 Autocorrelation

The bootstrap DGPs that we have discussed so far are not valid when applied to models with dependent disturbances having an unknown pattern of dependence. For such models, we wish to specify a bootstrap DGP which generates correlated disturbances that exhibit approximately the same pattern of dependence as the real disturbances, even though we do not know the process that actually generated them. There are two main approaches, neither of which is entirely satisfactory in all cases, unlike the case of the wild bootstrap for heteroskedasticity.

Block Bootstrap

The first principal method of dealing with dependent data is the **block bootstrap**, which was originally proposed by Künsch (1989). This method is by far the most widely used bootstrap in the presence of autocorrelation of unknown form. The idea is to divide the quantities that are being resampled, which might be either rescaled residuals or $[\boldsymbol{y}, \boldsymbol{X}]$ pairs, into blocks of l consecutive observations, and then resample the blocks. The blocks may be either overlapping or non-overlapping. In either case, the choice of block length, l, is evidently very important. If l is small, the bootstrap samples cannot possibly mimic non-trivial patterns of dependence in the original data, because these patterns are broken whenever one block ends and the next begins. However, if l is large, the bootstrap samples will tend to be excessively influenced by the random characteristics of the actual sample.

For the block bootstrap to work asymptotically, the block length must increase as the sample size n increases, but at a slower rate, which varies depending on what the bootstrap samples are to be used for. In some common cases, l should be proportional to $n^{1/3}$, but with a factor of proportionality that is, in practice, unknown. Unless the sample size is very large, it is generally impossible to find a value of l for which the bootstrap DGP provides a really good approximation to the unknown true DGP.

A variation of the block bootstrap is the **stationary bootstrap** proposed by Politis and Romano (1994), in which the block length is random rather than fixed. This procedure is commonly used in practice. However, Lahiri (1999) provides both theoretical arguments and limited simulation evidence which suggest that fixed block lengths are better than variable ones and that overlapping blocks are better than non-overlapping ones. Thus, at the present time, the procedure of choice appears to be the **moving-block bootstrap**, in which there are $n-l+1$ blocks, the first containing observations 1 through l, the second containing observations 2 through $l+1$, and the last containing observations $n-l+1$ through n.

It is possible to use block bootstrap methods with dynamic models. Let

$$\boldsymbol{Z}_t \equiv [y_t, y_{t-1}, \boldsymbol{X}_t].$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

For this model, we could construct $n-l+1$ overlapping blocks

$$\boldsymbol{Z}_1 \dots \boldsymbol{Z}_l, \quad \boldsymbol{Z}_2 \dots \boldsymbol{Z}_{l+1}, \dots\dots, \boldsymbol{Z}_{n-l+1} \dots \boldsymbol{Z}_n$$

and resample from them. This is the moving-block analog of the pairs bootstrap. When there are no exogenous variables and several lagged values of the dependent variable, the $\boldsymbol{Z}_t$ are themselves blocks of observations. Therefore, this method is sometimes referred to as the **block-of-blocks bootstrap**. Notice that, when the block size is 1, the block-of-blocks bootstrap is simply the pairs bootstrap adapted to dynamic models, as in Gonçalves and Kilian (2004).

Block bootstrap methods are conceptually simple. However, there are many different versions, most of which we have not discussed, and theoretical analysis of their properties tends to require advanced techniques. The biggest problem with block bootstrap methods is that they often do not work very well. We have already provided an intuitive explanation of why this is the case. From a theoretical perspective, the problem is that, even when the block bootstrap offers higher-order accuracy than asymptotic methods, it often does so to only a modest extent. The improvement is always of higher order in the independent case, where blocks should be of length 1, than in the dependent case, where the block size must be greater than 1 and must increase at an optimal rate with the sample size. See Hall, Horowitz, and Jing (1995) and Andrews (2004), among others.

There are several valuable, recent surveys of bootstrap methods for time-series data. These include Bühlmann (2002), Politis (2003), and Härdle, Horowitz, and Kreiss (2003). Surveys that are older or deal with methods for time-series data in less depth include Li and Maddala (1996), Davison and Hinkley (1997, Chapter 8), Berkowitz and Kilian (2000), Horowitz (2001), and Horowitz (2003).

Sieve Bootstrap

The second approach is a semiparametric one called the **sieve bootstrap**. The idea is to estimate a stationary autoregressive process of order p (AR(p)), and use this estimated process, perhaps together with resampled residuals from the estimation of the AR(p) model, to generate bootstrap samples.

An AR(p) process is a generalisation of the AR(1) process we saw in Section 4.2. Analogously to (4.14), a variable defined by an AR(p) process satisfies the equation

$$y_t = \rho_0 + \sum_{i=1}^{p} \rho_i y_{t-i} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{7.15}$$

where the ρ_i, $i = 0, 1, \dots, p$ are parameters that have to satisfy a rather complicated condition that need not trouble us here. We can see that (7.15)

Copyright © 2021, Russell Davidson and James G. MacKinnon

takes the form of a linear regression model, in which the only regressors are the constant and a set of p lags of the dependent variable.

Suppose we are concerned with a linear regression model, where the covariance matrix $\boldsymbol{\Omega}$ is no longer assumed to be diagonal. Instead, it is assumed that $\boldsymbol{\Omega}$ can be well approximated by the covariance matrix of a stationary AR(p) process, which implies that the diagonal elements are all the same.

In this case, the first step is to estimate the regression model, possibly after imposing restrictions on it, so as to generate a parameter vector $\hat{\boldsymbol{\beta}}$ and a vector of residuals $\hat{\boldsymbol{u}}$ with typical element $\hat{u}_t$. The next step is to estimate the AR(p) model

$$\hat{u}_t = \sum_{i=1}^{p} \rho_i \hat{u}_{t-i} + \varepsilon_t \tag{7.16}$$

for $t = p+1, \dots, n$. Note that, since the residuals sum to zero in most cases, there is no need for a constant in (7.16). In theory, the order p of this model should increase at a certain rate as the sample size increases. like the lag truncation parameter needed for a HAC covariance matrix; see Section 6.5.

Estimation of the AR(p) model yields residuals and an estimate $\hat{\sigma}_\varepsilon^2$ of the variance of the innovations ε_t, as well as the estimates $\hat{\rho}_i$. We may use these to set up a variety of possible bootstrap DGPs, all of which take the form

$$y_t^* = \boldsymbol{X}_t \hat{\boldsymbol{\beta}} + u_t^*.$$

There are two choices to be made, namely, the choice of parameter estimates $\hat{\boldsymbol{\beta}}$ and the generating process for the bootstrap disturbances u_t^*. The obvious choice for $\hat{\boldsymbol{\beta}}$ is just the (restricted) OLS estimates.

For observations after the first p, the bootstrap disturbances are generated as follows:

$$u_t^* = \sum_{i=1}^{p} \hat{\rho}_i u_{t-i}^* + \varepsilon_t^*, \quad t = p+1, \dots, n,$$

where the ε_t^* can either be drawn from the $\mathrm{N}(0, \hat{\sigma}_\varepsilon^2)$ distribution for a parametric bootstrap or resampled from the residuals $\hat{\varepsilon}_t$, preferably rescaled by the factor $\sqrt{n/(n-p)}$. First, of course, we must generate the first p bootstrap disturbances, the u_t^*, for $t = 1, \dots, p$.

The best way to do so is just to set $u_t^* = \hat{u}_t$ for the first p observations of each bootstrap sample. This is analogous to what we proposed in Section 7.4 for the bootstrap DGP used in conjunction with a dynamic model: We initialise with fixed starting values given by the real data.

The sieve bootstrap method has been used to improve the finite-sample properties of unit root tests by Park (2003) and Chang and Park (2003), but it has not yet been widely used in econometrics. The fact that it does not allow for heteroskedasticity is a limitation. Moreover, AR(p) processes do not provide good approximations to every time-series process that might arise in practice. For more detailed treatments, see Bühlmann (1997, 2002)), Choi and Hall (2000), and Park (2002).

Copyright © 2021, Russell Davidson and James G. MacKinnon

7.8 Bootstrap Confidence Sets

When exact confidence intervals are not available, and they generally are not, asymptotic ones are normally used. However, just as asymptotic tests do not always perform well in finite samples, neither do asymptotic confidence intervals. Since bootstrap P values and tests based on them often outperform their asymptotic counterparts, it seems natural to base confidence intervals on bootstrap tests when asymptotic intervals give poor coverage. There are a great many varieties of **bootstrap confidence intervals**; for a comprehensive discussion, see Davison and Hinkley (1997).

When we construct a bootstrap confidence interval, we wish to treat a family of tests, each corresponding to its own null hypothesis. Since, when we perform a bootstrap test, we must use a bootstrap DGP that satisfies the null hypothesis, it appears that we must use an infinite number of bootstrap DGPs if we are to consider the full family of tests, each with a different null. Fortunately, there is a clever trick that lets us avoid this difficulty.

It is, of course, essential for a bootstrap test that the bootstrap DGP should satisfy the null hypothesis under test. However, when the distribution of the test statistic does not depend on precisely which null is being tested, the same bootstrap distribution can be used for a whole family of tests with different nulls. If a family of test statistics is defined in terms of a pivotal random function $\tau(\boldsymbol{y}, \theta_0)$, then, by definition, the distribution of this function is independent of θ_0. Thus we could choose any value of θ_0 that the model allows for the bootstrap DGP, and the distribution of the test statistic, evaluated at θ_0, would always be the same. The important thing is to make sure that $\tau(\cdot)$ is evaluated at the *same* value of θ_0 as the one used to generate the bootstrap samples. Even if $\tau(\cdot)$ is only asymptotically pivotal, the effect of the choice of θ_0 on the distribution of the statistic should be slight if the sample size is reasonably large.

Suppose that we wish to construct a bootstrap confidence interval based on the t statistic $\hat{t}(\theta_0) \equiv \tau(\boldsymbol{y}, \theta_0) = (\hat{\theta} - \theta_0)/s_\theta$. The first step is to compute $\hat{\theta}$ and s_θ using the original data $\boldsymbol{y}$. Then we generate bootstrap samples using a DGP, which may be either parametric or based on resampling, characterized by $\hat{\theta}$ and by any other relevant estimates, such as the variance of disturbances, that may be needed. The resulting bootstrap DGP is thus quite independent of θ_0, but it does depend on the estimate $\hat{\theta}$.

We can now generate B bootstrap samples, $\boldsymbol{y}_b^*$, $b = 1, \dots, B$. For each of these, we compute an estimate θ_b^* and its standard error s_b^* in exactly the same way that we computed $\hat{\theta}$ and s_θ from the original data, and we then compute the bootstrap "t statistic"

$$t_b^* \equiv \tau(\boldsymbol{y}_b^*, \hat{\theta}) = \frac{\theta_b^* - \hat{\theta}}{s_b^*}. \tag{7.17}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

This is the statistic that tests the null hypothesis that $\theta = \hat{\theta}$, because $\hat{\theta}$ is the true value of θ for the bootstrap DGP. If $\tau(\cdot)$ is an exact pivot, the change of null from θ_0 to $\hat{\theta}$ makes no difference. If $\tau(\cdot)$ is an asymptotic pivot, there should usually be only a slight difference for values of θ_0 close to $\hat{\theta}$.

The limits of the bootstrap confidence interval depend on the quantiles of the EDF of the t_b^*. We can choose to construct either a symmetric confidence interval, by estimating a single critical value that applies to both tails, or an asymmetric one, by estimating two different critical values. When the distribution of the underlying test statistic $\tau(\boldsymbol{y}, \theta_0)$ is not symmetric, the latter interval should be more accurate.

P Values and Asymmetric Distributions

The above discussion of asymmetric confidence intervals raises the question of how to calculate P values for two-tailed tests based on statistics with asymmetric distributions. This rather tricky matter, which was treated briefly in Exercise 5.19, will turn out to be important when we discuss bootstrap confidence intervals in the next section.

If we denote by F the CDF used to calculate critical values or P values, the P value associated with a statistic τ should be $2F(\tau)$ if τ is in the lower tail, and $2\big(1 - F(\tau)\big)$ if it is in the upper tail, as seen in Exercise 5.19. In complete generality, we have that the P value is

$$p(\tau) = 2\min\big(F(\tau), 1 - F(\tau)\big). \tag{7.18}$$

A slight problem arises as to the point of separation between the left and the right sides of the distribution. This point is in fact the median, $q_{.50}$, for which $F(q_{.50}) = .50$ by definition, so that, if $\tau < q_{.50}$, the P value is $2F(\tau)$, and τ is consequently in the left-hand tail, while if $\tau > q_{.50}$, it is in the right-hand tail.

Asymmetric Bootstrap Confidence Intervals

Let us denote by $\hat{F}^*$ the EDF of the B bootstrap statistics t_b^*. For given θ_0, the bootstrap P value is, from (7.18),

$$\hat{p}\big(\hat{t}(\theta_0)\big) = 2\min\Big(\hat{F}^*\big(\hat{t}(\theta_0)\big),\ \ 1 - \hat{F}^*\big(\hat{t}(\theta_0)\big)\Big). \tag{7.19}$$

If this P value is greater than or equal to α, then θ_0 belongs to the $1 - \alpha$ confidence interval. If $\hat{F}^*$ were the CDF of a continuous distribution, we could express the confidence interval in terms of the quantiles of this distribution, just as in (6.13). In the limit as $B \to \infty$, the limiting distribution of the t_b^*, which we call the **ideal bootstrap distribution**, is usually continuous, and its quantiles define the ideal bootstrap confidence interval. However, since the distribution of the t_b^* is always discrete in practice, we must be a little more careful in our reasoning.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Suppose, to begin with, that $\hat{t}(\theta_0)$ is on the left side of the distribution. Then the bootstrap P value (7.19) is

$$2\hat{F}^*\big(\hat{t}(\theta_0)\big) = \frac{2}{B}\sum_{b=1}^{B} \mathbb{I}\big(t_b^* \le \hat{t}(\theta_0)\big) = \frac{2r(\theta_0)}{B},$$

where $r(\theta_0)$ is the number of bootstrap t statistics that are less than or equal to $\hat{t}(\theta_0)$. Thus θ_0 belongs to the equal-tail $1 - \alpha$ confidence interval if and only if $2r(\theta_0)/B \ge \alpha$, that is, if $r(\theta_0) \ge \alpha B/2$. Since $r(\theta_0)$ is an integer, while $\alpha B/2$ is not an integer, in general, this inequality is equivalent to $r(\theta_0) \ge r_{\alpha/2}$, where $r_{\alpha/2} = \lceil \alpha B/2 \rceil$ is the smallest integer not less than $\alpha B/2$, and $\lceil \cdot \rceil$ is the ceiling function we introduced in Section 7.2.

First, observe that $r(\theta_0)$ cannot exceed $r_{\alpha/2}$ for θ_0 sufficiently large. Since $\hat{t}(\theta_0) = (\hat{\theta} - \theta_0)/s_\theta$, it follows that $\hat{t}(\theta_0) \to -\infty$ as $\theta_0 \to \infty$. Accordingly, $r(\theta_0) \to 0$ as $\theta_0 \to \infty$. Therefore, there exists a greatest value of θ_0 for which $r(\theta_0) \ge r_{\alpha/2}$. This value must be the upper limit of the $1 - \alpha$ bootstrap confidence interval.

Suppose we sort the t_b^* from smallest to largest and denote by $c^*_{\alpha/2}$ the entry in the sorted list indexed by $r_{\alpha/2}$. Then, if $\hat{t}(\theta_0) = c^*_{\alpha/2}$, the number of the t_b^* less than or equal to $\hat{t}(\theta_0)$ is precisely $r_{\alpha/2}$. But if $\hat{t}(\theta_0)$ is smaller than $c^*_{\alpha/2}$ by however small an amount, this number is strictly less than $r_{\alpha/2}$. Thus θ_u, the upper limit of the confidence interval, is defined implicitly by $\hat{t}(\theta_u) = c^*_{\alpha/2}$. Explicitly, we have

$$\theta_u = \hat{\theta} - s_\theta c^*_{\alpha/2}.$$

As in the previous chapter, we see that the *upper* limit of the confidence interval is determined by the *lower* tail of the bootstrap distribution.

If the statistic is an exact pivot, then the probability that the true value of θ is greater than θ_u is exactly equal to $\alpha/2$ only if $\alpha(B + 1)/2$ is an integer. This follows by exactly the same argument as the one given in Section 7.3 for bootstrap P values. As an example, if $\alpha = .05$ and $B = 999$, we see that $\alpha(B + 1)/2 = 25$. In addition, since $\alpha B/2 = 24.975$, we see that $r_{\alpha/2} = 25$. The value of $c^*_{\alpha/2}$ is therefore the value of the 25^{th} bootstrap t statistic when they are sorted in ascending order.

In order to obtain the upper limit of the confidence interval, we began above with the assumption that $\hat{t}(\theta_0)$ is on the left side of the distribution. If we had begun by assuming that $\hat{t}(\theta_0)$ is on the right side of the distribution, we would have found that the lower limit of the confidence interval is

$$\theta_l = \hat{\theta} - s_\theta c^*_{1-(\alpha/2)},$$

where $c^*_{1-(\alpha/2)}$ is the entry indexed by $r_{1-(\alpha/2)}$ when the t_b^* are sorted in ascending order. For the example with $\alpha = .05$ and $B = 999$, this is the 975^{th} entry in the sorted list, since there are precisely 25 integers in the range $975 - 999$, just as there are in the range $1 - 25$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The asymmetric equal-tail bootstrap confidence interval can be written as

$$\big[\theta_l,\ \theta_u\big] = \big[\hat{\theta} - s_\theta c^*_{1-(\alpha/2)},\ \hat{\theta} - s_\theta c^*_{\alpha/2}\big]. \tag{7.20}$$

This interval bears a striking resemblance to the exact confidence interval (6.13). Clearly, $c^*_{1-(\alpha/2)}$ and $c^*_{\alpha/2}$, which are approximately the $1-(\alpha/2)$ and $\alpha/2$ quantiles of the EDF of the bootstrap tests, play the same roles as the $1-(\alpha/2)$ and $\alpha/2$ quantiles of the exact Student's t distribution.

Because Student's t distribution is symmetric, the confidence interval (6.13) is necessarily symmetric. In contrast, the interval (7.20) is almost never symmetric. Even if the distribution of the underlying test statistic happened to be symmetric, the bootstrap distribution based on finite B would almost never be. It is, of course, possible to construct a symmetric bootstrap confidence interval. We just need to invert a test for which the P value is not (7.18), but rather something like (5.08), which is based on the absolute value, or, equivalently, the square, of the t statistic.

The bootstrap confidence interval (7.20) is called a **studentized bootstrap** confidence interval. The name comes from the fact that a statistic is said to be **studentized** when it is the ratio of a random variable to its standard error, as is the ordinary t statistic. This type of confidence interval is also sometimes called a **percentile-*t*** or **bootstrap-*t*** confidence interval. Studentized bootstrap confidence intervals have good theoretical properties, and, as we have seen, they are quite easy to construct. If the assumptions of the classical normal linear model are violated and the empirical distribution of the t^*_b provides a better approximation to the actual distribution of the t statistic than does Student's t distribution, then the studentized bootstrap confidence interval should be more accurate than the usual interval based on asymptotic theory.

As we remarked above, there are a great many ways to compute bootstrap confidence intervals, and there is a good deal of controversy about the relative merits of different approaches. For an introduction to the voluminous literature, see DiCiccio and Efron (1996) and the associated discussion. Some of the approaches in the literature appear to be obsolete, mere relics of the way in which ideas about the bootstrap were developed, and others are too complicated to explain here. Even if we limit our attention to studentized bootstrap intervals, there are often several ways to proceed. Different ways of estimating standard errors inevitably lead to different confidence intervals, as do different ways of parametrizing a model. Thus, in practice, there is often quite a number of reasonable ways to construct studentized bootstrap confidence intervals.

Note that specifying the bootstrap DGP is not at all trivial if the disturbances are not assumed to be IID. In fact, this topic is quite advanced and has been the subject of much research: See Li and Maddala (1996) and Davison and Hinkley (1997), among others.

Theoretical results discussed in Hall (1992) and Davison and Hinkley (1997) suggest that studentized bootstrap confidence intervals generally work better

Copyright © 2021, Russell Davidson and James G. MacKinnon

than intervals based on asymptotic theory. However, their coverage can be quite unsatisfactory in finite samples if the quantity $(\hat{\theta} - \theta)/s_\theta$ is far from being pivotal, as can happen if the distributions of either $\hat{\theta}$ or s_θ depend strongly on the true unknown value of θ or on any other parameters of the model. When this is the case, the standard errors often fluctuate wildly among the bootstrap samples. Of course, the coverage of asymptotic confidence intervals is generally also unsatisfactory in such cases.

Asymptotic and Bootstrap Confidence Regions

When test statistics like (6.18), with known finite-sample distributions, are not available, the easiest way to construct an approximate confidence region is to base it on the Wald statistic (6.15), which can be used with any k–vector of parameter estimates $\hat{\boldsymbol{\theta}}$ that is root-n consistent and asymptotically normal and has a covariance matrix that can be consistently estimated by $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$. If c_α denotes the $1-\alpha$ quantile of the $\chi^2(k)$ distribution, then an approximate $1-\alpha$ confidence region is the set of all $\boldsymbol{\theta}_0$ such that

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \big(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\big)^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \le c_\alpha. \tag{7.21}$$

Like the exact confidence region defined by (6.19), this **asymptotic confidence region** is elliptical or ellipsoidal.

We can also use the statistic (6.15) to construct bootstrap confidence regions, making the same assumptions as were made above about $\hat{\boldsymbol{\theta}}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$. As we did for bootstrap confidence intervals, we use just one bootstrap DGP, either parametric or using resampling, characterized by the parameter vector $\hat{\boldsymbol{\theta}}$. For each of B bootstrap samples, indexed by j, we obtain a vector of parameter estimates $\boldsymbol{\theta}^*_b$ and an estimated covariance matrix $\mathrm{Var}^*(\boldsymbol{\theta}^*_b)$, in just the same way as $\hat{\boldsymbol{\theta}}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ were obtained from the original data. For each j, we compute the bootstrap "test statistic"

$$\tau^*_b \equiv (\boldsymbol{\theta}^*_b - \hat{\boldsymbol{\theta}})^\top \big(\mathrm{Var}^*(\boldsymbol{\theta}^*_b)\big)^{-1} (\boldsymbol{\theta}^*_b - \hat{\boldsymbol{\theta}}), \tag{7.22}$$

which is the multivariate analog of (7.17). We then find the bootstrap critical value c^*_α, which is the $1-\alpha$ quantile of the EDF of the τ^*_b. This is done by sorting the τ^*_b from smallest to largest and then taking the entry numbered $(B+1)(1-\alpha)$, assuming of course that $\alpha(B+1)$ is an integer. For example, if $B = 999$ and $\alpha = .05$, then c^*_α is the 950^{th} entry in the sorted list. The bootstrap confidence region is defined as the set of all $\boldsymbol{\theta}_0$ such that

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \big(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\big)^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \le c^*_\alpha. \tag{7.23}$$

It is no accident that the bootstrap confidence region defined by (7.23) looks very much like the asymptotic confidence region defined by (7.21). The only difference is that the critical value c_α, which appears on the right-hand side

Copyright © 2021, Russell Davidson and James G. MacKinnon

of (7.21), comes from the asymptotic distribution of the test statistic, while the critical value c^*_α, which appears on the right-hand side of (7.23), comes from the empirical distribution of the bootstrap samples. Both confidence regions have the same elliptical shape. When $c^*_\alpha > c_\alpha$, the region defined by (7.23) is larger than the region defined by (7.21), and the opposite is true when $c^*_\alpha < c_\alpha$.

Although this procedure is similar to the studentized bootstrap procedure discussed in Section 7.3, its true analog is the procedure for obtaining a *symmetric* bootstrap confidence interval that is the subject of Exercise 7.xxx. That procedure yields a symmetric interval because it is based on the square of the t statistic. Similarly, because this procedure is based on the quadratic form (6.15), the bootstrap confidence region defined by (7.23) is forced to have the same elliptical shape (but not the same size) as the asymptotic confidence region defined by (7.21). Of course, such a confidence region cannot be expected to work very well if the finite-sample distribution of $\hat{\boldsymbol{\theta}}$ does not in fact have contours that are approximately elliptical.

In view of the many ways in which bootstrap confidence intervals can be constructed, it should come as no surprise to learn that there are also many other ways to construct bootstrap confidence regions. See Davison and Hinkley (1997) for references and a discussion of some of these.

The bootstrap confidence interval for θ, (7.20), can also be transformed by g in order to obtain a bootstrap confidence interval for $\gamma \equiv g(\theta)$. The result is

$$\big[g(\hat\theta - s_\theta c^*_{1-(\alpha/2)}),\ \ g(\hat\theta - s_\theta c^*_{\alpha/2})\big], \tag{7.24}$$

where $c^*_{\alpha/2}$ and $c^*_{1-(\alpha/2)}$ are, as in (7.20), the entries indexed by $(\alpha/2)(B+1)$ and $(1-(\alpha/2))(B+1)$ in the sorted list of bootstrap t statistics t^*_b.

Yet another way to construct a bootstrap confidence interval is to bootstrap the t statistic for γ directly. Using the original data, we compute $\hat\theta$ and s_θ, and then $\hat\gamma$ and s_γ in terms of them. The bootstrap DGP is the same as the one used to obtain a bootstrap confidence interval for θ, but this time, for each bootstrap sample b, $b = 1, \ldots, B$, we compute γ^*_b and $(s_\gamma)^*_b$. The bootstrap "t statistics" $(\gamma^*_b - \hat\gamma)/(s_\gamma)^*_b$ are then sorted. If $(c_\gamma)^*_{\alpha/2}$ and $(c_\gamma)^*_{1-(\alpha/2)}$ denote the entries indexed by $(\alpha/2)(B+1)$ and $(1-(\alpha/2))(B+1)$ in the sorted list, then the (asymmetric) bootstrap confidence interval is

$$\big[\hat\gamma - s_\gamma(c_\gamma)^*_{1-(\alpha/2)},\ \ \hat\gamma - s_\gamma(c_\gamma)^*_{\alpha/2}\big]. \tag{7.25}$$

As readers are asked to check in Exercise 7.xxa, the intervals (7.24) and (7.25) are not the same.

Bootstrap Standard Errors

The delta method is not the only way to obtain standard errors and covariance matrices for functions of parameter estimates. The bootstrap can also be used for this purpose. Indeed, much of the early work on the bootstrap, such as Efron (1979), was largely concerned with bootstrap standard errors.

Suppose that, expanding on the work in the previous subsection, we wish to calculate the covariance matrix of the vector $\hat{\boldsymbol{\gamma}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}})$. A bootstrap procedure for doing this involves three steps:

1. Specify a bootstrap DGP, parametric or resampling, and use it to generate B bootstrap samples, $\boldsymbol{y}^*_b$.
2. For each bootstrap sample, use $\boldsymbol{y}^*_b$ to compute the parameter vector $\boldsymbol{\theta}^*_b$, and then use $\boldsymbol{\theta}^*_b$ to compute $\boldsymbol{\gamma}^*_b$.
3. Calculate $\bar{\boldsymbol{\gamma}}^*$, the mean of the $\boldsymbol{\gamma}^*_b$. Then calculate the estimated bootstrap covariance matrix,

$$\widehat{\mathrm{Var}}^*(\hat{\boldsymbol{\gamma}}) = \frac{1}{B-1}\sum_{b=1}^{B}(\boldsymbol{\gamma}^*_b - \bar{\boldsymbol{\gamma}}^*)(\boldsymbol{\gamma}^*_b - \bar{\boldsymbol{\gamma}}^*)^\top.$$

   If desired, bootstrap standard errors may be calculated as the square roots of the diagonal elements of this matrix.

Bootstrap standard errors, which may or may not be more accurate than ones based on asymptotic theory, can certainly be useful as descriptive statistics. However, using them for inference generally cannot be recommended. In many cases, calculating bootstrap standard errors is almost as much work as calculating studentized bootstrap confidence intervals. As we noted at the end of Section 7.3, there are theoretical reasons to believe that the latter yield more accurate inferences than confidence intervals based on asymptotic theory, including asymptotic confidence intervals that use bootstrap standard errors. Thus, if we are going to go to the trouble of calculating a large number of bootstrap estimates anyway, we can do better than just using them to compute bootstrap standard errors.

Copyright © 2021, Russell Davidson and James G. MacKinnon

7.9 Final Remarks

The bootstrap is a statistical technique capable of giving reliable inference for a wide variety of econometric models. In this chapter, the main focus is on inference based on the bootstrap. Although the bootstrap can be used for many other purposes, inference, in the form of hypothesis testing or of confidence sets, is the area in which use of the bootstrap has most clearly benefited econometric practice.

Although pivotal test statistics do arise from time to time, most test statistics in econometrics are not pivotal. The vast majority of them are, however, asymptotically pivotal. A statistic that is not an exact pivot cannot be used for a Monte Carlo test. However, approximate P values for statistics that are only asymptotically pivotal, or even non-pivotal, can still be obtained by

Copyright © 2021, Russell Davidson and James G. MacKinnon

bootstrapping. The difference between a Monte Carlo test and a bootstrap test is that for the former, the DGP is assumed to be known, whereas, for the latter, it is not. Unless the null hypothesis under test is a simple hypothesis, the DGP that generated the original data is unknown, and so it cannot be used to generate simulated data. The bootstrap DGP is an estimate of the unknown true DGP. The hope is that, if the bootstrap DGP is close, in some sense, to the true one, then data generated by the bootstrap DGP will be similar to data that would have been generated by the true DGP, if it were known. If so, then a simulated P value obtained by use of the bootstrap DGP is close enough to the true P value to allow accurate inference.

The actual implementation of a bootstrap test is identical to that of a Monte Carlo test. The only difference is that we do not (usually) just choose any convenient DGP in the null model, but rather one that can be considered a good estimate of the unknown true DGP.

Our theoretical understanding of the bootstrap is still incomplete. Many simulation experiments have shown that the bootstrap often performs much better than existing theories predict. Even so, there are some guidelines, here formulated more pretentiously as Golden Rules, that can help to ensure reliable bootstrap inference. These rules reflect the fact that, in inference, one wants as accurate a characterization as possible of the distribution, under the null hypothesis under test, of the test statistics on which inference is based.

7.10 Exercises

7.1 The file `bstats.txt` contains 999 bootstrap test statistics, sorted from smallest to largest and numbered for convenience. Use them to compute lower-tail, upper-tail, equal-tail, and symmetric bootstrap P values when the value of the actual test statistic is 2.197.

7.2 Suppose that we compute a bootstrap P value of 0.0603 using 199 bootstrap test statistics. If we could instead use an infinite number of bootstrap test statistics, we would obtain a bootstrap P value of p^*. Test the hypothesis that $p^* < 0.05$ against the alternative that $p^* \geq 0.05$.

7.3 Suppose the asymptotic distribution of a pivotal test statistic τ is $\mathrm{N}(0,1)$. In a sample of size n, the actual distribution is $\mathrm{N}(10/n, 1)$. What is the asymptotic P value for a two-tailed test based on the statistic $\hat{\tau} = -1.60$ when $n = 20$? Suppose you could perform an infinite number of bootstrap simulations. Then what would be the bootstrap P value based on the (incorrect) assumption that the distribution is symmetric around the origin? What would be the bootstrap P value without making any assumptions about the shape of the distribution? Based on these results, would you reject the null hypothesis at the .05 level? **Hint:** See Exercise 7.1.

7.4 The file **classical.data** contains 50 observations on three artificial variables, namely, $\boldsymbol{y}$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$. The data on $\boldsymbol{y}$ are generated by the classical linear regression model

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Test the hypothesis that $\sigma = 1.2$ at the .05 level. Also compute a P value for the test. **Hint:** See Exercise 4.19.

7.5 Using the data from the file **classical.data** again, estimate the regression model

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Compute a t statistic for the null hypothesis that $\beta_3 = 0$. On the basis of this test statistic, perform an exact test. Then perform parametric and semiparametric bootstrap tests using 99, 999, and 9,999 simulations. How do the two types of bootstrap P values correspond with the exact P value? How does this correspondence change as B increases?

7.6 If F is a strictly increasing CDF defined on an interval $[a, b]$ of the real line, where either or both of a and b may be infinite, then the inverse function F^{-1} is a well-defined mapping from $[0, 1]$ on to $[a, b]$. Show that, if the random variable X is a drawing from the $\mathrm{U}(0,1)$ distribution, then $F^{-1}(X)$ is a drawing from the distribution of which F is the CDF.

7.7 In Section 4.7, we saw that $\mathrm{Var}(\hat{u}_t) = (1 - h_t)\sigma_0^2$, where $\hat{u}_t$ is the t^{th} residual from the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, and h_t is the t^{th} diagonal element of the 'hat matrix' $\boldsymbol{P}_{\boldsymbol{X}}$; this was the result (4.58). Use this result to derive an alternative to (7.11) as a method of rescaling the residuals prior to resampling. Remember that the rescaled residuals must have mean 0.

7.8 This question uses data from the file **house-price-data.txt**, which contains 546 observations. Regress the logarithm of the house price on a constant, the logarithm of lot size, and the other ten explanatory variables, as in Exercise 6.15. One of the explanatory variables is the number of storeys, which can take on the values 1, 2, 3, and 4. Construct an unrestricted wild bootstrap .95 confidence interval for the difference in the expectation of the log price between a 3-storey house and a 2-storey house.

7.9 Consider again the data in the file **consumption.data** and the ADL model studied in Exercise 4.32, which is reproduced here for convenience:

$$c_t = \alpha + \beta c_{t-1} + \gamma_0 y_t + \gamma_1 y_{t-1} + u_t. \tag{4.90}$$

Compute a t statistic for the hypothesis that $\gamma_0 + \gamma_1 = 0$. On the basis of this test statistic, perform an asymptotic test, a parametric bootstrap test, and a resampling bootstrap test using residuals rescaled according to (7.11).

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 8

Instrumental Variables Estimation

8.1 Introduction

In Section 4.3 the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}$ was shown to be consistent under condition (4.13), according to which the expectation of the disturbance u_t associated with observation t is zero conditional on the regressors $\boldsymbol{X}_t$ for that same observation. As we saw in Section 5.5 this condition can also be expressed either by saying that the regressors $\boldsymbol{X}_t$ are predetermined or by saying that the disturbances u_t are innovations. When condition (4.13) does not hold, the consistency proof of Section 4.3 is not applicable, and the OLS estimator is in general both biased and inconsistent.

It is not always reasonable to assume that the disturbances are innovations. In fact, as we will see in the next section, there are commonly encountered situations in which the disturbances are necessarily correlated with some of the regressors for the same observation. Even in these circumstances, however, it is usually possible, although not always easy, to define an information set Ω_t for each observation such that

$$\mathrm{E}(u_t \,|\, \Omega_t) = 0. \tag{8.01}$$

Any regressor of which the value in period t is correlated with u_t cannot belong to Ω_t.

8.2 Correlation Between Disturbances and Regressors

We now briefly discuss two common situations in which the disturbances are correlated with the regressors and therefore do not have a zero expectation conditional on them. The first one, usually referred to by the name **errors in variables**, occurs whenever the independent variables in a regression model are measured with error. The second situation, often simply referred to as **simultaneity**, occurs whenever two or more endogenous variables are jointly determined by a system of simultaneous equations.

Errors in Variables

For a variety of reasons, many economic variables are measured with error. For example, macroeconomic time series are often based, in large part, on surveys, and they must therefore suffer from sampling variability. Whenever there are measurement errors, the values economists observe inevitably differ, to a greater or lesser extent, from the true values that economic agents presumably act upon. As we will see, measurement errors in the dependent variable of a regression model are generally of no great consequence, unless they are very large. However, measurement errors in the independent variables cause the disturbances to be correlated with the regressors that are measured with error, and this causes OLS to be inconsistent.

The problems caused by errors in variables can be seen quite clearly in the context of the simple linear regression model. Consider the model

$$y_t^\circ = \beta_1 + \beta_2 x_t^\circ + u_t^\circ, \quad u_t^\circ \sim \mathrm{IID}(0, \sigma^2), \tag{8.02}$$

where the variables x_t° and y_t° are not actually observed. We refer to them as **latent variables**. Instead, we observe

$$\begin{aligned} x_t &\equiv x_t^\circ + v_{1t}, \text{ and} \\ y_t &\equiv y_t^\circ + v_{2t}. \end{aligned} \tag{8.03}$$

Here v_{1t} and v_{2t} are measurement errors which are assumed, perhaps not realistically in some cases, to be IID with variances ω_1^2 and ω_2^2, respectively, and to be independent of x_t°, y_t°, and u_t°.

If we suppose that the true DGP is a special case of (8.02) along with (8.03), we see from (8.03) that $x_t^\circ = x_t - v_{1t}$ and $y_t^\circ = y_t - v_{2t}$. If we substitute these into (8.02), we find that

$$\begin{aligned} y_t &= \beta_1 + \beta_2(x_t - v_{1t}) + u_t^\circ + v_{2t} \\ &= \beta_1 + \beta_2 x_t + u_t^\circ + v_{2t} - \beta_2 v_{1t} \\ &= \beta_1 + \beta_2 x_t + u_t, \end{aligned} \tag{8.04}$$

where $u_t \equiv u_t^\circ + v_{2t} - \beta_2 v_{1t}$. Thus $\mathrm{Var}(u_t)$ is equal to $\sigma^2 + \omega_2^2 + \beta_2^2\omega_1^2$. The effect of the measurement error in the dependent variable is simply to increase the variance of the disturbances. Unless the increase is substantial, this is generally not a serious problem.

The measurement error in the independent variable also increases the variance of the disturbances, but it has another, much more severe, consequence as well. Because $x_t = x_t^\circ + v_{1t}$, and u_t depends on v_{1t}, u_t must be correlated with x_t whenever $\beta_2 \neq 0$. In fact, since the random part of x_t is v_{1t}, we see that

$$\mathrm{E}(u_t \,|\, x_t) = \mathrm{E}(u_t \,|\, v_{1t}) = -\beta_2 v_{1t}, \tag{8.05}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

because we assume that v_{1t} is independent of u_t° and v_{2t}. From (8.05), we can see, using the fact that $\mathrm{E}(u_t) = 0$ unconditionally, that

$$\begin{aligned}\mathrm{Cov}(x_t, u_t) &= \mathrm{E}(x_t u_t) = \mathrm{E}\big(x_t \mathrm{E}(u_t \mid x_t)\big) \\ &= -\mathrm{E}\big((x_t^\circ + v_{1t})\beta_2 v_{1t}\big) = -\beta_2 \omega_1^2.\end{aligned}$$

This covariance is negative if $\beta_2 > 0$ and positive if $\beta_2 < 0$, and, since it does not depend on the sample size n, it does not go away as n becomes large. An exactly similar argument shows that the assumption that $\mathrm{E}(u_t \mid \boldsymbol{X}_t) = 0$ is false whenever any element of $\boldsymbol{X}_t$ is measured with error. In consequence, the OLS estimator is biased and inconsistent.

Errors in variables are a potential problem whenever we try to estimate a consumption function, especially if we are using cross-section data. Many economic theories (for example, Friedman, 1957) suggest that household consumption depends on "permanent" income or "life-cycle" income, but surveys of household behavior almost never measure this. Instead, they typically provide somewhat inaccurate estimates of current income. If we think of y_t as measured consumption, x_t° as permanent income, and x_t as estimated current income, then the above analysis applies directly to the consumption function. The marginal propensity to consume is β_2, which must be positive, causing the correlation between u_t and x_t to be negative. As readers are asked to show in Exercise 8.1, the probability limit of $\hat{\beta}_2$ is less than the true value β_{20}. In consequence, the OLS estimator $\hat{\beta}_2$ is biased downward, even asymptotically.

Of course, if our objective is simply to estimate the relationship between the observed dependent variable y_t and the observed independent variable x_t, there is nothing wrong with using ordinary least squares to estimate equation (8.04). In that case, u_t would simply be *defined* as the difference between y_t and its expectation conditional on x_t. But our analysis shows that the OLS estimators of β_1 and β_2 in equation (8.04) are not consistent for the corresponding parameters of equation (8.02). In most cases, it is parameters like these that we want to estimate on the basis of economic theory.

There is an extensive literature on ways to avoid the inconsistency caused by errors in variables. See, among many others, Hausman and Watson (1985), Leamer (1987), and Dagenais and Dagenais (1997). The simplest and most widely-used approach is just to use an instrumental variables estimator.

Simultaneous Equations

Economic theory often suggests that two or more endogenous variables are determined simultaneously. In this situation, as we will see shortly, all of the endogenous variables must necessarily be correlated with the disturbances in all of the equations. This means that none of them may validly appear in the regression functions of models that are to be estimated by least squares.

A classic example, which well illustrates the econometric problems caused by simultaneity, is the determination of price and quantity for a commodity at

Copyright © 2021, Russell Davidson and James G. MacKinnon

the partial equilibrium of a competitive market. Suppose that q_t is quantity and p_t is price, both of which would often be in logarithms. A linear (or loglinear) model of demand and supply is

$$q_t = \gamma_d p_t + \boldsymbol{X}_t^d \boldsymbol{\beta}_d + u_t^d \tag{8.06}$$

$$q_t = \gamma_s p_t + \boldsymbol{X}_t^s \boldsymbol{\beta}_s + u_t^s, \tag{8.07}$$

where equation (8.06) is the demand function and equation (8.07) is the supply function. Here $\boldsymbol{X}_t^d$ and $\boldsymbol{X}_t^s$ are row vectors of observations on exogenous or predetermined variables that appear, respectively, in the demand and supply functions, $\boldsymbol{\beta}_d$ and $\boldsymbol{\beta}_s$ are corresponding vectors of parameters, γ_d and γ_s are scalar parameters, and u_t^d and u_t^s are the disturbances in the demand and supply functions. Economic theory predicts that, in most cases, $\gamma_d < 0$ and $\gamma_s > 0$, which is equivalent to saying that the demand curve slopes downward and the supply curve slopes upward.

Equations (8.06) and (8.07) are a pair of linear simultaneous equations for the two unknowns p_t and q_t. For that reason, these equations constitute what is called a **linear simultaneous equations model**. In this case, there are two dependent variables, quantity and price. For estimation purposes, the key feature of the model is that quantity depends on price in both equations.

Since there are two equations and two unknowns, it is straightforward to solve equations (8.06) and (8.07) for p_t and q_t. This is most easily done by rewriting them in matrix notation as

$$\begin{bmatrix} 1 & -\gamma_d \\ 1 & -\gamma_s \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t^d \boldsymbol{\beta}_d \\ \boldsymbol{X}_t^s \boldsymbol{\beta}_s \end{bmatrix} + \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix}. \tag{8.08}$$

The solution to (8.08), which exists whenever $\gamma_d \neq \gamma_s$, so that the matrix on the left-hand side of (8.08) is nonsingular, is

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} 1 & -\gamma_d \\ 1 & -\gamma_s \end{bmatrix}^{-1} \left(\begin{bmatrix} \boldsymbol{X}_t^d \boldsymbol{\beta}_d \\ \boldsymbol{X}_t^s \boldsymbol{\beta}_s \end{bmatrix} + \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix} \right). \tag{8.09}$$

It can be seen from this solution that p_t and q_t depend on both u_t^d and u_t^s, and on every exogenous and predetermined variable that appears in either the demand function, the supply function, or both. Therefore, p_t, which appears on the right-hand side of equations (8.06) and (8.07), must be correlated with the disturbances in both of those equations. If we rewrote one or both equations so that p_t was on the left-hand side and q_t was on the right-hand side, the problem would not go away, because q_t is also correlated with the disturbances in both equations.

It is easy to see that, whenever we have a linear simultaneous equations model, there must be correlation between all of the disturbances and all of the endogenous variables. If there are g endogenous variables and g equations, the

Copyright © 2021, Russell Davidson and James G. MacKinnon

solution looks very much like (8.09), with the inverse of a $g \times g$ matrix premultiplying the sum of a g–vector of linear combinations of the exogenous and predetermined variables and a g–vector of disturbances. If we want to estimate one equation out of such a system, the most popular approach is to use instrumental variables.

We have discussed two important situations in which the disturbances are necessarily correlated with some of the regressors, and the OLS estimator must consequently be inconsistent. This provides a strong motivation to employ estimators that do not suffer from this type of inconsistency. In the remainder of this chapter, we therefore discuss the method of instrumental variables. This method can be used whenever the disturbances are correlated with one or more of the explanatory variables, regardless of how that correlation may have arisen.

8.3 Instrumental Variables Estimation

For most of this chapter, we will focus on the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2\mathbf{I}, \tag{8.10}$$

where at least one of the explanatory variables in the $n \times k$ matrix $\boldsymbol{X}$ is assumed not to be predetermined with respect to the disturbances. Suppose that, for each $t = 1, \dots, n$, condition (8.01) is satisfied for some suitable information set Ω_t, and that we can form an $n \times k$ matrix $\boldsymbol{W}$ with typical row $\boldsymbol{W}_t$ such that all its elements belong to Ω_t. The k variables given by the k columns of $\boldsymbol{W}$ are called **instrumental variables**, or simply **instruments**. Later, we will allow for the possibility that the number of instruments may exceed the number of regressors.

Instrumental variables may be either exogenous or predetermined, and, for a reason that will be explained later, they should always include any columns of $\boldsymbol{X}$ that are exogenous or predetermined. Finding suitable instruments may be quite easy in some cases, but it can be extremely difficult in others. Many empirical controversies in economics are essentially disputes about whether or not certain variables constitute valid instruments.

The Simple IV Estimator

The condition (8.01) together with the requirement that $\boldsymbol{W}_t \in \Omega_t$ implies that

$$\mathrm{E}(u_t \,|\, \boldsymbol{W}_t) = 0. \tag{8.11}$$

Thus $\boldsymbol{W}_t^\top(y_t - \boldsymbol{X}_t\boldsymbol{\beta})$ is a a vector of elementary zero functions for the model (8.10). We can therefore construct a set of unbiased estimating equations:

$$\boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}. \tag{8.12}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Since there are k equations and k unknowns, we can solve equations (8.12) directly to obtain the **simple IV estimator**

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} \equiv (\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{W}^\top\boldsymbol{y}. \tag{8.13}$$

For identification by any given sample, it is necessary that $\boldsymbol{W}^\top\boldsymbol{X}$ should be nonsingular. If this condition were not satisfied, equations (8.12) would have no unique solution. In particular, this rules out any linear dependence in the columns of $\boldsymbol{W}$.

This well-known estimator has a long history (see Morgan, 1990). It is in general biased, just like the OLS estimate in a dynamic model with a set of regressors including lags of the dependent variable. It is however consistent under any sensible asymptotic construction like the "more of the same" construction recommended in Section 4.3. This leads us to make the assumption that

$$\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{X} \text{ is deterministic and nonsingular,} \tag{8.14}$$

which is certainly a consequence of any decent asymptotic construction.

It is easy to see directly that the simple IV estimator (8.13) is consistent, and, in so doing, to see that condition (8.11) can be weakened slightly. If the model (8.10) is correctly specified, with true parameter vector $\boldsymbol{\beta}_0$, then it follows that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{IV}} &= (\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\boldsymbol{\beta}_0 + (\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{W}^\top\boldsymbol{u} \\ &= \boldsymbol{\beta}_0 + (n^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}n^{-1}\boldsymbol{W}^\top\boldsymbol{u}. \end{aligned} \tag{8.15}$$

Given the assumption (8.14) of asymptotic identification, it is clear that $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ is consistent if and only if

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{u} = \mathbf{0}, \tag{8.16}$$

Although this follows directly from (8.11), it may hold in circumstances in which (8.11) is violated. We usually refer to the condition (8.16) by saying that the disturbances are **asymptotically uncorrelated** with the instruments. The weaker condition (8.16) is what is required for the consistency of the IV estimator.

In order to derive the asymptotic covariance matrix of $\boldsymbol{\beta}_{\mathrm{IV}}$, we make a further assumption,

$$\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W} \text{ is deterministic and nonsingular.} \tag{8.17}$$

If the model (8.10) is correctly specified with true parameter vector $\boldsymbol{\beta}_0$ and true disturbance variance σ_0^2, we see from (8.15) that

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}} - \boldsymbol{\beta}_0) = (n^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1}n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The factor $n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u}$ can be handled in exactly the same way as was $n^{-1/2}\boldsymbol{X}^\top\boldsymbol{u}$ in Section 5.5; see (5.48), (5.49), and (5.50). This leads to

$$n^{-1/2}\boldsymbol{W}^\top\boldsymbol{u} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}}),$$

Along with assumption (8.14), this lets us conclude that

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\big(\boldsymbol{0}, \sigma_0^2 (\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}})^{-1} \boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{W}} (\boldsymbol{S}_{\boldsymbol{W}^\top\boldsymbol{X}}^\top)^{-1}\big)$$

The limiting covariance matrix above is σ_0^2 times

$$\begin{aligned}
&\operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{W}^\top\boldsymbol{X})^{-1} \operatorname*{plim}_{n\to\infty} n^{-1}\boldsymbol{W}^\top\boldsymbol{W} \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{X}^\top\boldsymbol{W})^{-1} \\
&= \operatorname*{plim}_{n\to\infty} n(\boldsymbol{W}^\top\boldsymbol{X})^{-1}\boldsymbol{W}^\top\boldsymbol{W}(\boldsymbol{X}^\top\boldsymbol{W})^{-1} \\
&= \operatorname*{plim}_{n\to\infty} n\big[\boldsymbol{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{X}\big]^{-1} \\
&= \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})^{-1}.
\end{aligned}$$

The asymptotic covariance matrix of the simple IV estimator is thus

$$\sigma_0^2 \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})^{-1}. \tag{8.18}$$

The Generalized IV Estimator

In practice, the information set Ω_t is very frequently specified by providing a list of l instrumental variables that suggest themselves for various reasons. Therefore, we now drop the assumption that the number of instruments is equal to the number of parameters and let $\boldsymbol{W}$ denote an $n \times l$ matrix of instruments. Often, l is greater than k, the number of regressors in the model (8.10). In this case, the model is said to be **overidentified**, because, in general, there is more than one way to formulate estimating equations like (8.12) using the available instruments. If $l = k$, the model (8.10) is said to be **just identified** or **exactly identified**, because there is only one way to formulate the estimating equations. If $l < k$, it is said to be **underidentified**, because there are fewer estimating equations than parameters to be estimated, and equations (8.12) therefore have no unique solution.

If any instruments at all are available, it is normally possible to generate an arbitrarily large collection of them, because *any* deterministic nonlinear function of the l components of the t^{th} row $\boldsymbol{W}_t$ of $\boldsymbol{W}$ can be used as the t^{th} component of a new instrument.[1] If (8.10) is underidentified, some such

[1] This procedure would not work if, for example, all of the original instruments were binary variables.

Copyright © 2021, Russell Davidson and James G. MacKinnon

procedure is necessary if we wish to obtain consistent estimates of all the elements of $\boldsymbol{\beta}$. Alternatively, we would have to impose at least $k-l$ restrictions on $\boldsymbol{\beta}$ so as to reduce the number of independent parameters that must be estimated to no more than the number of instruments.

For models that are just identified or overidentified, it is often desirable to limit the set of potential instruments to deterministic *linear* functions of the instruments in $\boldsymbol{W}$, rather than allowing arbitrary deterministic functions. We will see shortly that this is not only reasonable but optimal for linear simultaneous equation models. With this restriction, the IV estimator is unique for a just identified model, because there is only one k–dimensional linear space $\mathcal{S}(\boldsymbol{W})$ that can be spanned by the $k = l$ instruments, and, as we saw earlier, the IV estimator for a given model depends only on the space spanned by the instruments.

We can always treat an overidentified model as if it were just identified by choosing exactly k linear combinations of the l columns of $\boldsymbol{W}$. The challenge is to choose these linear combinations optimally. Formally, we seek an $l \times k$ matrix $\boldsymbol{J}$ such that the $n \times k$ matrix $\boldsymbol{WJ}$ is a valid instrument matrix and such that, by using $\boldsymbol{J}$, the asymptotic covariance matrix of the estimator is minimized in the class of IV estimators which use an $n \times k$ matrix of instruments that are linear combinations of the columns of $\boldsymbol{W}$.

There are three requirements that the matrix $\boldsymbol{J}$ must satisfy. The first of these is that it should have full column rank of k. Otherwise, the space spanned by the columns of $\boldsymbol{WJ}$ would have dimension less than k, and the model would be underidentified. The second requirement is that $\boldsymbol{J}$ should be at least asymptotically deterministic. If not, it is possible that condition (8.16) applied to $\boldsymbol{WJ}$ could fail to hold. The last requirement is that $\boldsymbol{J}$ should be chosen to minimize the asymptotic covariance matrix of the resulting IV estimator, and we now explain how this may be achieved.

Efficiency Considerations

First of all, notice that, since (8.18) depends on $\boldsymbol{W}$ only through the orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{W}}$, all that matters is the space $\mathcal{S}(\boldsymbol{W})$ spanned by the instrumental variables. In fact, as readers are asked to show in Exercise 8.2, the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ itself depends on $\boldsymbol{W}$ only through $\boldsymbol{P}_{\boldsymbol{W}}$. This fact is closely related to the result that, for ordinary least squares, fitted values and residuals depend only on the space $\mathcal{S}(\boldsymbol{X})$ spanned by the regressors.

Suppose first that we are at liberty to choose for instruments any variables at all that satisfy the predeterminedness condition (8.11). Then, under reasonable and plausible conditions, we can characterize the **optimal instruments** for IV estimation of the model (8.10). By this, we mean the instruments that minimize the asymptotic covariance matrix (8.18), in the usual sense that any other choice of instruments leads to an asymptotic covariance matrix that differs from the optimal one by a positive semidefinite matrix.

Copyright © 2021, Russell Davidson and James G. MacKinnon

In order to determine the optimal instruments, we must know the data-generating process. In the context of a simultaneous equations model, a single equation like (8.10), even if we know the values of the parameters, cannot be a complete description of the DGP, because at least some of the variables in the matrix $\boldsymbol{X}$ are endogenous. For the DGP to be fully specified, we must know how all the endogenous variables are generated. For the demand-supply model given by equations (8.06) and (8.07), both of those equations are needed to specify the DGP. For a more complicated simultaneous equations model with g endogenous variables, we would need g equations. For the simple errors-in-variables model discussed in Section 8.2, we need equations (8.03) as well as equation (8.02) in order to specify the DGP fully.

Quite generally, we can suppose that the explanatory variables in the regression model (8.10) satisfy the relation

$$\boldsymbol{X} = \bar{\boldsymbol{X}} + \boldsymbol{V}, \qquad \mathrm{E}(\boldsymbol{V}_t \,|\, \Omega_t) = \boldsymbol{0}, \tag{8.19}$$

where the t^{th} row of $\bar{\boldsymbol{X}}$ is $\bar{\boldsymbol{X}}_t = \mathrm{E}(\boldsymbol{X}_t \,|\, \Omega_t)$, and $\boldsymbol{X}_t$ is the t^{th} row of $\boldsymbol{X}$. Thus equation (8.19) can be interpreted as saying that $\bar{\boldsymbol{X}}_t$ is the expectation of $\boldsymbol{X}_t$ conditional on the information set Ω_t. It turns out that the $n \times k$ matrix $\bar{\boldsymbol{X}}$ provides the optimal instruments for (8.10). Of course, in practice, $\bar{\boldsymbol{X}}$ is never observed, and it should be replaced by something that estimates it consistently.

To see that $\bar{\boldsymbol{X}}$ provides the optimal matrix of instruments, it is, as usual, easier to reason in terms of precision matrices rather than covariance matrices. For any valid choice of instruments, the precision matrix corresponding to (8.18) is $1/\sigma^2$ times

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} = \operatorname*{plim}_{n\to\infty} \big(n^{-1} \boldsymbol{X}^\top \boldsymbol{W} (n^{-1} \boldsymbol{W}^\top \boldsymbol{W})^{-1} n^{-1} \boldsymbol{W}^\top \boldsymbol{X}\big). \tag{8.20}$$

Using (8.19) and a law of large numbers, we see that

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{W} &= \lim_{n\to\infty} \frac{1}{n} \mathrm{E}(\boldsymbol{X}^\top \boldsymbol{W}) \\ &= \lim_{n\to\infty} \frac{1}{n} \mathrm{E}(\bar{\boldsymbol{X}}^\top \boldsymbol{W}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \bar{\boldsymbol{X}}^\top \boldsymbol{W}. \end{aligned} \tag{8.21}$$

The second equality holds because $\mathrm{E}(\boldsymbol{V}^\top \boldsymbol{W}) = \boldsymbol{O}$, since, by the construction in (8.19), $\boldsymbol{V}_t$ has zero expectation conditional on $\boldsymbol{W}_t$. The last equality is just an LLN in reverse. Similarly, we find that $\operatorname{plim} n^{-1} \boldsymbol{W}^\top \boldsymbol{X} = \operatorname{plim} n^{-1} \boldsymbol{W}^\top \bar{\boldsymbol{X}}$. Thus the precision matrix (8.20) becomes

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}. \tag{8.22}$$

If we make the choice $\boldsymbol{W} = \bar{\boldsymbol{X}}$, then (8.22) reduces to $\operatorname{plim} n^{-1} \bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}}$. The difference between this and (8.22) itself is just $\operatorname{plim} n^{-1} \bar{\boldsymbol{X}}^\top \boldsymbol{M}_{\boldsymbol{W}} \bar{\boldsymbol{X}}$, which is a

Copyright © 2021, Russell Davidson and James G. MacKinnon

positive semidefinite matrix. This shows that $\bar{\boldsymbol{X}}$ is indeed the optimal choice of instrumental variables by the criterion of asymptotic variance.

We mentioned earlier that all the explanatory variables in (8.10) that are exogenous or predetermined should be included in the matrix $\boldsymbol{W}$ of instrumental variables. It is now clear why this is so. If we denote by $\boldsymbol{Z}$ the submatrix of $\boldsymbol{X}$ containing the exogenous or predetermined variables, then $\bar{\boldsymbol{Z}} = \boldsymbol{Z}$, because the row $\boldsymbol{Z}_t$ is already contained in Ω_t. Thus $\boldsymbol{Z}$ is a submatrix of the matrix $\bar{\boldsymbol{X}}$ of optimal instruments. As such, it should always be a submatrix of the matrix of instruments $\boldsymbol{W}$ used for estimation, even if $\boldsymbol{W}$ is not actually equal to $\bar{\boldsymbol{X}}$.

Since the explanatory variables $\boldsymbol{X}$ satisfy (8.19), it follows from (8.18) and (8.21) that the asymptotic covariance matrix of the IV estimator computed using $\boldsymbol{WJ}$ as instrument matrix is

$$\sigma_0^2 \operatorname*{plim}_{n\to\infty} (n^{-1} \bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{WJ}} \bar{\boldsymbol{X}})^{-1}. \tag{8.23}$$

The t^{th} row $\bar{\boldsymbol{X}}_t$ of $\bar{\boldsymbol{X}}$ belongs to Ω_t by construction, and so each element of $\bar{\boldsymbol{X}}_t$ is a deterministic function of variables in the information set Ω_t. However, the deterministic functions are not necessarily linear functions of $\boldsymbol{W}_t$. Thus, in general, it is impossible to find a matrix $\boldsymbol{J}$ such that $\bar{\boldsymbol{X}} = \boldsymbol{WJ}$, as would be needed for $\boldsymbol{WJ}$ to constitute a set of truly optimal instruments. A natural second-best solution is to project $\bar{\boldsymbol{X}}$ orthogonally on to the space $\mathcal{S}(\boldsymbol{W})$. This yields the matrix of instruments

$$\boldsymbol{WJ} = \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}} = \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \bar{\boldsymbol{X}}, \tag{8.24}$$

which implies that

$$\boldsymbol{J} = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \bar{\boldsymbol{X}}. \tag{8.25}$$

We now show that these instruments are indeed optimal under the constraint that the instruments should be linear in $\boldsymbol{W}_t$.

By substituting $\boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}$ for $\boldsymbol{WJ}$ in (8.23), the asymptotic covariance matrix becomes

$$\sigma_0^2 \operatorname*{plim}_{n\to\infty} (n^{-1} \bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}} \bar{\boldsymbol{X}})^{-1}.$$

If we write out the projection matrix $\boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}}$ explicitly, we find that

$$\bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}} \bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}} (\bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}})^{-1} \bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}. \tag{8.26}$$

Thus, the precision matrix for the estimator that uses instruments $\boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}$ is proportional to $\bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{W}} \bar{\boldsymbol{X}}$. For the estimator with $\boldsymbol{WJ}$ as instruments, the precision matrix is proportional to $\bar{\boldsymbol{X}}^\top \boldsymbol{P}_{\boldsymbol{WJ}} \bar{\boldsymbol{X}}$. The difference between the two precision matrices is therefore proportional to

$$\bar{\boldsymbol{X}}^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{WJ}}) \bar{\boldsymbol{X}}. \tag{8.27}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The k–dimensional subspace $\mathcal{S}(\boldsymbol{WJ})$, which is the image of the orthogonal projection $\boldsymbol{P_{WJ}}$, is a subspace of the l–dimensional space $\mathcal{S}(\boldsymbol{W})$, which is the image of $\boldsymbol{P_W}$. Thus, by the result in Exercise 3.18, the difference $\boldsymbol{P_W} - \boldsymbol{P_{WJ}}$ is itself an orthogonal projection matrix. This implies that the difference (8.27) is a positive semidefinite matrix, and so we can conclude that (8.24) is indeed the optimal choice of instruments of the form $\boldsymbol{WJ}$.

At this point, we come up against the difficulty that the optimal instrument choice is infeasible, because we do not know $\bar{\boldsymbol{X}}$. But notice that, from the definition (8.25) of the matrix $\boldsymbol{J}$, we have that

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} \boldsymbol{J} &= \operatorname*{plim}_{n\to\infty} (n^{-1}\boldsymbol{W}^\top\boldsymbol{W})^{-1} n^{-1}\boldsymbol{W}^\top\bar{\boldsymbol{X}} \\ &= \operatorname*{plim}_{n\to\infty} (n^{-1}\boldsymbol{W}^\top\boldsymbol{W})^{-1} n^{-1}\boldsymbol{W}^\top\boldsymbol{X}, \end{aligned} \tag{8.28}$$

by (8.21). This suggests, correctly, that we can use $\boldsymbol{P_W X}$ instead of $\boldsymbol{P_W}\bar{\boldsymbol{X}}$ without changing the asymptotic properties of the estimator.

If we use $\boldsymbol{P_W X}$ as the matrix of instrumental variables, the estimating equations (8.12) that define the estimator become

$$\boldsymbol{X}^\top\boldsymbol{P_W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}, \tag{8.29}$$

which can be solved to yield the **generalized IV estimator**, or **GIV estimator**,

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\boldsymbol{X}^\top\boldsymbol{P_W X})^{-1}\boldsymbol{X}^\top\boldsymbol{P_W y}, \tag{8.30}$$

which is sometimes just abbreviated as **GIVE**. The estimator (8.30) is indeed a generalization of the simple estimator (8.13), as readers are asked to verify in Exercise 8.3. For this reason, we will usually refer to the IV estimator without distinguishing the simple from the generalized case.

The generalized IV estimator (8.30) can also be obtained by minimizing the **IV criterion function**, which has many properties in common with the sum of squared residuals for models estimated by least squares. This function is defined as follows:

$$Q(\boldsymbol{\beta}, \boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{P_W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{8.31}$$

Minimizing $Q(\boldsymbol{\beta}, \boldsymbol{y})$ with respect to $\boldsymbol{\beta}$ yields the estimator (8.30), as readers are asked to show in Exercise 8.4.

Identifiability and Consistency of the IV Estimator

It is clear from (8.30) that the generalized IV estimator needs the matrix $\boldsymbol{X}^\top\boldsymbol{P_W X}$ to be invertible. A condition of this sort is called an **identification condition**. For the OLS estimator, the identification condition is that $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible, and we saw that this is equivalent to the requirement of linear

Copyright © 2021, Russell Davidson and James G. MacKinnon

independence of the columns of $\boldsymbol{X}$. The analogous requirement for the IV estimator is that the columns of $\boldsymbol{P_W X}$ should be linearly independent, which implies that the estimating equations (8.29) have a unique solution. When this condition is satisfied by a given data set, we say that the parameters $\boldsymbol{\beta}$ are **identified** by that data set.

A different condition is needed if we want to show that $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ is consistent. It is an asymptotic counterpart to identification by a finite data set, and is called **asymptotic identification**. Consider what happens to the estimating functions in equations (8.29) as $n \to \infty$ with a sensible asymptotic construction. Under any DGP in the model (8.10), we define a vector of functions as follows:

$$\boldsymbol{\alpha}(\boldsymbol{\beta}) \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{P_W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{P_W X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}_0$ is the true parameter vector. Note that $\operatorname{plim}_{n\to\infty} n^{-1}\boldsymbol{X}^\top\boldsymbol{P_W u} = \boldsymbol{0}$, on account of (8.16). Clearly $\boldsymbol{\alpha}(\boldsymbol{\beta}_0) = \boldsymbol{0}$.

For asymptotic identification, we require the condition that $\boldsymbol{\alpha}(\boldsymbol{\beta}) \neq \boldsymbol{0}$ for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. For this condition to fail, there must exist $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ such that $\operatorname{plim}_{n\to\infty} n^{-1}\boldsymbol{X}^\top\boldsymbol{P_W X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = \boldsymbol{0}$. But this implies that $\operatorname{plim}_{n\to\infty} n^{-1}\boldsymbol{X}^\top\boldsymbol{P_W X}$ is singular. Consequently, what we require for asymptotic identification is that

$$\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{P_W X}} \equiv \operatorname*{plim}_{n\to\infty} n^{-1}\boldsymbol{X}^\top\boldsymbol{P_W X} \text{ is deterministic and nonsingular.}$$

Note that this does not necessarily follow from (8.14) and (8.17), since the former holds only for a just-identified model with $l = k$.

Asymptotic identification is sufficient for consistency. Because we are dealing here with linear models, there is no need for a sophisticated proof of this fact; see Exercise 8.6. The key assumption is, of course, (8.16). If this assumption did not hold, because any of the instruments was asymptotically correlated with the disturbances, the IV estimator would not be consistent.

Asymptotic Distribution of the IV Estimator

Like every estimator that we have studied, the IV estimator is asymptotically normally distributed with an asymptotic covariance matrix that can be estimated consistently. The asymptotic covariance matrix for the simple IV estimator, expression (8.18), turns out to be valid for the generalized IV estimator as well. To see this, we replace $\boldsymbol{W}$ in (8.18) by the asymptotically optimal instruments $\boldsymbol{P_W X}$. As in (8.26), we find that

$$\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{P_W X}}\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{P_W X}(\boldsymbol{X}^\top\boldsymbol{P_W X})^{-1}\boldsymbol{X}^\top\boldsymbol{P_W X} = \boldsymbol{X}^\top\boldsymbol{P_W X},$$

from which it follows that (8.18) is unchanged if $\boldsymbol{W}$ is replaced by $\boldsymbol{P_W X}$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

It can also be shown directly that (8.18) is the asymptotic covariance matrix of the generalized IV estimator. From (8.30), it follows that

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}_0) = (n^{-1}\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} n^{-1/2} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}. \tag{8.32}$$

Under reasonable assumptions, a central limit theorem can be applied to the expression $n^{-1/2}\boldsymbol{W}^\top \boldsymbol{u}$, which allows us to conclude that the asymptotic distribution of this expression is multivariate normal, with expectation zero and covariance matrix

$$\lim_{n\to\infty} \frac{1}{n} \boldsymbol{W}^\top \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) \boldsymbol{W} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}, \tag{8.33}$$

since we assume that $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma_0^2 \mathbf{I}$. With this result, it can be shown quite simply that (8.18) is the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{IV}}$; see Exercise 8.7.

In practice, since σ_0^2 is unknown, we use

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_{\text{IV}}) = \hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \tag{8.34}$$

to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{IV}}$. Here $\hat{\sigma}^2$ is $1/n$ times the sum of the squares of the components of the residual vector $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. In contrast to the OLS case, there is no good reason to divide by anything other than n when estimating σ^2. Because IV estimation minimizes the IV criterion function and not the sum of squared residuals, IV residuals are not necessarily too small. Nevertheless, many regression packages divide by $n - k$ instead of by n.

The choice of instruments usually affects the asymptotic covariance matrix of the IV estimator. If some or all of the columns of $\bar{\boldsymbol{X}}$ are not contained in the span $\mathcal{S}(\boldsymbol{W})$ of the instruments, an efficiency gain is potentially available if that span is made larger. Readers are asked in Exercise 8.8 to demonstrate formally that adding an extra instrument by appending a new column to $\boldsymbol{W}$ must, in general, reduce the asymptotic covariance matrix. Of course, it cannot be made smaller than the lower bound $\sigma_0^2(\bar{\boldsymbol{X}}^\top \bar{\boldsymbol{X}})^{-1}$, which is attained if the optimal instruments $\bar{\boldsymbol{X}}$ are available.

When all the regressors can validly be used as instruments, we have $\bar{\boldsymbol{X}} = \boldsymbol{X}$, and the efficient IV estimator coincides with the OLS estimator, as the Gauss-Markov Theorem predicts.

Two-Stage Least Squares

The IV estimator (8.30) is commonly known as the **two-stage least-squares**, or **2SLS**, estimator, because, before the days of good econometrics software packages, it was often calculated in two stages using OLS regressions. In the first stage, each column $\boldsymbol{x}_i$, $i = 1, \ldots, k$, of $\boldsymbol{X}$ is regressed on $\boldsymbol{W}$, if necessary. If a regressor $\boldsymbol{x}_i$ is a valid instrument, it is already (or should be) one of the columns of $\boldsymbol{W}$. In that case, since $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{x}_i = \boldsymbol{x}_i$, no first-stage regression is needed, and we say that such a regressor serves as its own instrument.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The fitted values from the first-stage regressions, plus the actual values of any regressors that serve as their own instruments, are collected to form the matrix $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}$. Then the second-stage regression,

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} \boldsymbol{\beta} + \text{residuals}, \tag{8.35}$$

is used to obtain the 2SLS estimates. Because $\boldsymbol{P}_{\boldsymbol{W}}$ is an idempotent matrix, the OLS estimate of $\boldsymbol{\beta}$ from this second-stage regression is

$$\hat{\boldsymbol{\beta}}_{\text{2sls}} = (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y},$$

which is identical to (8.30), the generalized IV estimator $\hat{\boldsymbol{\beta}}_{\text{IV}}$.

If this two-stage procedure is used, some care must be taken when estimating the standard error of the regression and the covariance matrix of the parameter estimates. The OLS estimate of σ^2 from regression (8.35) is

$$s^2 = \frac{\|\boldsymbol{y} - \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} \hat{\boldsymbol{\beta}}_{\text{IV}}\|^2}{n - k}. \tag{8.36}$$

In contrast, the estimate that was used in the estimated IV covariance matrix (8.34) is

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_{\text{IV}}\|^2}{n}. \tag{8.37}$$

These two estimates of σ^2 are not asymptotically equivalent, and s^2 is not consistent. The reason is that the residuals from regression (8.35) do not tend to the corresponding disturbances as $n \to \infty$, because the regressors in (8.35) are not the true explanatory variables. Therefore, $1/(n - k)$ times the sum of squared residuals is not a consistent estimator of σ^2. Of course, no regression package providing IV or 2SLS estimation would ever use (8.36) to estimate σ^2. Instead, it would use (8.37), or at least something that is asymptotically equivalent to it.

One clever way to get a consistent covariance matrix estimate is to run the artificial regression

$$\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_{\text{IV}} = \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} \boldsymbol{b} + \text{residuals}. \tag{8.38}$$

The regressand is just the vector of residuals from the IV estimation, and so it is orthogonal to the regressors on account of the estimating equations (8.29). The estimates of the artificial parameters $\boldsymbol{b}$ are thus all zero, and so the vector of residuals is the regressand unaltered, and the sum of squared residuals is the numerator of the consistent estimator (8.37). The estimated covariance matrix is therefore exactly (8.34), multiplied by $(n - k)/n$ if the regression package reports s^2 with a denominator of $n - k$.

Two-stage least squares was invented by Theil (1953) and Basmann (1957) at a time when computers were very primitive. Consequently, despite the

Copyright © 2021, Russell Davidson and James G. MacKinnon

classic papers of Durbin (1954) and Sargan (1958) on instrumental variables estimation, the term "two-stage least squares" came to be very widely used in econometrics, even when the estimator is not actually computed in two stages. We prefer to think of two-stage least squares as simply a particular way to compute the generalized IV estimator, and we will use $\hat{\boldsymbol{\beta}}_{\text{IV}}$ rather than $\hat{\boldsymbol{\beta}}_{\text{2sls}}$ to denote that estimator.

8.4 Finite-Sample Properties of IV Estimators

Unfortunately, the finite-sample distributions of IV estimators are much more complicated than the asymptotic ones. Indeed, except in very special cases, these distributions are unknowable in practice. Although it is consistent, the IV estimator for just identified models has a distribution with such thick tails that its expectation does not even exist. With overidentified models, the expectation of the estimator exists, but it is in general different from the true parameter value, so that the estimator is biased, often very substantially so. In consequence, investigators can easily make serious errors of inference when interpreting IV estimates.

The biases in the OLS estimates of a model like (8.10) arise because the disturbances are correlated with some of the regressors. The IV estimator solves this problem asymptotically, because the projections of the regressors on to $\mathcal{S}(\boldsymbol{W})$ are asymptotically uncorrelated with the disturbances. However, there must always still be some correlation in finite samples, and this causes the IV estimator to be biased.

Systems of Equations

In order to understand the finite-sample properties of the IV estimator, we need to consider the model (8.10) as part of a system of equations. We therefore change notation somewhat and rewrite (8.10) as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{Y}\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2\mathbf{I}, \tag{8.39}$$

where the matrix of regressors $\boldsymbol{X}$ has been partitioned into an $n \times k_1$ matrix of exogenous and predetermined variables, $\boldsymbol{Z}$, and an $n \times k_2$ matrix of endogenous variables, $\boldsymbol{Y}$, and the vector $\boldsymbol{\beta}$ has been partitioned conformably into two subvectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. There are assumed to be $l \geq k$ instruments, of which k_1 are the columns of the matrix $\boldsymbol{Z}$.

The model (8.39) is not fully specified, because it says nothing about how the matrix $\boldsymbol{Y}$ is generated. For each observation t, $t = 1, \ldots, n$, the value y_t of the dependent variable and the values $\boldsymbol{Y}_t$ of the other endogenous variables are assumed to be determined by a set of linear simultaneous equations. The variables in the matrix $\boldsymbol{Y}$ are called **current endogenous variables**, because they are determined simultaneously, row by row, along with $\boldsymbol{y}$. Suppose that

Copyright © 2021, Russell Davidson and James G. MacKinnon

all the exogenous and predetermined explanatory variables in the full set of simultaneous equations are included in the $n \times l$ instrument matrix $\boldsymbol{W}$, of which the first k_1 columns are those of $\boldsymbol{Z}$. Then, as can easily be seen by analogy with the explicit result (8.09) for the demand-supply model, we have for each endogenous variable $\boldsymbol{y}_i$, $i = 0, 1, \ldots, k_2$, that

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{\pi}_i + \boldsymbol{v}_i, \quad \mathrm{E}(v_{ti} \,|\, \boldsymbol{W}_t) = 0. \tag{8.40}$$

Here $\boldsymbol{y}_0 \equiv \boldsymbol{y}$, and the $\boldsymbol{y}_i$, for $i = 1, \ldots, k_2$, are the columns of $\boldsymbol{Y}$. The $\boldsymbol{\pi}_i$ are l–vectors of unknown coefficients, the $\boldsymbol{v}_i$ are n–vectors of disturbances that are innovations with respect to the instruments, v_{ti} is the t^{th} element of $\boldsymbol{v}_i$, and $\boldsymbol{W}_t$ is the t^{th} row of $\boldsymbol{W}$.

Equations like (8.40), which have only exogenous and predetermined variables on the right-hand side, are called **reduced-form equations**, in contrast with equations like (8.39), which are called **structural equations**. Writing a model as a set of reduced-form equations emphasizes the fact that all the endogenous variables are generated by similar mechanisms. In general, the disturbances for the various reduced-form equations display **contemporaneous correlation**: If v_{ti} denotes a typical element of the vector $\boldsymbol{v}_i$, then, for observation t, the reduced-form disturbances v_{ti} are generally correlated among themselves and correlated with the disturbance u_t of the structural equation.

A Simple Example

In order to gain additional intuition about the properties of the IV estimator in finite samples, we consider the very simplest nontrivial example, in which the dependent variable $\boldsymbol{y}$ is explained by only one variable, which we denote by $\boldsymbol{x}$. The regressor $\boldsymbol{x}$ is endogenous, and there is available exactly one exogenous instrument, $\boldsymbol{w}$. In order to keep the example reasonably simple, we suppose that all the disturbances, for both $\boldsymbol{y}$ and $\boldsymbol{x}$, are normally distributed. Thus the DGP that simultaneously determines $\boldsymbol{x}$ and $\boldsymbol{y}$ can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x}\beta_0 + \sigma_u\boldsymbol{u}, \\ \boldsymbol{x} &= \boldsymbol{w}\pi_0 + \sigma_v\boldsymbol{v}, \end{aligned} \tag{8.41}$$

where the second equation is analogous to (8.40). By explicitly writing σ_u and σ_v as the standard deviations of the disturbances, we can define the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ to be multivariate standard normal, that is, distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$. There is contemporaneous correlation of $\boldsymbol{u}$ and $\boldsymbol{v}$. Therefore, $\mathrm{E}(u_t v_t) = \rho$ for some correlation coefficient ρ such that $-1 < \rho < 1$. The result of Exercise 5.4 shows that the expectation of u_t conditional on v_t is ρv_t, and so we can write $\boldsymbol{u} = \rho\boldsymbol{v} + \boldsymbol{u}_1$, where $\boldsymbol{u}_1$ has expectation zero conditional on $\boldsymbol{v}$.

In this simple, just identified, setup, the IV estimator of the parameter β is

$$\hat{\beta}_{\text{IV}} = (\boldsymbol{w}^\top\boldsymbol{x})^{-1}\boldsymbol{w}^\top\boldsymbol{y} = \beta_0 + \sigma_u(\boldsymbol{w}^\top\boldsymbol{x})^{-1}\boldsymbol{w}^\top\boldsymbol{u}. \tag{8.42}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

This expression is clearly unchanged if the instrument $\boldsymbol{w}$ is multiplied by an arbitrary scalar, and so we can, without loss of generality, rescale $\boldsymbol{w}$ so that $\boldsymbol{w}^\top \boldsymbol{w} = 1$. Then, using the second equation in (8.41), we find that

$$\hat{\beta}_{\mathrm{IV}} - \beta_0 = \frac{\sigma_u \boldsymbol{w}^\top \boldsymbol{u}}{\pi_0 + \sigma_v \boldsymbol{w}^\top \boldsymbol{v}} = \frac{\sigma_u \boldsymbol{w}^\top (\rho \boldsymbol{v} + \boldsymbol{u}_1)}{\pi_0 + \sigma_v \boldsymbol{w}^\top \boldsymbol{v}}.$$

Let us now compute the expectation of this expression conditional on $\boldsymbol{v}$. Since, by construction, $\mathrm{E}(\boldsymbol{u}_1 \,|\, \boldsymbol{v}) = \boldsymbol{0}$, we obtain

$$\mathrm{E}(\hat{\beta}_{\mathrm{IV}} - \beta_0 \,|\, \boldsymbol{v}) = \frac{\rho \sigma_u}{\sigma_v} \frac{z}{a + z}, \tag{8.43}$$

where we have made the definitions $a \equiv \pi_0/\sigma_v$, and $z \equiv \boldsymbol{w}^\top \boldsymbol{v}$. Given our rescaling of $\boldsymbol{w}$, it is easy to see that $z \sim \mathrm{N}(0, 1)$.

When $\rho = 0$, the right-hand side of equation (8.43) vanishes, and so, conditional on $\boldsymbol{v}$, $\hat{\beta}_{\mathrm{IV}}$ is unbiased. In fact, since $\boldsymbol{v}$ is independent of $\boldsymbol{u}$ in this case, and $\boldsymbol{w}$ is exogenous, it follows that $\boldsymbol{x}$ is itself exogenous. With both $\boldsymbol{x}$ and $\boldsymbol{w}$ exogenous, the IV estimator is like the estimators dealt with in Exercise 4.25, which are unbiased conditional on these exogenous variables. If $\rho \neq 0$, however, $\boldsymbol{x}$ is not exogenous, and the estimator is biased conditional on $\boldsymbol{v}$. The *unconditional* expectation of the estimator does not even exist. To see this, let us try to calculate the expectation of the random variable $z/(a+z)$. If the expectation existed, it would be

$$\mathrm{E}\Big(\frac{z}{a+z}\Big) = \int_{-\infty}^{\infty} \frac{x}{a+x}\, \phi(x)\, dx, \tag{8.44}$$

where, as usual, $\phi(\cdot)$ is the density of the standard normal distribution. It is a fairly simple calculus exercise to show that the integral in (8.44) diverges in the neighborhood of $x = -a$.

If $\pi_0 = 0$, then $a = 0$. In this extreme case, the model is not asymptotically identified, and $\boldsymbol{x} = \sigma_v \boldsymbol{v}$ is just noise, as though it were a disturbance. As a consequence, $\boldsymbol{w}$ is not a valid instrument, and the IV estimator is inconsistent.

When $a \neq 0$, which is the usual case, the IV estimator (8.42) is neither biased nor unbiased, because it has no expectation for any finite sample size n. This may seem to contradict the result according to which $\hat{\beta}_{\mathrm{IV}}$ is asymptotically normal, since all the moments of the normal distribution exist. However, the fact that a sequence of random variables converges to a limiting random variable does not necessarily imply that the *moments* of the variables in the sequence converge to those of the limiting variable; see Davidson and MacKinnon (1993, Section 4.5). The estimator (8.42) is a case in point. Fortunately, this possible failure to converge of the moments does not extend to the CDFs of the random variables, which do indeed converge to that of the limit. Consequently, P values and the upper and lower limits of confidence

Copyright © 2021, Russell Davidson and James G. MacKinnon

intervals computed with the asymptotic distribution are legitimate approximations, in the sense that they become more and more accurate as the sample size increases.

A less simple calculation can be used to show that, in the overidentified case, the first $l - k$ moments of $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ exist; see Kinal (1980). This is consistent with the result we have just obtained for an exactly identified model, where $l - k = 0$, and the IV estimator has no moments at all. When the mean of $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ exists, it is almost never equal to $\boldsymbol{\beta}_0$. Readers will have a much clearer idea of the impact of the existence or nonexistence of moments, and of the bias of the IV estimator, if they work carefully through Exercises 8.10 to 8.13, in which they are asked to generate by simulation the EDFs of the estimator in different situations.

8.5 Hypothesis Testing

Because the finite-sample distributions of IV estimators are almost never known, exact tests of hypotheses based on such estimators are almost never available. However, large-sample tests can be performed in a variety of ways. Many of the methods of performing these tests are very similar to methods that we have already discussed in Chapter 5.

Asymptotic t and Wald Statistics

When there is just one restriction, the easiest approach is simply to compute an asymptotic t test. For example, if we wish to test the hypothesis that $\beta_i = \beta_{0i}$, where β_i is one of the regression parameters, then a suitable test statistic is

$$t_{\beta_i} = \frac{\hat{\beta}_i - \beta_{i0}}{\big(\widehat{\mathrm{Var}}(\hat{\beta}_i)\big)^{1/2}}, \tag{8.45}$$

where $\hat{\beta}_i$ is the IV estimate of β_i, and $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ is the i^{th} diagonal element of the estimated covariance matrix, (8.34). This test statistic does not follow Student's t distribution in finite samples, but it is asymptotically distributed as $\mathrm{N}(0, 1)$ under the null hypothesis.

For testing restrictions on two or more parameters, the natural analog of (8.45) is a Wald statistic. Suppose that $\boldsymbol{\beta}$ is partitioned as $[\boldsymbol{\beta}_1 \;\; \boldsymbol{\beta}_2]$, and we wish to test the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$. Then the appropriate Wald statistic is

$$W_{\boldsymbol{\beta}_2} = (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})^\top \big(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2)\big)^{-1} (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20}), \tag{8.46}$$

where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_2)$ is the submatrix of (8.34) that corresponds to the vector $\boldsymbol{\beta}_2$. This Wald statistic can be thought of as a generalization of the asymptotic t statistic: When $\boldsymbol{\beta}_2$ is a scalar, the square root of (8.46) is (8.45).

Copyright © 2021, Russell Davidson and James G. MacKinnon

Linear Restrictions

If the restrictions to be tested are all linear restrictions, there is no further loss of generality if we suppose that they are all zero restrictions. Thus the null and alternative hypotheses can be written as

$$H_0: \quad \boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}, \text{ and} \tag{8.47}$$

$$H_1: \quad \boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \tag{8.48}$$

where the matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are, respectively, $n \times k_1$ and $n \times k_2$, $\boldsymbol{\beta}_1$ is a k_1–vector, and $\boldsymbol{\beta}_2$ is a k_2–vector. As elsewhere in this chapter, it is assumed that $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma^2\mathbf{I}$. Any or all of the columns of $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ may be correlated with the disturbances. It is assumed that there exists an $n \times l$ matrix $\boldsymbol{W}$ of instruments, which are asymptotically uncorrelated with the disturbances, and that $l \geq k = k_1 + k_2$.

There is a very convenient was to implement a test of the null (8.47) against the alternative (8.48), described and justified in the following theorem.

Theorem 8.1.

A statistic that is a version of the Wald statistic $W_{\boldsymbol{\beta}_2}$ in (8.46), with $\boldsymbol{\beta}_{20} = \mathbf{0}$, is the explained sum of squares from the artificial regression

$$\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2\boldsymbol{b}_2 + \text{residuals}, \tag{8.49}$$

where $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are artificial parameter vectors, divided by any consistent estimator of the variance σ^2 of the disturbances.

Proof:

The proof proceeds in three stages. First, we derive an explicit expression for the Wald statistic. Second, we derive another explicit expression for the explained sum of squares from (8.49). Finally, we demonstrate that, except for possibly different estimators of σ^2, the two expressions are algebraically identical.

(1.) With IV estimation, we cannot use the FWL theorem directly to get a closed-form expression for $\hat{\boldsymbol{\beta}}_2$. However we can apply the theorem to the second-stage regression (8.35), since it is estimated by OLS. With the partition of $\boldsymbol{X}$, (8.35) becomes

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals},$$

and so the appropriate FWL regression is

$$\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals},$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

which gives the closed-form expression we seek:

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{y}. \tag{8.50}$$

By arguments similar to those that led to (8.18), we can see that the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_2$ is

$$\sigma^2 \operatorname*{plim}_{n\to\infty}(n^{-1}\boldsymbol{X}_2^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2)^{-1}. \tag{8.51}$$

By use of (8.50) and (8.51), the Wald statistic can be written as

$$\begin{aligned} W_{\boldsymbol{\beta}_2} &= \frac{1}{\hat{\sigma}^2}\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{y} \\ &= \frac{1}{\hat{\sigma}^2}\boldsymbol{y}^\top\boldsymbol{P}_{\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_2}\boldsymbol{y}, \end{aligned} \tag{8.52}$$

where $\hat{\sigma}^2$ is some consistent estimate of σ^2. Now

$$\boldsymbol{M}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{W}} = \boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}} = \boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1},$$

and so

$$\hat{\sigma}^2 W_{\boldsymbol{\beta}_2} = \boldsymbol{y}^\top\boldsymbol{P}_{(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1})\boldsymbol{X}_2}\boldsymbol{y}. \tag{8.53}$$

Note that the orthogonal projection matrix in the middle of the quadratic form above projects on to the k_2–dimensional space spanned by the columns of the matrix $(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1})\boldsymbol{X}_2$.

(2.) Recall that a consistent estimate of the covariance matrix of the IV estimator can be obtained by running the artificial regression (8.38), and, from that, we can extract a consistent estimate of σ^2. For the null hypothesis regression, (8.38) becomes

$$\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1\boldsymbol{b}_1 + \text{residuals}, \tag{8.54}$$

where $\tilde{\boldsymbol{\beta}}_1$ is the restricted estimator for H_0. By analogy with this, we construct the artificial regression (8.49) in which we append extra regressors to correspond to the columns of $\boldsymbol{X}_2$. The explained sum of squares from (8.49) is

$$(\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1)^\top\boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}}(\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1), \tag{8.55}$$

that is, the squared norm of the projection of the regressand $\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1$ on to the span of the regressors, that is, the columns of $\boldsymbol{P}_{\boldsymbol{W}}[\boldsymbol{X}_1 \ \ \boldsymbol{X}_2] = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}$.

Now $\tilde{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}$, and so

$$\begin{aligned} \boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1) &= (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{y} \\ &= (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}_1})\boldsymbol{y}. \end{aligned}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Thus expression (8.55) is

$$\boldsymbol{y}^\top(\boldsymbol{P_W} - \boldsymbol{P_{P_W X_1}})\boldsymbol{P_{P_W X}}(\boldsymbol{P_W} - \boldsymbol{P_{P_W X_1}})\boldsymbol{y} = \boldsymbol{y}^\top(\boldsymbol{P_{P_W X}} - \boldsymbol{P_{P_W X_1}})\boldsymbol{y} \quad (8.56)$$

because $\boldsymbol{P_W P_{P_W X}} = \boldsymbol{P_{P_W X}}$ and $\boldsymbol{P_W P_{P_W X_1}} = \boldsymbol{P_{P_W X_1}}$.

(3.) Although it is not obvious at first sight, we now show that the projection in (8.53) is equal to $\boldsymbol{P_{P_W X}} - \boldsymbol{P_{P_W X_1}}$, which is itself an orthogonal projection matrix on account of the result of Exercise 3.18. The dimensions are correct: we noted above that the projection in (8.53) projects on to a space of dimension k_2, and $\boldsymbol{P_{P_W X}} - \boldsymbol{P_{P_W X_1}}$ projects on to a space of dimension $k - k_1 = k_2$. For what we claim to be true, it is necessary and sufficient that

$$\boldsymbol{P_{P_W X}} = \boldsymbol{P_{P_W X_1}} + \boldsymbol{P}_{(\boldsymbol{P_W} - \boldsymbol{P_{P_W X_1}})\boldsymbol{X}_2}, \quad (8.57)$$

and, for this, it is enough to show that the images of the two projections on the right-hand side of (8.57) are orthogonal. We have

$$\begin{aligned} &\boldsymbol{X}_1^\top \boldsymbol{P_W}(\boldsymbol{P_W} - \boldsymbol{P_{P_W X_1}})\boldsymbol{X}_2 \\ &= \boldsymbol{X}_1^\top \boldsymbol{P_W X}_2 - \boldsymbol{X}_1^\top \boldsymbol{P_W X}_1(\boldsymbol{X}_1^\top \boldsymbol{P_W X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{P_W X}_2 \\ &= \boldsymbol{X}_1^\top \boldsymbol{P_W X}_2 - \boldsymbol{X}_1^\top \boldsymbol{P_W X}_2 = \boldsymbol{O}. \end{aligned}$$

We conclude that the quadratic forms in (8.53) and (8.56) are equal. ■

Remarks:

The total sum of squares from the artificial regression (8.49), divided by n, is the usual (consistent) estimator (8.37) of σ^2 from the IV estimation of the null-hypothesis model. Thus an admissible version of $W_{\boldsymbol{\beta}_2}$ is n times the ratio of the explained sum of squares to the total sum of squares from (8.49), which is just n times the uncentered R^2 from that regression. Alternatively, another asymptotically equivalent statistic is n times the ratio of the explained sum of squares to the sum of squared residuals. This is justified because, under the null hypothesis, the sum of squared residuals, being $O_p(n)$, dominates the explained sum of squares, which is $O_p(1)$, in the denominator of R^2, namely the total sum of squares.

The result that the explained sum of squares is $O_p(1)$ follows if we can show that (8.56) divided by n tends to zero in probability as $n \to \infty$. Observe first that

$$(\boldsymbol{P_{P_W X}} - \boldsymbol{P_{P_W X_1}})\boldsymbol{X}_1 = (\boldsymbol{P_{P_W X}} - \boldsymbol{P_{P_W X_1}})\boldsymbol{P_W X}_1 = \boldsymbol{O}. \quad (8.58)$$

This follows because $\boldsymbol{P_W X}_1$ is in the image of both $\boldsymbol{P_{P_W X}}$ and $\boldsymbol{P_{P_W X_1}}$ and so both terms in the middle expression in (8.58) are just equal to $\boldsymbol{P_W X}_1$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

If we now replace $\boldsymbol{y}$ in (8.56) by $\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}$, according to (8.47), the terms involving $\boldsymbol{X}_1$ vanish, so that (8.56) becomes

$$\boldsymbol{u}^\top(\boldsymbol{P_{P_W X}} - \boldsymbol{P_{P_W X_1}})\boldsymbol{u}. \quad (8.59)$$

Although $\boldsymbol{X}$ is random, we have seen that we can replace it asymptotically by $\bar{\boldsymbol{X}}$ when it occurs in the matrix product $\boldsymbol{W}^\top\boldsymbol{X}$, as is does in (8.59). Then, if we make the replacement, it follows by Theorem 5.1, part 2, that this expression is asymptotically equal to σ^2 times a $\chi^2(k_2)$ variable. When it is divided by n, the limit is zero.

The same matrix of instruments is assumed to be used for the estimation of both H_0 and H_1. While this assumption is natural if we start by estimating H_1 and then impose restrictions on it, it may not be so natural if we start by estimating H_0 and then estimate a less restricted model. A matrix of instruments that would be entirely appropriate for estimating H_0 may be inappropriate for estimating H_1, either because it omits some columns of $\boldsymbol{X}_2$ that are known to be uncorrelated with the disturbances, or because the number of instruments is greater than k_1 but less than $k_1 + k_2$. It is however essential for a test that the $\boldsymbol{W}$ matrix used should be appropriate and should be used for estimating H_1 as well as H_0.

A temptation that should be resisted is to compute an F statistic based on the SSRs obtained by IV estimation of (8.47) and (8.48). Such a "real" F statistic is not valid, even asymptotically. The problem is not with SSR_0, which is $\|\boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1\|^2$ both from the IV estimation of (8.47) and the OLS estimation of (8.54). But the sum of squared residuals from the IV estimation of (8.48) is the squared norm of the vector $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ (no need to partition $\boldsymbol{X}$ or $\boldsymbol{\beta}$). Since $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{P_W X})^{-1}\boldsymbol{X}^\top\boldsymbol{P_W y}$ by (8.30), this vector is

$$(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{P_W X})^{-1}\boldsymbol{X}^\top\boldsymbol{P_W})\boldsymbol{y}. \quad (8.60)$$

The $n \times n$ matrix in parentheses above can very easily be seem to be idempotent, but it is manifestly not symmetric. It is therefore an **oblique projection**. If we write $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{P_W X})^{-1}\boldsymbol{X}^\top\boldsymbol{P_W}$, it is clear that $\boldsymbol{P}$ projects on to $\mathcal{S}(\boldsymbol{X})$. However, $\mathbf{I} - \boldsymbol{P}$ does not project on to $\mathcal{S}^\perp(\boldsymbol{X})$. Rather, its image is $\mathcal{S}^\perp(\boldsymbol{P_W X})$; see Exercise 3.10. The obliqueness of the projection highlights a property of IV estimation, namely that the explained sum of squares and the sum of squared residuals do *not* add up to the total sum of squares. The Theorem relies on the fact that this property does hold for the artificial regressions (8.54) and (8.49), and so using the wrong SSR_1 gives an invalid F statistic.

Tests Based on Criterion Functions

The heart of the problem is that IV estimates are not obtained by minimizing the SSR, but rather the IV criterion function (8.31). The proper IV analog

Copyright © 2021, Russell Davidson and James G. MacKinnon

for the F statistic is a statistic based on the difference between the values of this criterion function evaluated at the restricted and unrestricted estimates. At the unrestricted estimates $\hat{\boldsymbol{\beta}}$, we obtain

$$Q(\hat{\boldsymbol{\beta}}, \boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{8.61}$$

Using the explicit expression (8.30) for the IV estimator, we see that (8.61) is equal to

$$\begin{aligned} &\boldsymbol{y}^\top (\mathbf{I} - \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top) \boldsymbol{P}_{\boldsymbol{W}} (\mathbf{I} - \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{y} \\ &\quad = \boldsymbol{y}^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}}) \boldsymbol{y} \\ &\quad = \boldsymbol{y}^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}}) \boldsymbol{y}. \end{aligned} \tag{8.62}$$

Observe that the presence of the factor of $\boldsymbol{P}_{\boldsymbol{W}}$ in the middle of this expression converts the oblique projection in (8.60) into an orthogonal projection. If Q is now evaluated at the restricted estimates $\tilde{\boldsymbol{\beta}}$, an exactly similar calculation shows that

$$Q(\tilde{\boldsymbol{\beta}}, \boldsymbol{y}) = \boldsymbol{y}^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_1}) \boldsymbol{y}. \tag{8.63}$$

The difference between (8.63) and (8.62) is thus

$$Q(\tilde{\boldsymbol{\beta}}, \boldsymbol{y}) - Q(\hat{\boldsymbol{\beta}}, \boldsymbol{y}) = \boldsymbol{y}^\top (\boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}_1}) \boldsymbol{y}. \tag{8.64}$$

This is precisely the expression (8.56) which is the numerator of the Wald statistic. Thus we can obtain an asymptotically correct test statistic by dividing (8.64) by any consistent estimate of σ^2.

The only practical difficulty in computing (8.64) is that some regression packages do not report the minimized value of the IV criterion function. However, this value is very easy to compute, since for any IV regression, restricted or unrestricted, it is equal to the explained sum of squares from a regression of the vector of IV residuals on the instruments $\boldsymbol{W}$, as can be seen at once from equation (8.61).

Heteroskedasticity and Autocorrelation Robust Tests

The test statistics discussed so far are valid only under the assumptions that the disturbances are serially uncorrelated and homoskedastic. If we are prepared to use an HCCME, the second of these assumptions can be relaxed; both if we use a HAC estimator. If $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is an $n \times n$ matrix, then it can readily be seen from equation (8.32) that the asymptotic covariance matrix of the vector $n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}} - \boldsymbol{\beta}_0)$ has the sandwich form

$$\Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}\Big)^{-1} \Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{\Omega} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}\Big) \Big(\operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}\Big)^{-1}. \tag{8.65}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Not surprisingly, this looks very much like expression (6.23) for OLS estimation, except that $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}$ replaces $\boldsymbol{X}$, and (8.65) involves probability limits rather than ordinary limits because the matrices $\boldsymbol{X}$, and possibly also $\boldsymbol{W}$, are now assumed to be stochastic.

Once again, the artificial regression is useful for computing an HCCME or a HAC estimator. For the basic model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, the artificial regression is given by (8.38), repeated here for convenience:

$$\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}\boldsymbol{b} + \text{residuals}.$$

Since the estimated artificial parameters $\boldsymbol{b}$ are zero, the IV residuals are also the residuals from this artificial regression. Then any version of an HCCME or HAC estimator computed from (8.38) uses the correct residuals for whatever version of the inconsistent estimate $\hat{\boldsymbol{\Omega}}$ is chosen. Thus a robust estimator from (8.38) takes the form

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}) \equiv (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{\Omega}} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1}, \tag{8.66}$$

and it is immediate that this matrix times the sample size n tends to the limit (8.65) as $n \to \infty$.

Once the matrix (8.66) has been calculated with an appropriate choice of $\hat{\boldsymbol{\Omega}}$, we can compute Wald tests that are robust to heteroskedasticity or to both heteroskedasticity and autocorrelation of unknown form. We simply use (8.45) for a test of a single linear restriction, or (8.46) for a test of two or more restrictions, with (8.66) replacing the ordinary covariance matrix estimator. Alternatively, a robust Wald test can be performed by any test of the artificial hypothesis that $\boldsymbol{b}_2 = \mathbf{0}$ in (8.49) that uses an HCCME or HAC estimator. Of course, it must be remembered that all these tests are based on asymptotic theory, and there is good reason to believe that this theory may often provide a poor guide to their performance in finite samples.

8.6 Testing Overidentifying Restrictions

The **degree of overidentification** of an overidentified linear regression model is defined to be $l - k$, where, as usual, l is the number of instruments, and k is the number of regressors. Such a model implicitly incorporates $l - k$ **overidentifying restrictions**. These arise because the generalized IV estimator implicitly uses only k **effective instruments**, namely, the k columns of $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}$. It does this because it is not possible, in general, to solve the l estimating equations (8.12) for only k unknowns.

In order for a set of instruments to be valid, a sufficient condition is (8.11), according to which the disturbance u_t has expectation zero conditional on $\boldsymbol{W}_t$, the l–vector of current instruments. When this condition is not satisfied, the

Copyright © 2021, Russell Davidson and James G. MacKinnon

IV estimator risks being inconsistent. But, if we use for estimation only the k effective instruments in the matrix $\boldsymbol{P_W X}$, it is only those k instruments that need to satisfy condition (8.11). Let $\boldsymbol{W}^*$ be an $n \times (l - k)$ matrix of **extra instruments** such that $\mathcal{S}(\boldsymbol{W}) = \mathcal{S}(\boldsymbol{P_W X}, \boldsymbol{W}^*)$. This means that the l–dimensional span of the full set of instruments is generated by linear combinations of the effective instruments, $\boldsymbol{P_W X}$, and the extra instruments, $\boldsymbol{W}^*$. The overidentifying restrictions require that the extra instruments should also satisfy (8.11). Unlike the conditions for the effective instruments, the overidentifying restrictions can, and always should, be tested.

The matrix $\boldsymbol{W}^*$ is not uniquely determined, but we will see in a moment that this does not matter. For any specific choice of $\boldsymbol{W}^*$, what we wish to test is the set of conditions

$$\mathrm{E}(\boldsymbol{W}_t^* u_t) = \boldsymbol{0}. \tag{8.67}$$

Although we do not observe the u_t, we can estimate the vector $\boldsymbol{u}$ by the vector of IV residuals $\hat{\boldsymbol{u}}$. Thus, in order to make our test operational, we form the sample analog of condition (8.67), which is

$$\frac{1}{n} (\boldsymbol{W}^*)^\top \hat{\boldsymbol{u}}, \tag{8.68}$$

and check whether this quantity is significantly different from zero.

The model we wish to test is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \quad \mathrm{E}(\boldsymbol{W}^\top \boldsymbol{u}) = \boldsymbol{0}. \tag{8.69}$$

Testing the overidentifying restrictions implicit in this model is equivalent to testing it against the alternative model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{W}^* \boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \quad \mathrm{E}(\boldsymbol{W}^\top \boldsymbol{u}) = \boldsymbol{0}. \tag{8.70}$$

This alternative model is constructed in such a way that it is just identified: There are precisely l coefficients to estimate, namely, the k elements of $\boldsymbol{\beta}$ and the $l - k$ elements of $\boldsymbol{\gamma}$, and there are precisely l instruments.

Let $\hat{\boldsymbol{u}}$ denote the residuals from the IV estimation of the null model (8.69). The artificial regression for testing the null against the alternative (8.70) is constructed just like (8.49):

$$\hat{\boldsymbol{u}} = \boldsymbol{P_W X} \boldsymbol{b}_1 + \boldsymbol{W}^* \boldsymbol{b}_2 + \text{residuals}. \tag{8.71}$$

As before, the numerator of the Wald statistic based on this artificial regression is given by (8.59), where what is written as $\boldsymbol{P_W X}$ there becomes, in our current notation, $[\boldsymbol{P_W X} \;\; \boldsymbol{W}^*]$, and $\boldsymbol{P_W X}_1$ there becomes $\boldsymbol{P_W X}$ here. Since $\mathcal{S}(\boldsymbol{W}) = \mathcal{S}(\boldsymbol{P_W X}, \boldsymbol{W}^*)$, the projection denoted $\boldsymbol{P}_{\boldsymbol{P_W X}}$ in (8.59) is simply $\boldsymbol{P_W}$, and the difference of the two projections in (8.59) becomes $\boldsymbol{P_W} - \boldsymbol{P}_{\boldsymbol{P_W X}}$.

One possible choice for $\boldsymbol{W}^*$ would be a matrix the columns of which were all orthogonal to those of $\boldsymbol{P_W X}$. Such a matrix could be constructed from an

Copyright © 2021, Russell Davidson and James G. MacKinnon

arbitrary $\boldsymbol{W}^*$ by multiplying it by $\boldsymbol{M}_{\boldsymbol{P_W X}}$. With such a choice, the orthogonality of $\boldsymbol{P_W X}$ and $\boldsymbol{W}^*$ means that, by the result in Exercise 3.18,

$$\boldsymbol{P_W} - \boldsymbol{P}_{\boldsymbol{P_W X}} = \boldsymbol{P}_{\boldsymbol{W}^*}.$$

With this choice for $\boldsymbol{W}^*$, (8.59) becomes $\boldsymbol{u}^\top \boldsymbol{P}_{\boldsymbol{W}^*} \boldsymbol{u}$, which makes it clear that the test of the null (8.69) against the alternative (8.70) tests whether $(\boldsymbol{W}^*)^\top \boldsymbol{u}$ is significantly different from zero, as we wanted, based on (8.68).

As we claimed above, implementing a test of the overidentifying restrictions does not require a specific choice of the matrix $\boldsymbol{W}^*$, and in fact it does not require us to construct $\boldsymbol{W}^*$ explicitly at all. This is because the explained sum of squares from (8.71) is the same as that from the regression

$$\hat{\boldsymbol{u}} = \boldsymbol{W}\boldsymbol{b} + \text{residuals}. \tag{8.72}$$

Hence the numerator of the Wald statistic is the explained sum of squares from (8.72), which is just $\hat{\boldsymbol{u}}^\top \boldsymbol{P_W} \hat{\boldsymbol{u}}$. The statistic itself is this divided by a consistent estimate of the variance of the disturbances. One such estimate is $n^{-1} \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}$, the usual estimate of $\hat{\sigma}^2$ from IV estimation. It is also, of course, the total sum of squares from (8.72), divided by n. Thus one way to compute the test statistic is to regress the residuals $\hat{\boldsymbol{u}}$ from IV estimation of the original model (8.69) on the full set of instruments, and use n times the uncentered R^2 from this regression as the test statistic. If the model (8.69) is correctly specified, the asymptotic distribution of the statistic is $\chi^2(l - k)$.

Another very easy way to test the overidentifying restrictions is to use a test statistic based on the IV criterion function. Since the alternative model (8.70) is just identified, the minimized IV criterion function for it is exactly zero. To see this, note that, for any just identified model, the IV residuals are orthogonal to the full set of instruments by the estimating equations (8.12) used with just identified models. Therefore, when the criterion function (8.31) is evaluated at the IV estimates $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$, it becomes $\hat{\boldsymbol{u}}^\top \boldsymbol{P_W} \hat{\boldsymbol{u}}$, as before. Thus an appropriate test statistic is just the criterion function $Q(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}, \boldsymbol{y})$ for the original model (8.69), divided by the estimate of the variance of the disturbances from this same model. A test based on this statistic is often called a **Sargan test**, after Sargan (1958). The test statistic is numerically identical to the one based on regression (8.72), as readers are asked to show in Exercise 8.16.

Although (8.70) is a simple enough model, it actually represents two conceptually different alternatives, because there are two situations in which the "true" parameter vector $\boldsymbol{\gamma}$ in (8.70) could be nonzero. One possibility is that the model (8.69) is correctly specified, but some of the instruments are asymptotically correlated with the disturbances and are therefore not valid instruments. The other possibility is that (8.69) is not correctly specified, and some of the instruments (or, possibly, other variables that are correlated with them) have incorrectly been omitted from the regression function. In either

Copyright © 2021, Russell Davidson and James G. MacKinnon

case, the overidentification test statistic leads us to reject the null hypothesis whenever the sample size is large enough.

Even if we do not know quite how to interpret a significant value of the overidentification test statistic, it is always a good idea to compute it. If it is significantly larger than it should be by chance under the null hypothesis, one should be extremely cautious in interpreting the estimates, because it is quite likely either that the model is specified incorrectly or that some of the instruments are invalid.

8.7 Durbin-Wu-Hausman Tests

In many cases, we do not know whether we actually need to use instrumental variables. For example, we may suspect that some variables are measured with error, but we may not know whether the errors are large enough to cause enough inconsistency for us to worry about. Or we may suspect that certain explanatory variables are endogenous, but we may not be at all sure of our suspicions, and we may not know how much inconsistency would result if they were justified. In such a case, it may or may not be perfectly reasonable to employ OLS estimation.

If the regressors are valid instruments, then, as we saw in Section 8.3, they are also the optimal instruments. Consequently, the OLS estimator, which is consistent in this case, is preferable to an IV estimator computed with some other valid instrument matrix $\boldsymbol{W}$. In view of this, it would evidently be very useful to be able to test the null hypothesis that the disturbances are uncorrelated with all the regressors against the alternative that they are correlated with some of the regressors, although not with the instruments $\boldsymbol{W}$. In this section, we discuss a simple procedure that can be used to perform such a test. This procedure dates back to a famous paper by Durbin (1954), and it was subsequently extended by Wu (1973) and Hausman (1978). We will therefore refer to all tests of this general type as **Durbin-Wu-Hausman tests**, or **DWH tests**.

The null and alternative hypotheses for the DWH test can be expressed as

$$H_0\colon \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad \text{E}(\boldsymbol{X}^\top\boldsymbol{u}) = \boldsymbol{0}, \text{ and} \tag{8.73}$$

$$H_1\colon \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad \text{E}(\boldsymbol{W}^\top\boldsymbol{u}) = \boldsymbol{0}. \tag{8.74}$$

Under H_1, the IV estimator $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is consistent, but the OLS estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is not. Under H_0, both are consistent. Thus, $\operatorname{plim}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}})$ is zero under the null and nonzero under the alternative. The idea of the DWH test is to check whether the difference $\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}}$ is significantly different from zero in the available sample. This difference, which is sometimes called the **vector of contrasts**, can be written as

$$\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}. \tag{8.75}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Expression (8.75) is not very useful as it stands, but it can be converted into a much more useful expression by means of a trick that is often useful in econometrics. We pretend that the first factor of $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is common to both estimators, and take it out as a common factor. This gives

$$\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})^{-1}\big(\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\big).$$

Now we can find some genuinely common factors in the two terms of the rightmost factor of this expression. Taking them out yields

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\boldsymbol{y} \\ &= (\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}. \end{aligned} \tag{8.76}$$

The first factor in expression (8.76) is a positive definite matrix, by the identification condition. Therefore, testing whether $\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}}$ is significantly different from zero is equivalent to testing whether the vector $\boldsymbol{X}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ is significantly different from zero.

Under H_0, the preferred estimation technique is OLS, and the OLS residuals are given by the vector $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$. Therefore, we wish to test whether the k columns of the matrix $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}$ are orthogonal to this vector of residuals. Let us partition the matrix of regressors $\boldsymbol{X} = [\boldsymbol{Z}\ \boldsymbol{Y}]$, where the k_1 columns of $\boldsymbol{Z}$ are included in the matrix of instruments $\boldsymbol{W}$, and the $k_2 = k - k_1$ columns of $\boldsymbol{Y}$ are treated as potentially endogenous. By construction, OLS residuals are orthogonal to all the columns of $\boldsymbol{X}$, in particular to those of $\boldsymbol{Z}$. For these regressors, there is therefore nothing to test: The relation

$$\boldsymbol{Z}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{Z}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{0}$$

holds identically, because $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Z} = \boldsymbol{Z}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Z} = \boldsymbol{O}$. The test is thus concerned only with the k_2 elements of $\boldsymbol{Y}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$, which are not in general identically zero, but should not differ from it significantly under H_0.

The easiest way to test whether $\boldsymbol{Y}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ is significantly different from zero is to use an F test for the k_2 restrictions $\boldsymbol{\delta} = \boldsymbol{0}$ in the OLS regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y}\boldsymbol{\delta} + \boldsymbol{u}. \tag{8.77}$$

The OLS estimates of $\boldsymbol{\delta}$ from (8.77) are, by the FWL Theorem, the same as those from the FWL regression of $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ on $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y}$, that is,

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{Y}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}.$$

Since the inverted matrix is positive definite, we see that testing whether $\boldsymbol{\delta} = \boldsymbol{0}$ is equivalent to testing whether $\boldsymbol{Y}^\top\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{0}$, as desired. This conclusion could have been foreseen by considering the threefold orthogonal decomposition that is implicitly performed by an F test; recall Section 5.4.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The DWH test can also be implemented by means of another F test, which yields exactly the same test statistic; see Exercise 8.17 for details.

The F test based on (8.77) has k_2 and $n-k-k_2$ degrees of freedom. Under H_0, if we assume that $\boldsymbol{X}$ and $\boldsymbol{W}$ are not merely predetermined but also exogenous, and that the disturbances $\boldsymbol{u}$ are multivariate normal, the F statistic does indeed have the $F(k_2, n-k-k_2)$ distribution. Under H_0 as it is expressed in (8.73), its asymptotic distribution is $F(k_2, \infty)$, and k_2 times the statistic is asymptotically distributed as $\chi^2(k_2)$.

If the null hypothesis (8.73) is rejected, we are faced with the same sort of ambiguity of interpretation as for the test of overidentifying restrictions. One possibility is that at least some columns of $\boldsymbol{Y}$ are indeed endogenous, but in such a way that the alternative model (8.74) is correctly specified. But we can equally well take (8.77) literally as a model with exogenous or predetermined regressors. In that case, the nature of the misspecification of (8.73) is not that $\boldsymbol{Y}$ is endogenous, but rather that the linear combinations of the instruments given by the columns of $\boldsymbol{P_W Y}$ have explanatory power for the dependent variable $\boldsymbol{y}$ over and above that of $\boldsymbol{X}$. Without further investigation, there is no way to choose between these alternative interpretations.

8.8 Bootstrap Tests

The difficulty with using the bootstrap for models estimated by IV is that there is more than one endogenous variable. The bootstrap DGP must therefore be formulated in such a way as to generate samples containing bootstrap realizations of both the main dependent variable $\boldsymbol{y}$ and the endogenous explanatory variables, which we denote by $\boldsymbol{Y}$.

As we saw in Section 8.4 the single equation (8.39) is not a complete specification of a model. We can complete it in various ways, of which the easiest is to use equations (8.40) for $i = 1, \ldots, k_2$. This introduces k_2 vectors $\boldsymbol{\pi}_i$, each containing l parameters. In addition, we must specify the *joint* distribution of the disturbances $\boldsymbol{u}$ in the equation for $\boldsymbol{y}$ and the $\boldsymbol{v}_i$ in the equations for $\boldsymbol{Y}$. We can write the reduced-form equations for the endogenous explanatory variables in matrix form as

$$\boldsymbol{Y} = \boldsymbol{W\Pi}_2 + \boldsymbol{V}_2, \tag{8.78}$$

where $\boldsymbol{\Pi}_2$ is an $l \times k_2$ matrix, the columns of which are the $\boldsymbol{\pi}_i$ of (8.40), and $\boldsymbol{V}_2$ is an $n \times k_2$ matrix of disturbances, the columns of which are the $\boldsymbol{v}_i$ of (8.40). It is convenient to group all the disturbances together into one matrix, and so we define the $n \times (k_2 + 1)$ matrix $\boldsymbol{V}$ as $[\boldsymbol{u} \;\; \boldsymbol{V}_2]$. If $\boldsymbol{V}_t$ denotes a typical row of $\boldsymbol{V}$, then we will assume that

$$\mathrm{E}(\boldsymbol{V}_t \boldsymbol{V}_t^\top) = \boldsymbol{\Sigma}, \tag{8.79}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $\boldsymbol{\Sigma}$ is a $(k_2+1) \times (k_2+1)$ covariance matrix, the upper left-hand element of which is σ^2, the variance of the disturbances in $\boldsymbol{u}$. Together, (8.39), (8.78), and (8.79) constitute a model that, although not quite fully specified (because the distribution of the disturbances is not given in full, only the first two moments), can serve as a basis for various bootstrap procedures.

Suppose that we wish to develop bootstrap versions of the tests considered in Section 8.5 where the null and alternative hypotheses are given by (8.47) and (8.48), respectively. For concreteness, we consider the test implemented by use of the artificial regression (8.49), although the same principles apply to other forms of test, such as the asymptotic t test (8.45), or tests based on the IV criterion function. Note that we now have two different partitions of the matrix $\boldsymbol{X}$ of explanatory variables. First, there is the partition $\boldsymbol{X} = [\boldsymbol{Z} \;\; \boldsymbol{Y}]$, in which $\boldsymbol{Z}$ contains the exogenous or predetermined variables, and $\boldsymbol{Y}$ contains the endogenous ones that are modeled explicitly by (8.78). Then there is the partition $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, in which we separate the variables $\boldsymbol{X}_1$ included under the null from the variables $\boldsymbol{X}_2$ that appear only under the alternative. In general, these two partitions are not related. We can expect that, in most cases, some columns of $\boldsymbol{Y}$ are contained in $\boldsymbol{X}_1$ and some in $\boldsymbol{X}_2$, and similarly for $\boldsymbol{Z}$.

The first step, as usual, is the estimation by IV of the model (8.47) that represents the null hypothesis. From this we obtain the constrained parameter estimates $\tilde{\boldsymbol{\beta}}_1$ and residuals $\tilde{\boldsymbol{u}}$. Next, we use $\tilde{\boldsymbol{\beta}}_1$ to formulate and run the artificial regression (8.49), and compute the statistic as n times the uncentered R^2. Then, in order to estimate all the other parameters of the extended model, we run the k_2 reduced-form regressions represented by (8.78), obtaining OLS estimates and residuals that we denote respectively by $\hat{\boldsymbol{\Pi}}_2$ and $\hat{\boldsymbol{V}}_2$. We will write $\hat{\boldsymbol{V}}$ to denote $[\tilde{\boldsymbol{u}} \;\; \hat{\boldsymbol{V}}_2]$.

For the bootstrap DGP, suppose first that all the instruments are exogenous. In that case, they are used unchanged in the bootstrap DGP. At this point, we must choose between a parametric and a resampling bootstrap. Since the latter is slightly easier, we discuss it first. In most cases, both $\boldsymbol{X}$ and $\boldsymbol{W}$ include a constant, and the residuals $\tilde{\boldsymbol{u}}$ and $\hat{\boldsymbol{V}}$ are centered. If not, as we discussed in Section 7.4, they must be centered before proceeding further. Because we wish the bootstrap DGP to retain the contemporaneous covariance structure of $\boldsymbol{V}$, the bootstrap disturbances are drawn as complete rows $\boldsymbol{V}_t^*$ by resampling entire rows of $\hat{\boldsymbol{V}}$, in a way analogous to what is done with the pairs bootstrap. In this way, we draw our bootstrap disturbances from the joint empirical distribution of the $\hat{\boldsymbol{V}}_t$. With models estimated by least squares, it is desirable to rescale residuals before they are resampled; again see Section 7.4. Since the columns of $\hat{\boldsymbol{V}}_2$ are least squares residuals, it is probably desirable to rescale them. However, there is no justification for rescaling the vector $\tilde{\boldsymbol{u}}$.

For the parametric bootstrap, we must actually estimate $\boldsymbol{\Sigma}$. The easiest way to do so is to form the matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \hat{\boldsymbol{V}}^\top \hat{\boldsymbol{V}}.$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Since $\tilde{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\Pi}}_2$ are consistent estimators, it follows that $\hat{\boldsymbol{V}}$ is also consistent for $\boldsymbol{V}$. We can then apply a law of large numbers to each element of $\hat{\boldsymbol{\Sigma}}$ in order to show that it converges as $n \to \infty$ to the corresponding element of the true $\boldsymbol{\Sigma}$. The row vectors of parametric bootstrap disturbances $\boldsymbol{V}_t^*$ are then independent drawings from the multivariate normal distribution with expectation zero and covariance matrix $\hat{\boldsymbol{\Sigma}}$. In order to make these drawings, the easiest method is to form a $(k_2+1)\times(k_2+1)$ matrix $\hat{\boldsymbol{A}}$ such that $\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^\top = \hat{\boldsymbol{\Sigma}}$. Usually, $\hat{\boldsymbol{A}}$ is chosen to be upper or lower triangular; recall the discussion of the multivariate normal distribution in Section 5.3. Then, if a random number generator is used to draw (k_2+1)–vectors $\boldsymbol{v}^*$ from $\mathrm{N}(\mathbf{0}, \mathbf{I})$, we see that $\hat{\boldsymbol{A}}\boldsymbol{v}^*$ is a drawing from $\mathrm{N}(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$, as desired.

The rest of the implementation is the same for both the parametric and the resampling bootstrap. For each bootstrap replication, the endogenous explanatory variables are first generated by the bootstrap reduced-form equations

$$\boldsymbol{Y}^* = \boldsymbol{W}\hat{\boldsymbol{\Pi}}_2 + \boldsymbol{V}_2^*, \tag{8.80}$$

where $\hat{\boldsymbol{\Pi}}_2$ and $\boldsymbol{V}_2^*$ are just the matrices $\hat{\boldsymbol{\Pi}}$ and $\boldsymbol{V}^*$ without their first columns. Then the main dependent variable is generated so as to satisfy the null hypothesis:

$$\boldsymbol{y}^* = \boldsymbol{X}_1^*\tilde{\boldsymbol{\beta}}_1 + \boldsymbol{u}^*.$$

Here the star on $\boldsymbol{X}_1^*$ indicates that some of the regressors in $\boldsymbol{X}_1$ may be endogenous, and so must have been simulated using (8.80). The bootstrap disturbances $\boldsymbol{u}^*$ are just the first column of $\boldsymbol{V}^*$. For each bootstrap sample, the testing artificial regression is run, and a bootstrap statistic is computed as n times the uncentered R^2. Then, as usual, the bootstrap P value is the proportion of bootstrap statistics greater than the statistic computed from the original data.

An alternative and probably preferable approach to the parametric and resampling bootstraps discussed above is a version of the **wild bootstrap**; see Section 7.6. Instead of resampling rows of $\hat{\boldsymbol{V}}$, we create the row vector of bootstrap disturbances for observation t by the formula $\boldsymbol{V}_t^* = s_t^*\hat{\boldsymbol{V}}_t$, where the s_t^* are IID drawings from a distribution with expectation zero and variance one, such as Mammen's two-point distribution (7.13) or the Rademacher distribution (7.14). This approach retains the contemporaneous covariance structure, and also allows for this structure to vary across observations; a phenomenon that extends the idea of heteroskedasticity to the multivariate case.

Bootstrap tests of overidentifying restrictions follow the same lines. Since the null hypothesis for such a test is just the model being estimated, the only extra work needed is the estimation of the reduced-form model (8.78) for the endogenous explanatory variables. Bootstrap disturbances are generated by a parametric, resampling, or wild bootstrap, and the residuals from the IV estimation using the bootstrap data are regressed on the full set of instruments. As usual, the simplest test statistic is the nR^2 from this regression.

Copyright © 2021, Russell Davidson and James G. MacKinnon

It is particularly easy to bootstrap DWH tests, because for them the null hypothesis is that none of the explanatory variables is endogenous. It is therefore quite unnecessary to model them by (8.78), and bootstrap data are generated as for any other model to be estimated by least squares. Note that, if we are prepared to make the strong assumptions of the classical normal linear model under the null, the bootstrap is quite unnecessary, because, as we saw in the previous section, the test statistic has a known finite-sample distribution.

If some of the non-endogenous explanatory variables are lagged dependent variables, or lags of the endogenous explanatory variables, bootstrap samples must be generated recursively, as for the case of the ordinary regression model with a lagged dependent variable, for which the recursive bootstrap DGP was (7.09). Especially if lags of endogenous explanatory variables are involved, this may become quite complicated.

It is worth issuing a warning that, for a number of reasons well beyond the scope of this chapter, the bootstrap methods outlined above cannot be expected to work as well as the bootstrap methods for regression models discussed in earlier chapters. Some reasons for this are discussed in Dufour (1997). Bootstrapping of simultaneous equations models has been and still is an active topic of research, and many new methods have been proposed. One of several versions of the wild bootstrap proposed in Davidson and MacKinnon (2010) appears particularly promising.

Wild Cluster Bootstrap

At this point, we can take up the problem, mentioned but not treated in the last chapter, of how to construct bootstrap DGPs in the presence of clustering. Even more so than with IV estimation, bootstrapping with clustering is an active research topic; see MacKinnon (2015) and MacKinnon and Webb (2016). It seems fair to say that, at this time of writing, there is no single approach that works effectively for all sets of clustered data, but one that does work well when there is a good number of rather small clusters is the **wild cluster bootstrap**. It uses exactly the same principle as the wild bootstrap for IV estimation. For the g^{th} cluster, $g = 1, \ldots, G$, the vector $\boldsymbol{u}_g^*$ of bootstrap disturbances is $s_g^*\hat{\boldsymbol{u}}_g$, with, as before, the s_g^* IID drawings from a distribution with expectation zero and variance one, and the $\hat{\boldsymbol{u}}_g$ are vectors of residuals for the observations in cluster g.

The wild cluster bootstrap preserves the covariance structure within clusters. Where it may fail to work well is when some clusters contain many observations, because then there is too little variability in the bootstrap samples to give rise to an adequate representation of the distribution implicit in the true DGP. Different issues, which we cannot treat here, arise as well when clusters are of widely different sizes.

Copyright © 2021, Russell Davidson and James G. MacKinnon

8.9 Final Remarks

Although it is formally very similar to other estimators that we have studied, the IV estimator does involve several important new concepts. These include the idea of an instrumental variable, the notion of forming a set of instruments optimally as weighted combinations of a larger number of instruments when that number exceeds the number of parameters, and the concept of overidentifying restrictions.

The optimality of the generalized IV estimator depends critically on the fairly strong assumption that the disturbances are homoskedastic and serially uncorrelated. When this assumption is relaxed, it may be possible to obtain estimators that are more efficient than the GIV estimator.

8.10 Exercises

8.1 Consider a very simple consumption function, of the form

$$c_i = \beta_1 + \beta_2 y_i^* + u_i^*, \quad u_i^* \sim \text{IID}(0, \sigma^2),$$

where c_i is the logarithm of consumption by household i, and y_i^* is the permanent income of household i, which is not observed. Instead, we observe current income y_i, which is equal to $y_i^* + v_i$, where $v_i \sim \text{IID}(0, \omega^2)$ is assumed to be uncorrelated with y_i^* and u_i^*. Therefore, we run the regression

$$c_i = \beta_1 + \beta_2 y_i + u_i.$$

Under the plausible assumption that the true value β_{20} is positive, show that y_i is negatively correlated with u_i. Using this result, evaluate the plim of the OLS estimator $\hat{\beta}_2$, and show that this plim is less than β_{20}.

8.2 Consider the simple IV estimator (8.13), computed first with an $n \times k$ matrix $\boldsymbol{W}$ of instrumental variables, and then with another $n \times k$ matrix $\boldsymbol{WJ}$, where $\boldsymbol{J}$ is a $k \times k$ nonsingular matrix. Show that the two estimators coincide. Why does this fact show that (8.13) depends on $\boldsymbol{W}$ only through the orthogonal projection matrix $\boldsymbol{P_W}$?

8.3 Show that, if the matrix of instrumental variables $\boldsymbol{W}$ is $n \times k$, with the same dimensions as the matrix $\boldsymbol{X}$ of explanatory variables, then the generalized IV estimator (8.30) is identical to the simple IV estimator (8.13).

8.4 Show that minimizing the criterion function (8.31) with respect to $\boldsymbol{\beta}$ yields the generalized IV estimator (8.30).

⋆**8.5** Under the usual assumptions of this chapter, including (8.16), show that the plim of

$$\frac{1}{n} Q(\boldsymbol{\beta}_0, \boldsymbol{y}) = \frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^\top \boldsymbol{P_W} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)$$

is zero if $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$. Under the same assumptions, along with the asymptotic identification condition that $\boldsymbol{S_{X^\top W}} (\boldsymbol{S_{W^\top W}})^{-1} \boldsymbol{S_{W^\top X}}$ has full rank, show further that $\operatorname{plim} n^{-1} Q(\boldsymbol{\beta}, \boldsymbol{y})$ is strictly positive for $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

8.6 Under assumption (8.16) and the asymptotic identification condition that $\boldsymbol{S_{X^\top W}} (\boldsymbol{S_{W^\top W}})^{-1} \boldsymbol{S_{W^\top X}}$ has full rank, show that the GIV estimator $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is consistent by explicitly computing the probability limit of the estimator for a DGP such that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$.

8.7 Suppose that you can apply a central limit theorem to the vector $n^{-1/2} \boldsymbol{W}^\top \boldsymbol{u}$, with the result that it is asymptotically multivariate normal, with expectation $\boldsymbol{0}$ and covariance matrix (8.33). Use equation (8.32) to demonstrate explicitly that, if $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$, then $n^{1/2}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}_0)$ is asymptotically normal with expectation $\boldsymbol{0}$ and covariance matrix (8.18).

8.8 Suppose that $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are, respectively, $n \times l_1$ and $n \times l_2$ matrices of instruments, and that $\boldsymbol{W}_2$ consists of $\boldsymbol{W}_1$ plus $l_2 - l_1$ additional columns. Prove that the generalized IV estimator using $\boldsymbol{W}_2$ is asymptotically more efficient than the generalized IV estimator using $\boldsymbol{W}_1$. To do this, you need to show that the matrix $(\boldsymbol{X}^\top \boldsymbol{P_{W_1}} \boldsymbol{X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{P_{W_2}} \boldsymbol{X})^{-1}$ is positive semidefinite. **Hint:** see Exercise 4.14.

8.9 Show that the simple IV estimator defined in (8.42) is unbiased when the data are generated by (8.41) with $\sigma_v = 0$. Interpret this result.

8.10 Use the DGP (8.41) to generate at least 1000 sets of simulated data for $\boldsymbol{x}$ and $\boldsymbol{y}$ with sample size $n = 10$, using normally distributed disturbances and parameter values $\sigma_u = \sigma_v = 1$, $\pi_0 = 1$, $\beta_0 = 0$, and $\rho = 0.5$. For the exogenous instrument $\boldsymbol{w}$, use independent drawings from the standard normal distribution, and then rescale $\boldsymbol{w}$ so that $\boldsymbol{w}^\top \boldsymbol{w}$ is equal to n, rather than 1 as in Section 10.4.

For each simulated data set, compute the IV estimator (8.42). Then draw the empirical distribution of the realizations of the estimator on the same plot as the CDF of the normal distribution with expectation zero and variance $\sigma_u^2 / n\pi_0^2$. Explain why this is an appropriate way to compare the finite-sample and asymptotic distributions of the estimator.

In addition, for each simulated data set, compute the OLS estimator, and plot the EDF of the realizations of this estimator on the same axes as the EDF of the realizations of the IV estimator.

8.11 Redo Exercise 8.10 for a sample size of $n = 100$. If you have enough computer time available, redo it yet again for $n = 1000$, in order to see how quickly or slowly the finite-sample distribution tends to the asymptotic distribution.

8.12 Redo the simulations of Exercise 8.10, for $n = 10$, generating the exogenous instrument $\boldsymbol{w}$ as follows. For the first experiment, use independent drawings from the uniform distribution on $[-1, 1]$. For the second, use drawings from the AR(1) process $w_t = \alpha w_{t-1} + \varepsilon_t$, where $w_0 = 0$, $\alpha = 0.8$, and the ε_t are independent drawings from $\text{N}(0, 1)$. In all cases, rescale $\boldsymbol{w}$ so that $\boldsymbol{w}^\top \boldsymbol{w} = n$. To what extent does the empirical distribution of $\hat{\beta}_{\text{IV}}$ appear to depend on the properties of $\boldsymbol{w}$? What theoretical explanation can you think of for your results?

8.13 Include one more instrument in the simulations of Exercise 8.10. Continue to use the same DGP for $\boldsymbol{y}$ and $\boldsymbol{x}$, but replace the simple IV estimator by the generalized one, based on two instruments $\boldsymbol{w}$ and $\boldsymbol{z}$, where $\boldsymbol{z}$ is generated independently of everything else in the simulation. See if you can verify the theoretical prediction that the overidentified estimator computed with two

Copyright © 2021, Russell Davidson and James G. MacKinnon

instruments is more biased, but has thinner tails, than the just identified estimator.

Repeat the simulations twice more, first with two additional instruments and then with four. What happens to the distribution of the estimator as the number of instruments increases?

8.14 Verify that $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ is the OLS estimator for model (8.10) when the regressor matrix is $\boldsymbol{X} = [\boldsymbol{Z} \ \boldsymbol{Y}] = \boldsymbol{W\Pi}$, with the matrix $\boldsymbol{V}$ in (8.81) equal to $\mathbf{O}$. Is this estimator consistent? Explain.

$$\begin{aligned} \boldsymbol{P_W X} &= [\boldsymbol{Z} \ \ \boldsymbol{P_W Y}] = [\boldsymbol{Z} \ \ \boldsymbol{P_W}(\boldsymbol{W\Pi}_2 + \boldsymbol{V}_2)] \\ &= [\boldsymbol{Z} \ \ \boldsymbol{W\Pi}_2 + \boldsymbol{P_W V}_2] = \boldsymbol{W\Pi} + \boldsymbol{P_W V}. \end{aligned} \tag{8.81}$$

8.15 Sketch a proof of the result that the scalar

$$\frac{1}{\sigma_0^2} \boldsymbol{u}^\top (\boldsymbol{P}_{\boldsymbol{P_W X}} - \boldsymbol{P}_{\boldsymbol{P_W X}_1}) \boldsymbol{u},$$

which is expression (8.59) divided by σ_0^2, is asymptotically distributed as $\chi^2(k_2)$ whenever the random vector $\boldsymbol{u}$ is $\mathrm{IID}(\mathbf{0}, \sigma_0^2 \mathbf{I})$ and is asymptotically uncorrelated with the instruments $\boldsymbol{W}$. Here $\boldsymbol{X}$ has k columns, $\boldsymbol{X}_1$ has k_1 columns, and $k_2 = k - k_1$.

8.16 Show that nR^2 from the artificial regression (8.72) is equal to the Sargan test statistic, that is, the minimized IV criterion function for model (8.69) divided by the IV estimate of the variance of the disturbances of that model.

8.17 Consider the following OLS regression, where the variables have the same interpretation as in Section 10.7 on DWH tests:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{M_W Y\zeta} + \boldsymbol{u}. \tag{8.82}$$

Show that an F test of the restrictions $\boldsymbol{\zeta} = \mathbf{0}$ in (8.82) is numerically identical to the F test for $\boldsymbol{\delta} = \mathbf{0}$ in (8.77). Show further that the OLS estimator of $\boldsymbol{\beta}$ from (8.82) is identical to the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ obtained by estimating (8.74) by instrumental variables.

8.18 Show that the difference between the generalized IV estimator $\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ and the OLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, for which an explicit expression is given in equation (8.76), has zero covariance with $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ itself. For simplicity, you may treat the matrix $\boldsymbol{X}$ as fixed.

8.19 The file **money.data** contains seasonally adjusted quarterly data for the logarithm of the real money supply, m_t, real GDP, y_t, and the 3-month Treasury Bill rate, r_t, for Canada for the period 1967:1 to 1998:4. Using these data, estimate the model

$$m_t = \beta_1 + \beta_2 r_t + \beta_3 y_t + \beta_4 m_{t-1} + \beta_5 m_{t-2} + u_t \tag{8.83}$$

by OLS for the period 1968:1 to 1998:4. Then perform a DWH test for the hypothesis that the interest rate, r_t, can be treated as exogenous, using r_{t-1} and r_{t-2} as additional instruments.

8.20 Estimate equation (8.83) by generalized instrumental variables, treating r_t as endogenous and using r_{t-1} and r_{t-2} as additional instruments. Are the

Copyright © 2021, Russell Davidson and James G. MacKinnon

estimates much different from the OLS ones? Verify that the IV estimates may also be obtained by OLS estimation of equation (8.82). Are the reported standard errors the same? Explain why or why not.

8.21 Perform a Sargan test of the overidentifying restrictions for the IV estimation you performed in Exercise 8.20. How do you interpret the results of this test?

8.22 The file **demand-supply.data** contains 120 artificial observations on a demand-supply model similar to equations (8.06)–(8.07). The demand equation is

$$q_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \gamma p_t + u_t, \tag{8.84}$$

where q_t is the log of quantity, p_t is the log of price, x_{t2} is the log of income, and x_{t3} is a dummy variable that accounts for regular demand shifts.

Estimate equation (8.84) by OLS and 2SLS, using the variables x_{t4} and x_{t5} as additional instruments. Does OLS estimation appear to be valid here? Does 2SLS estimation appear to be valid here? Perform whatever tests are appropriate to answer these questions.

Reverse the roles of q_t and p_t in equation (8.84) and estimate the new equation by OLS and 2SLS. How are the two estimates of the coefficient of q_t in the new equation related to the corresponding estimates of γ from the original equation? What do these results suggest about the validity of the OLS and 2SLS estimates?

Copyright © 2021, Russell Davidson and James G. MacKinnon

Chapter 9

Generalized Least Squares and Related Topics

9.1 Introduction

If the parameters of a regression model are to be estimated efficiently by least squares, the disturbances must be white noise, that is, be uncorrelated and have the same variance. These assumptions are needed to prove the Gauss-Markov Theorem. Moreover, the usual estimators of the covariance matrices of the OLS estimator are not valid when these assumptions do not hold, although alternative "sandwich" covariance matrix estimators that are asymptotically valid may be available (see Sections 6.4, 6.5, and 6.6). Thus it is clear that we need new estimation methods to handle regression models with disturbances that are heteroskedastic, serially correlated, or both. We develop some of these methods in this chapter.

We will consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{9.01}$$

where $\boldsymbol{\Omega}$, the covariance matrix of the disturbances, is a positive definite $n \times n$ matrix. If $\boldsymbol{\Omega}$ is equal to $\sigma^2\mathbf{I}$, then (9.01) is just the linear regression model (4.03), with white-noise disturbances. If $\boldsymbol{\Omega}$ is diagonal with nonconstant diagonal elements, then the disturbances are still uncorrelated, but they are heteroskedastic. If $\boldsymbol{\Omega}$ is not diagonal, then u_i and u_j are correlated whenever ω_{ij}, the ij^{th} element of $\boldsymbol{\Omega}$, is nonzero. In econometrics, covariance matrices that are not diagonal are most commonly encountered with time-series data, and the correlations are usually highest for observations that are close in time.

In the next section, we obtain an efficient estimator for the vector $\boldsymbol{\beta}$ in the model (9.01) by transforming the regression so that it satisfies the conditions of the Gauss-Markov theorem. This efficient estimator is called the **generalized least squares**, or **GLS**, estimator. Although it is easy to write down the GLS estimator, it is not always easy to compute it. In Section 9.3, we therefore discuss ways of computing GLS estimates, including the particularly simple case of weighted least squares. In the following section, we relax the often implausible assumption that the matrix $\boldsymbol{\Omega}$ is completely known. Section 9.5 discusses some aspects of heteroskedasticity. Sections 9.6 through 9.9 deal with various aspects of serial correlation, including autoregressive and moving-average processes, testing for serial correlation, GLS estimation of models with serially correlated disturbances, and specification tests for models with serially correlated disturbances. Finally, Section 9.10 discusses error-components models for panel data.

9.2 The GLS Estimator

In order to obtain an efficient estimator of the parameter vector $\boldsymbol{\beta}$ of the linear regression model (9.01), we transform the model so that the transformed model satisfies the conditions of the Gauss-Markov theorem. Estimating the transformed model by OLS therefore yields efficient estimates. The transformation is expressed in terms of an $n \times n$ matrix $\boldsymbol{\Psi}$, which is usually triangular, that satisfies the equation

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top. \tag{9.02}$$

As we discussed in Section 4.4, such a matrix can always be found, often by using Crout's algorithm.[1] Premultiplying (9.01) by $\boldsymbol{\Psi}^\top$ gives

$$\boldsymbol{\Psi}^\top\boldsymbol{y} = \boldsymbol{\Psi}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Psi}^\top\boldsymbol{u}. \tag{9.03}$$

Because the covariance matrix $\boldsymbol{\Omega}$ is nonsingular, the matrix $\boldsymbol{\Psi}$ must be as well, and so the transformed regression model (9.03) is perfectly equivalent to the original model (9.01). The OLS estimator of $\boldsymbol{\beta}$ from regression (9.03) is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{y}. \tag{9.04}$$

This estimator is called the **generalized least squares**, or **GLS**, estimator of $\boldsymbol{\beta}$.

It is not difficult to show that the covariance matrix of the transformed vector of disturbances $\boldsymbol{\Psi}^\top\boldsymbol{u}$ is simply the identity matrix:

$$\begin{aligned}\mathrm{E}(\boldsymbol{\Psi}^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{\Psi}) &= \boldsymbol{\Psi}^\top\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{\Psi} = \boldsymbol{\Psi}^\top\boldsymbol{\Omega}\boldsymbol{\Psi} \\ &= \boldsymbol{\Psi}^\top(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi} = \boldsymbol{\Psi}^\top(\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi} = \mathbf{I}.\end{aligned}$$

The second equality in the second line here uses a result about the inverse of a product of square matrices that was proved in Exercise 2.17.

[1] For computation, it is easier to use the algorithm for $\boldsymbol{\Omega}$, not $\boldsymbol{\Omega}^{-1}$, and invert the result to obtain $\boldsymbol{\Psi}$. Inverting a triangular matrix is numerically simpler than inverting a symmetric matrix.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Since $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is just the OLS estimator from (9.03), its covariance matrix can be found directly from the standard formula for the OLS covariance matrix, expression (4.38), if we replace $\boldsymbol{X}$ by $\boldsymbol{\Psi}^\top \boldsymbol{X}$ and σ_0^2 by 1:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = (\boldsymbol{X}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{X})^{-1} = (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1}. \tag{9.05}$$

In order for (9.05) to be valid, the conditions of the Gauss-Markov theorem must be satisfied. Here, this means that $\boldsymbol{\Omega}$ must be the covariance matrix of $\boldsymbol{u}$ *conditional* on the explanatory variables $\boldsymbol{X}$. It is thus permissible for $\boldsymbol{\Omega}$ to depend on $\boldsymbol{X}$, or indeed on any other exogenous variables.

The generalized least squares estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ can also be obtained by minimizing the **GLS criterion function**

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{9.06}$$

which is just the sum of squared residuals from the transformed regression (9.03). This criterion function can be thought of as a generalization of the SSR function in which the squares and cross products of the residuals from the original regression (9.01) are weighted by the inverse of the matrix $\boldsymbol{\Omega}$. The effect of such a weighting scheme is clearest when $\boldsymbol{\Omega}$ is a diagonal matrix: In that case, each observation is simply given a weight proportional to the inverse of the variance of its disturbance.

Efficiency of the GLS Estimator

The GLS estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ defined in (9.04) is also the solution of the set of estimating equations

$$\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{GLS}}) = \boldsymbol{0}. \tag{9.07}$$

These estimating equations are equivalent to the first-order conditions for the minimization of the GLS criterion function (9.06).

It is interesting to compare the GLS estimator with other estimators. A general estimator for the linear regression model (9.01) is defined in terms of an $n \times k$ matrix of exogenous variables $\boldsymbol{W}$, where k is the dimension of $\boldsymbol{\beta}$, by the equations

$$\boldsymbol{W}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}. \tag{9.08}$$

Since there are k equations and k unknowns, we can solve (9.08) to obtain the estimator

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} \equiv (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{y}. \tag{9.09}$$

The GLS estimator (9.04) is evidently a special case of this estimator, with $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}\boldsymbol{X}$.

Under certain assumptions, the estimator (9.09) is unbiased for the model (9.01). Suppose that the DGP is a special case of that model, with parameter vector $\boldsymbol{\beta}_0$ and known covariance matrix $\boldsymbol{\Omega}$. We assume that $\boldsymbol{X}$ and

Copyright © 2021, Russell Davidson and James G. MacKinnon

$\boldsymbol{W}$ are exogenous, which implies that $\text{E}(\boldsymbol{u} \,|\, \boldsymbol{X}, \boldsymbol{W}) = \boldsymbol{0}$. This rather strong assumption, which is analogous to the assumption (4.11), is necessary for the unbiasedness of $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ and makes it unnecessary to resort to asymptotic analysis. If we merely wanted to prove that $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ is consistent, we could, as in Section 8.3, get away with the much weaker assumption that $\text{E}(u_t \,|\, \boldsymbol{W}_t) = 0$, or, weaker still, that $\text{plim}_{n\to\infty} n^{-1}\boldsymbol{W}^\top \boldsymbol{u} = \boldsymbol{0}$; recall (8.16).

Substituting $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$ for $\boldsymbol{y}$ in (9.09), we see that

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} = \boldsymbol{\beta}_0 + (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{u}.$$

Therefore, the covariance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ is

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}) &= \text{E}\big((\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} - \boldsymbol{\beta}_0)^\top\big) \\ &= \text{E}\big((\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{W} (\boldsymbol{X}^\top \boldsymbol{W})^{-1}\big) \\ &= (\boldsymbol{W}^\top \boldsymbol{X})^{-1} \boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W} (\boldsymbol{X}^\top \boldsymbol{W})^{-1}. \end{aligned} \tag{9.10}$$

As we would expect, this is a sandwich covariance matrix. When $\boldsymbol{W} = \boldsymbol{X}$, we have the OLS estimator, and $\text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}})$ reduces to expression (6.22).

The efficiency of the GLS estimator can be verified by showing that the difference between (9.10), the covariance matrix for the estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ defined in (9.09), and (9.05), the covariance matrix for the GLS estimator, is a positive semidefinite matrix. As was shown in Exercise 4.14, this difference is positive semidefinite if and only if the difference between the inverse of (9.05) and the inverse of (9.10), that is, the matrix

$$\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X} - \boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{X}, \tag{9.11}$$

is positive semidefinite. In Exercise 9.2, readers are invited to show that this is indeed the case.

The GLS estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is typically more efficient than the more general estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ for all elements of $\boldsymbol{\beta}$, because it is only in very special cases that the matrix (9.11) has any zero diagonal elements. Because the OLS estimator $\hat{\boldsymbol{\beta}}$ is just $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ when $\boldsymbol{W} = \boldsymbol{X}$, we conclude that the GLS estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ in most cases is more efficient, and is never less efficient, than the OLS estimator $\hat{\boldsymbol{\beta}}$.

9.3 Computing GLS Estimates

At first glance, the formula (9.04) for the GLS estimator seems quite simple. To calculate $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ when $\boldsymbol{\Omega}$ is known, we apparently just have to invert $\boldsymbol{\Omega}$, form the matrix $\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X}$ and invert it, then form the vector $\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{y}$, and, finally, postmultiply the inverse of $\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X}$ by $\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{y}$. However, GLS

Copyright © 2021, Russell Davidson and James G. MacKinnon

estimation is not nearly as easy as it looks. The procedure just described may work acceptably when the sample size n is small, but it rapidly becomes computationally infeasible as n becomes large. The problem is that $\boldsymbol{\Omega}$ is an $n \times n$ matrix. When $n = 1000$, simply storing $\boldsymbol{\Omega}$ and its inverse typically requires 16 MB of memory; when $n = 10{,}000$, storing both these matrices requires 1600 MB. Even if enough memory were available, computing GLS estimates in this naive way would be enormously expensive.

Practical procedures for GLS estimation require us to know quite a lot about the structure of the covariance matrix $\boldsymbol{\Omega}$ and its inverse. GLS estimation is easy to do if the matrix $\boldsymbol{\Psi}$, defined in (9.02), is known and has a form that allows us to calculate $\boldsymbol{\Psi}^\top \boldsymbol{x}$, for any vector $\boldsymbol{x}$, without having to store $\boldsymbol{\Psi}$ itself in memory. If so, we can easily formulate the transformed model (9.03) and estimate it by OLS.

There is one important difference between (9.03) and the usual linear regression model. For the latter, the variance of the disturbances is unknown, while for the former, it is known to be 1. Since we can obtain OLS estimates without knowing the variance of the disturbances, this suggests that we should not need to know everything about $\boldsymbol{\Omega}$ in order to obtain GLS estimates. Suppose that $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{\Delta}$, where the $n \times n$ matrix $\boldsymbol{\Delta}$ is known to the investigator, but the positive scalar σ^2 is unknown. Then if we replace $\boldsymbol{\Omega}$ by $\boldsymbol{\Delta}$ in the definition (9.02) of $\boldsymbol{\Psi}$, we can still run regression (9.03), but the disturbances now have variance σ^2 instead of variance 1. When we run this modified regression, we obtain the estimate

$$(\boldsymbol{X}^\top \boldsymbol{\Delta}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Delta}^{-1} \boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{y} = \hat{\boldsymbol{\beta}}_{\text{GLS}},$$

where the equality follows immediately from the fact that $\sigma^2/\sigma^2 = 1$. Thus the GLS estimates are the same whether we use $\boldsymbol{\Omega}$ or $\boldsymbol{\Delta}$, that is, whether or not we know σ^2. However, if σ^2 is known, we can use the true covariance matrix (9.05). Otherwise, we must fall back on the estimated covariance matrix

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = s^2 (\boldsymbol{X}^\top \boldsymbol{\Delta}^{-1} \boldsymbol{X})^{-1},$$

where s^2 is the usual OLS estimate (4.63) of the error variance from the transformed regression.

Weighted Least Squares

It is particularly easy to obtain GLS estimates when the disturbances are heteroskedastic but uncorrelated. This implies that the matrix $\boldsymbol{\Omega}$ is diagonal. Let ω_t^2 denote the t^{th} diagonal element of $\boldsymbol{\Omega}$. Then $\boldsymbol{\Omega}^{-1}$ is a diagonal matrix with t^{th} diagonal element ω_t^{-2}, and $\boldsymbol{\Psi}$ can be chosen as the diagonal matrix with t^{th} diagonal element ω_t^{-1}. Thus we see that, for a typical observation, regression (9.03) can be written as

$$\omega_t^{-1} y_t = \omega_t^{-1} \boldsymbol{X}_t \boldsymbol{\beta} + \omega_t^{-1} u_t. \tag{9.12}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

This regression is to be estimated by OLS. The regressand and regressors are simply the dependent and independent variables multiplied by ω_t^{-1}, and the variance of the disturbance is clearly 1.

For obvious reasons, this special case of GLS estimation is often called **weighted least squares**, or **WLS**. The weight given to each observation when we run regression (9.12) is ω_t^{-1}. Observations for which the variance of the disturbance is large are given low weights, and observations for which it is small are given high weights. In practice, if $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{\Delta}$, with $\boldsymbol{\Delta}$ known but σ^2 unknown, regression (9.12) remains valid, provided we reinterpret ω_t^2 as the t^{th} diagonal element of $\boldsymbol{\Delta}$ and recognize that the variance of the disturbances is now σ^2 instead of 1.

There are various ways of determining the weights to be used in weighted least squares estimation. In the simplest case, either theory or preliminary testing may suggest that $\text{E}(u_t^2)$ is proportional to z_t^2, where z_t is some variable that we observe. For instance, z_t might be a variable like population or national income. In this case, z_t plays the role of ω_t in equation (9.12), because we want to weight the t^{th} observation by z_t^{-1}. Another possibility is that the data we actually observe were obtained by grouping data on different numbers of individual units. For example, suppose that the disturbances for the ungrouped data have constant variance, but that observation t is the average of N_t individual observations, where N_t varies. This implies that the variance of u_t must then be proportional to $1/N_t$. Thus, in this case, $N_t^{1/2}$ plays the role of ω_t^{-1} in equation (9.12). If the grouped data were sums instead of averages, the variance of u_t would be proportional to N_t, and $N_t^{-1/2}$ would play the role of ω_t^{-1}.

Weighted least squares estimation can easily be performed using any program for OLS estimation. When one is using such a procedure, it is important to remember that all the variables in the regression, *including the constant term*, must be multiplied by the same weights. Thus if, for example, the original regression is

$$y_t = \beta_1 + \beta_2 x_t + u_t,$$

the weighted regression is

$$y_t/\omega_t = \beta_1 (1/\omega_t) + \beta_2 (x_t/\omega_t) + u_t/\omega_t.$$

Here the regressand is y_t/ω_t, the regressor that corresponds to the constant term is $1/\omega_t$, and the regressor that corresponds to x_t is x_t/ω_t.

It is possible to report summary statistics like R^2, ESS, and SSR either in terms of the dependent variable y_t or in terms of the transformed regressand y_t/ω_t. However, it really only makes sense to report R^2 in terms of the transformed regressand. As we saw in Section 4.9, R^2 is valid as a measure of goodness of fit only when the residuals are orthogonal to the fitted values. This is true for the residuals and fitted values from OLS estimation of the weighted regression (9.12), but it is not true if those residuals and fitted values

Copyright © 2021, Russell Davidson and James G. MacKinnon

are subsequently multiplied by the ω_t in order to make them comparable with the original dependent variable.

9.4 Feasible Generalized Least Squares

In practice, the covariance matrix $\boldsymbol{\Omega}$ is often not known even up to a scalar factor. This makes it impossible to compute GLS estimates. However, in many cases it is reasonable to suppose that $\boldsymbol{\Omega}$, or $\boldsymbol{\Delta}$, depends in a known way on a vector of unknown parameters $\boldsymbol{\gamma}$. If so, it may be possible to estimate $\boldsymbol{\gamma}$ consistently, so as to obtain $\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$, say. Then $\boldsymbol{\Psi}(\hat{\boldsymbol{\gamma}})$ can be defined as in (9.02), and GLS estimates computed conditional on $\boldsymbol{\Psi}(\hat{\boldsymbol{\gamma}})$. This type of procedure is called **feasible generalized least squares**, or **feasible GLS**, because it is feasible in many cases when ordinary GLS is not.

As a simple example, suppose we want to obtain feasible GLS estimates of the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad \mathrm{E}(u_t^2) = \exp(\boldsymbol{Z}_t\boldsymbol{\gamma}), \tag{9.13}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are, respectively, a k–vector and an l–vector of unknown parameters, and $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$ are conformably dimensioned row vectors of observations on exogenous or predetermined variables that belong to the information set on which we are conditioning. Some or all of the elements of $\boldsymbol{Z}_t$ may well belong to $\boldsymbol{X}_t$. The function $\exp(\boldsymbol{Z}_t\boldsymbol{\gamma})$ is an example of a **skedastic function**. In the same way that a regression function determines the conditional expectation of a random variable, a skedastic function determines its conditional variance. The skedastic function $\exp(\boldsymbol{Z}_t\boldsymbol{\gamma})$ has the property that it is positive for any vector $\boldsymbol{\gamma}$. This is a desirable property for any skedastic function to have, since negative estimated variances would be highly inconvenient.

In order to obtain consistent estimates of $\boldsymbol{\gamma}$, usually we must first obtain consistent estimates of the disturbances in (9.13). The obvious way to do so is to start by computing OLS estimates $\hat{\boldsymbol{\beta}}$. This allows us to calculate a vector of OLS residuals with typical element $\hat{u}_t$. We can then run the auxiliary linear regression

$$\log \hat{u}_t^2 = \boldsymbol{Z}_t\boldsymbol{\gamma} + v_t, \tag{9.14}$$

over observations $t = 1, \ldots, n$ to find the OLS estimates $\hat{\boldsymbol{\gamma}}$. These estimates are then used to compute

$$\hat{\omega}_t = \big(\exp(\boldsymbol{Z}_t\hat{\boldsymbol{\gamma}})\big)^{1/2}$$

for all t. Finally, feasible GLS estimates of $\boldsymbol{\beta}$ are obtained by using ordinary least squares to estimate regression (9.12), with the estimates $\hat{\omega}_t$ replacing the unknown ω_t. This is an example of **feasible weighted least squares**.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Why Feasible GLS Works Asymptotically

Under suitable regularity conditions, it can be shown that this type of procedure yields a feasible GLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ that is consistent and asymptotically equivalent to the GLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$. We will not attempt to provide a rigorous proof of this proposition; for that, see Amemiya (1973a). However, we will try to provide an intuitive explanation of why it is true.

If we substitute $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$ for $\boldsymbol{y}$ into expression (9.04), the formula for the GLS estimator, we find that

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \boldsymbol{\beta}_0 + (\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{u},$$

from which we see that

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} - \boldsymbol{\beta}_0) = \big(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X}\big)^{-1}n^{-1/2}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{u}. \tag{9.15}$$

Under standard assumptions, the first matrix on the right-hand side tends in probability to a nonstochastic $k \times k$ matrix with full rank as $n \to \infty$, while the vector that postmultiplies it tends in distribution to a multivariate normal distribution.

For the feasible GLS estimator, the analog of equation (9.15) is

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{F}} - \boldsymbol{\beta}_0) = \big(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\gamma}})\boldsymbol{X}\big)^{-1}n^{-1/2}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\gamma}})\boldsymbol{u}. \tag{9.16}$$

The right-hand sides of expressions (9.16) and (9.15) look very similar, and it is clear that the latter must be asymptotically equivalent to the former if

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\gamma}})\boldsymbol{X} = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{X} \tag{9.17}$$

and

$$n^{-1/2}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\gamma}})\boldsymbol{u} \stackrel{a}{=} n^{-1/2}\boldsymbol{X}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{u}; \tag{9.18}$$

recall that things that depend on the sample size n are asymptotically equal if the difference between them converges to zero in probability.

A rigorous statement and proof of the conditions under which equations (9.17) and (9.18) hold is beyond the scope of this book. If they are to hold, it is desirable that $\hat{\boldsymbol{\gamma}}$ should be a consistent estimator of $\boldsymbol{\gamma}$, and this requires that the OLS estimator $\hat{\boldsymbol{\beta}}$ should be consistent. For example, it can be shown that the estimator obtained by running regression (9.14) would be consistent if the regressand depended on u_t rather than $\hat{u}_t$. Since the regressand is actually $\hat{u}_t$, it is necessary that the residuals $\hat{u}_t$ should consistently estimate the disturbances u_t. This in turn requires that $\hat{\boldsymbol{\beta}}$ should be consistent for $\boldsymbol{\beta}_0$. Thus, in general, we cannot expect $\hat{\boldsymbol{\gamma}}$ to be consistent if we do not start with a consistent estimator of $\boldsymbol{\beta}$.

Unfortunately, as we will see later, if $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ is not diagonal, then the OLS estimator $\hat{\boldsymbol{\beta}}$ is, in general, not consistent whenever any element of $\boldsymbol{X}_t$ is a

Copyright © 2021, Russell Davidson and James G. MacKinnon

lagged dependent variable. A lagged dependent variable is predetermined with respect to disturbances that are innovations, but not with respect to disturbances that are serially correlated. With GLS or feasible GLS estimation, the problem does not arise, because, if the model is correctly specified, the transformed explanatory variables are predetermined with respect to the transformed disturbances. When the OLS estimator is inconsistent, we must obtain a consistent estimator of $\boldsymbol{\gamma}$ in some other way.

Whether or not feasible GLS is a desirable estimation method in practice depends on how good an estimate of $\boldsymbol{\Omega}$ can be obtained. If $\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$ is a very good estimate, then feasible GLS has essentially the same properties as GLS itself, and inferences based on the GLS covariance matrix (9.05), with $\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$ replacing $\boldsymbol{\Omega}$, should be reasonably reliable, even though they are not exact in finite samples. Note that condition (9.17), in addition to being necessary for the validity of feasible GLS, guarantees that the feasible GLS covariance matrix estimator converges as $n \to \infty$ to the true GLS covariance matrix. On the other hand, if $\boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}})$ is a poor estimate, feasible GLS estimates may have quite different properties from real GLS estimates, and inferences may be quite misleading.

It is entirely possible to iterate a feasible GLS procedure. The estimator $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ can be used to compute a new set of residuals, which can be used to obtain a second-round estimate of $\boldsymbol{\gamma}$, which can be used to calculate second-round feasible GLS estimates, and so on. This procedure can either be stopped after a predetermined number of rounds or continued until convergence is achieved (if it ever is achieved). Iteration does not change the asymptotic distribution of the feasible GLS estimator, but it does change its finite-sample distribution.

9.5 Heteroskedasticity

There are two situations in which the disturbances are heteroskedastic but serially uncorrelated. In the first, the form of the heteroskedasticity is completely unknown, while, in the second, the skedastic function is known except for the values of some parameters that can be estimated consistently. Concerning the case of heteroskedasticity of unknown form, we saw in Section 6.4 how to compute asymptotically valid covariance matrix estimates for OLS parameter estimates. The fact that these HCCMEs are sandwich covariance matrices makes it clear that, although they are consistent under standard regularity conditions, OLS is not efficient when the disturbances are heteroskedastic.

If the variances of all the disturbances are known, at least up to a scalar factor, then efficient estimates can be obtained by weighted least squares, which we discussed in Section 9.3. For a linear model, we need to multiply all of the variables by ω_t^{-1}, the inverse of the standard error of u_t, and then use ordinary least squares. The usual OLS covariance matrix is perfectly valid, although

Copyright © 2021, Russell Davidson and James G. MacKinnon

it is desirable to replace s^2 by 1 if the variances are completely known, since in that case $s^2 \to 1$ as $n \to \infty$.

If the form of the heteroskedasticity is known, but the skedastic function depends on unknown parameters, then we can use feasible weighted least squares and still achieve asymptotic efficiency. An example of such a procedure was discussed in the previous section. As we have seen, it makes no difference asymptotically whether the ω_t are known or merely estimated consistently, although it can certainly make a substantial difference in finite samples. Asymptotically, at least, the usual OLS covariance matrix is just as valid with feasible WLS as with WLS.

Testing for Heteroskedasticity

In some cases, it may be clear from the specification of the model that the disturbances must exhibit a particular pattern of heteroskedasticity. In many cases, however, we may hope that the disturbances are homoskedastic but be prepared to admit the possibility that they are not. In such cases, if we have no information on the form of the skedastic function, it may be prudent to employ an HCCME, especially if the sample size is large. In a number of simulation experiments, Andrews (1991) has shown that, when the disturbances are homoskedastic, use of an HCCME, rather than the usual OLS covariance matrix, frequently has little cost. However, as we saw in Exercise 6.12 this is not always true. In finite samples, tests and confidence intervals based on HCCMEs are somewhat less reliable than ones based on the usual OLS covariance matrix when the latter is appropriate.

If we have information on the form of the skedastic function, we might well wish to use weighted least squares. Before doing so, it is advisable to perform a **specification test** of the null hypothesis that the disturbances are homoskedastic against whatever heteroskedastic alternatives may seem reasonable. There are many ways to perform this type of specification test. The simplest approach that is widely applicable, and the only one that we will discuss, involves running an artificial regression in which the regressand is the vector of squared residuals from the model under test.

A reasonably general model of heteroskedasticity conditional on some explanatory variables $\boldsymbol{Z}_t$ is

$$\mathrm{E}(u_t^2 \,|\, \Omega_t) = h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma}), \tag{9.19}$$

where the skedastic function $h(\cdot)$ is a nonlinear function that can take on only positive values, $\boldsymbol{Z}_t$ is a $1 \times r$ vector of observations on exogenous or predetermined variables that belong to the information set Ω_t, δ is a scalar parameter, and $\boldsymbol{\gamma}$ is an r–vector of parameters. Under the null hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$, the function $h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma})$ collapses to $h(\delta)$, a constant. One plausible specification of the skedastic function is

$$h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma}) = \exp(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma}) = \exp(\delta)\exp(\boldsymbol{Z}_t\boldsymbol{\gamma}).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

Under this specification, the variance of u_t reduces to the constant $\sigma^2 \equiv \exp(\delta)$ when $\boldsymbol{\gamma} = \mathbf{0}$. Since, as we will see, one of the advantages of the tests proposed here is that they do not depend on the functional form of $h(\cdot)$, there is no need for us to consider specifications less general than (9.19).

If we define v_t as the difference between u_t^2 and its conditional expectation, we can rewrite equation (9.19) as

$$u_t^2 = h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma}) + v_t, \tag{9.20}$$

which has the form of a regression model. While we would not expect the disturbance v_t to be as well behaved as the disturbances in most regression models, since the distribution of u_t^2 is almost always skewed to the right, it does have zero expectation by definition, and we will assume that it has a finite, and constant, variance. This assumption would probably be excessively strong if $\boldsymbol{\gamma}$ were nonzero, but it seems perfectly reasonable to assume that the variance of v_t is constant under the null hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$.

Suppose, to begin with, that we actually observe the u_t. In order to turn (9.20) into a linear regression model, we replace the nonlinear function h by a first-order Taylor approximation; see Section 6.8. We have, approximately, that

$$h(\delta + \boldsymbol{Z}_t\boldsymbol{\gamma}) = h(\delta) + h'(\delta)\boldsymbol{Z}_t\boldsymbol{\gamma},$$

and, by substituting this into (9.20), we find a linear regression model

$$u_t^2 = h(\delta) + h'(\delta)\boldsymbol{Z}_t\boldsymbol{\gamma} + v_t. \tag{9.21}$$

Here $h(\delta)$ is just a constant, and the constant factor $h'(\delta)$ can harmlessly be incorporated into the definition of the vector of coefficients of $\boldsymbol{Z}_t$, and so (9.21) is equivalent to the regression

$$u_t^2 = b_\delta + \boldsymbol{Z}_t\boldsymbol{b}_{\boldsymbol{\gamma}} + v_t \tag{9.22}$$

where we implicitly define the new parameters b_δ and $\boldsymbol{b}_{\boldsymbol{\gamma}}$. Observe that regression (9.22) indeed does not depend on the functional form of $h(\cdot)$. The null hypothesis of homoskedasticity can be expressed as $\boldsymbol{b}_{\boldsymbol{\gamma}} = \mathbf{0}$, and can be tested by, for instance, the ordinary F statistic, or by n times the centered R^2 from this regression, which is asymptotically distributed as $\chi^2(r)$.

In practice, of course, we do not actually observe the u_t. However, as we noted in Section 6.4, least squares residuals converge asymptotically to the corresponding disturbances when the model is correctly specified. Thus it seems plausible that the test should still be asymptotically valid if we replace u_t^2 in regression (9.22) by $\hat{u}_t^2$, the t^{th} squared residual from least squares estimation of the model under test. The test regression then becomes

$$\hat{u}_t^2 = b_\delta + \boldsymbol{Z}_t\boldsymbol{b}_{\boldsymbol{\gamma}} + \text{residual}. \tag{9.23}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

It can be shown that replacing u_t^2 by $\hat{u}_t^2$ does not change the asymptotic distribution of the F and nR^2 statistics for testing the hypothesis $\boldsymbol{b}_{\boldsymbol{\gamma}} = \mathbf{0}$; see Davidson and MacKinnon (1993, Section 11.5). Of course, since the finite-sample distributions of these test statistics may differ substantially from their asymptotic ones, it is a very good idea to bootstrap them when the sample size is small or moderate. This will be discussed further in Section 9.7.

Tests based on regression (9.23) require us to choose $\boldsymbol{Z}_t$, and there are many ways to do so. One approach is to include functions of some of the original regressors. As we saw in Section 5.5, there are circumstances in which the usual OLS covariance matrix is valid even when there is heteroskedasticity. White (1980) showed that, in a linear regression model, if $\mathrm{E}(u_t^2)$ is constant conditional on the squares and cross-products of all the regressors, then there is no need to use an HCCME; see Section 6.4. He therefore suggested that $\boldsymbol{Z}_t$ should consist of the squares and cross-products of all the regressors, because, asymptotically, such a test rejects the null whenever heteroskedasticity causes the usual OLS covariance matrix to be invalid. However, unless the number of regressors is very small, this suggestion results in r, the dimension of $\boldsymbol{Z}_t$, being very large. As a consequence, the test is likely to have poor finite-sample properties and low power, unless the sample size is quite large.

If economic theory does not tell us how to choose $\boldsymbol{Z}_t$, there is no simple, mechanical rule for choosing it. The more variables that are included in $\boldsymbol{Z}_t$, the greater is likely to be their ability to explain any observed pattern of heteroskedasticity, but the larger is the number of degrees of freedom for the test statistic. Adding a variable that helps substantially to explain the u_t^2 must surely increase the power of the test. However, adding variables with little explanatory power may simply dilute test power by increasing the number of degrees of freedom without increasing the noncentrality parameter; recall the discussion in Section 5.8. This is most easily seen in the context of χ^2 tests, where the critical values increase monotonically with the number of degrees of freedom. For a test with, say, $r + 1$ degrees of freedom to have as much power as a test with r degrees of freedom, the noncentrality parameter for the former test must be a certain amount larger than the noncentrality parameter for the latter.

9.6 Autoregressive and Moving-Average Processes

The disturbances for nearby observations may be correlated, or may appear to be correlated, in any sort of regression model, but this phenomenon is most commonly encountered in models estimated with time-series data, where it is known as **serial correlation** or **autocorrelation**. In practice, what appears to be serial correlation may instead be evidence of a misspecified model, as we discuss in Section 9.9. In some circumstances, though, it is natural to model the serial correlation by assuming that the disturbances follow some sort of

Copyright © 2021, Russell Davidson and James G. MacKinnon

stochastic process. Such a process defines a sequence of random variables. Some of the stochastic processes that are commonly used to model serial correlation will be discussed in this section.

If there is reason to believe that serial correlation may be present, the first step is usually to test the null hypothesis that the disturbances are serially uncorrelated against a plausible alternative that involves serial correlation. Several ways of doing this will be discussed in the next section. The second step, if evidence of serial correlation is found, is to estimate a model that accounts for it. An estimation method based on GLS will be discussed in Section 9.8. The final step, which is extremely important but is often omitted, is to verify that the model which accounts for serial correlation is compatible with the data. Some techniques for doing so will be discussed in Section 9.9.

The AR(1) Process

One of the simplest and most commonly used stochastic processes is the **first-order autoregressive process**, or **AR(1) process**. We have already encountered regression models with disturbances that follow such a process in Section 4.2. The AR(1) process can be written as

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2), \quad |\rho| < 1. \tag{9.24}$$

The disturbance at time t is equal to some fraction ρ of that at time $t-1$, with the sign changed if $\rho < 0$, plus the innovation ε_t. Since it is assumed that ε_t is independent of ε_s for all $s \neq t$, ε_t evidently is an innovation, according to the definition of that term in Section 5.5.

The condition in equation (9.24) that $|\rho| < 1$ is called a **stationarity condition**, because it is necessary for the AR(1) process to be **stationary**. There are several definitions of stationarity in time series analysis. According to the one that interests us here, a series with typical element u_t is stationary if the unconditional expectation $\text{E}(u_t)$ and the unconditional variance $\text{Var}(u_t)$ exist and are independent of t, and if the covariance $\text{Cov}(u_t, u_{t-j})$ is also, for any given j, independent of t. This particular definition is sometimes referred to as **covariance stationarity**, or **wide-sense stationarity**. Another term used to describe stationarity is **time-translation invariance**.

Suppose that, although we begin to observe the series only once $t = 1$, the series has been in existence for an infinite time. We can then compute the variance of u_t by substituting successively for u_{t-1}, u_{t-2}, u_{t-3}, and so on in (9.24). We see that

$$u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \cdots. \tag{9.25}$$

Using the fact that the innovations ε_t, $\varepsilon_{t-1}, \ldots$ are independent, and therefore uncorrelated, the variance of u_t is seen to be

$$\sigma_u^2 \equiv \text{Var}(u_t) = \sigma_\varepsilon^2 + \rho^2\sigma_\varepsilon^2 + \rho^4\sigma_\varepsilon^2 + \rho^6\sigma_\varepsilon^2 + \cdots = \frac{\sigma_\varepsilon^2}{1-\rho^2}. \tag{9.26}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

The last expression here is indeed independent of t, as required for a stationary process, but the last equality can be true only if the stationarity condition $|\rho| < 1$ holds, since that condition is necessary for the infinite series $1 + \rho^2 + \rho^4 + \rho^6 + \cdots$ to converge. In addition, if $|\rho| > 1$, the last expression in (9.26) is negative, and so cannot be a variance. In most econometric applications, where u_t is the disturbance appended to a regression model, the stationarity condition is a very reasonable condition to impose, since, without it, the variance of the disturbances would increase without limit as the sample size was increased.

It is not necessary to make the rather strange assumption that u_t exists for negative values of t all the way to $-\infty$. If we suppose that the expectation and variance of u_1 are respectively 0 and $\sigma_\varepsilon^2/(1-\rho^2)$, then we see at once that $\text{E}(u_2) = \text{E}(\rho u_1) + \text{E}(\varepsilon_2) = 0$, and that

$$\text{Var}(u_2) = \text{Var}(\rho u_1 + \varepsilon_2) = \sigma_\varepsilon^2\left(\frac{\rho^2}{1-\rho^2} + 1\right) = \frac{\sigma_\varepsilon^2}{1-\rho^2} = \text{Var}(u_1),$$

where the second equality uses the fact that ε_2, because it is an innovation, is uncorrelated with u_1. A simple recursive argument then shows that $\text{Var}(u_t) = \sigma_\varepsilon^2/(1-\rho^2)$ for all t.

The argument in (9.26) shows that $\sigma_u^2 \equiv \sigma_\varepsilon^2/(1-\rho^2)$ is the only admissible value for $\text{Var}(u_t)$ if the series is stationary. Consequently, if the variance of u_1 is *not* equal to σ_u^2, then the series cannot be stationary. However, if the stationarity condition is satisfied, $\text{Var}(u_t)$ must tend to σ_u^2 as t becomes large. This can be seen by repeating the calculation in (9.26), but recognizing that the series has only a finite number of terms. As t grows, the number of terms becomes large, and the value of the finite sum tends to the value of the infinite series, which is the stationary variance σ_u^2.

It is not difficult to see that, for the AR(1) process (9.24), the covariance of u_t and u_{t-1} is independent of t if $\text{Var}(u_t) = \sigma_u^2$ for all t.

$$\text{Cov}(u_t, u_{t-1}) = \text{E}(u_t u_{t-1}) = \text{E}\big((\rho u_{t-1} + \varepsilon_t)u_{t-1}\big) = \rho\sigma_u^2.$$

In order to compute the correlation of u_t and u_{t-1}, we divide $\text{Cov}(u_t, u_{t-1})$ by the square root of the product of the variances of u_t and u_{t-1}, that is, by σ_u^2. We then find that the correlation of u_t and u_{t-1} is just ρ.

The j^{th} order **autocovariance** of the AR(1) process is both the covariance of u_t and u_{t-j} and the covariance of u_t and u_{t+j}. As readers are asked to demonstrate in Exercise 9.4, under the assumption that $\text{Var}(u_1) = \sigma_u^2$, this autocovariance is equal to $\rho^j\sigma_u^2$, independently of t. It follows that the AR(1) process (9.24) is indeed covariance stationary if $\text{Var}(u_1) = \sigma_u^2$. The correlation between u_t and u_{t-j} is of course just ρ^j. Since ρ^j tends to zero quite rapidly as j increases, except when $|\rho|$ is very close to 1, this result implies that an AR(1) process generally exhibits small correlations between observations

Copyright © 2021, Russell Davidson and James G. MacKinnon

that are far removed in time, but it may exhibit large correlations between observations that are close in time. Since this is precisely the pattern that is frequently observed in the residuals of regression models estimated using time-series data, it is not surprising that the AR(1) process is often used to account for serial correlation in such models.

If we combine the result (9.26) with the result proved in Exercise 9.4, we see that, if the AR(1) process (9.24) is stationary, the covariance matrix of the vector $\boldsymbol{u}$, which is called the **autocovariance matrix** of the AR(1) process, can be written as

$$\boldsymbol{\Omega}(\rho) = \frac{\sigma_\varepsilon^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}. \tag{9.27}$$

All the u_t have the same variance, σ_u^2, which by (9.26) is the first factor on the right-hand side of (9.27). It follows that the second factor is the matrix of correlations of the disturbances, or **autocorrelation matrix**, which we denote $\boldsymbol{\Delta}(\rho)$. We will need to make use of (9.27) in Section 9.8 when we discuss GLS estimation of regression models with AR(1) disturbances.

Higher-Order Autoregressive Processes

Although the AR(1) process is very useful, it is quite restrictive. A much more general stochastic process is the **p^{th} order autoregressive process**, or **AR(p) process**,

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_p u_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2). \tag{9.28}$$

For such a process, u_t depends on up to p lagged values of itself, as well as on ε_t. The AR(p) process (9.28) can also be expressed as

$$\big(1 - \rho_1 \text{L} - \rho_2 \text{L}^2 - \cdots - \rho_p \text{L}^p\big) u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2), \tag{9.29}$$

where L denotes the **lag operator**. The lag operator L has the property that when L multiplies anything with a time subscript, this subscript is lagged one period. Thus $\text{L}u_t = u_{t-1}$, $\text{L}^2 u_t = u_{t-2}$, $\text{L}^3 u_t = u_{t-3}$, and so on. The expression in parentheses in (9.29) is a polynomial in the lag operator L, with coefficients 1 and $-\rho_1, \ldots, -\rho_p$. If we make the definition

$$\rho(z) \equiv \rho_1 z + \rho_2 z^2 + \cdots + \rho_p z^p \tag{9.30}$$

for arbitrary z, we can write the AR(p) process (9.29) very compactly as

$$\big(1 - \rho(\text{L})\big) u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2).$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

This compact notation is useful, but it does have two disadvantages: The order of the process, p, is not apparent, and there is no way of expressing any restrictions on the ρ_i.

The stationarity condition for an AR(p) process may be expressed in several ways. One of them, based on the definition (9.30), is that all the roots of the polynomial equation

$$1 - \rho(z) = 0 \tag{9.31}$$

must lie **outside the unit circle**. This simply means that all of the (possibly complex) roots of equation (9.31) must be greater than 1 in absolute value.[2] This condition can lead to quite complicated restrictions on the ρ_i for general AR(p) processes. The stationarity condition that $|\rho_1| < 1$ for an AR(1) process is evidently a consequence of this condition. In that case, (9.31) reduces to the equation $1 - \rho_1 z = 0$, the unique root of which is $z = 1/\rho_1$, and this root is greater than 1 in absolute value if and only if $|\rho_1| < 1$. As with the AR(1) process, the stationarity condition for an AR(p) process is necessary but not sufficient. Stationarity requires in addition that the variances and covariances of $u_1, \ldots, u_p$ should be equal to their stationary values. If not, it remains true that $\text{Var}(u_t)$ and $\text{Cov}(u_t, u_{t-j})$ tend to their stationary values for large t if the stationarity condition is satisfied, and so processes that are not stationary but satisfy the necessary condition may be called **asymptotically stationary**.

In practice, when an AR(p) process is used to model the disturbances of a regression model, p is usually chosen to be quite small. By far the most popular choice is the AR(1) process, but AR(2) and AR(4) processes are also encountered reasonably frequently. AR(4) processes are particularly attractive for quarterly data, because seasonality may cause correlation between disturbances that are four periods apart.

Moving-Average Processes

Autoregressive processes are not the only way to model stationary time series. Another type of stochastic process is the **moving-average**, or **MA**, process. The simplest of these is the **first-order moving-average**, or **MA(1)**, process

$$u_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1}, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2), \tag{9.32}$$

in which the disturbance u_t is a weighted average of two successive innovations, ε_t and ε_{t-1}.

It is not difficult to calculate the autocovariance matrix for an MA(1) process. From (9.32), we see that the variance of u_t is

$$\sigma_u^2 \equiv \text{E}\big((\varepsilon_t + \alpha_1 \varepsilon_{t-1})^2\big) = \sigma_\varepsilon^2 + \alpha_1^2 \sigma_\varepsilon^2 = (1 + \alpha_1^2)\sigma_\varepsilon^2,$$

[2] For a complex number $a + b\text{i}$, a and b real, the absolute value is $(a^2 + b^2)^{1/2}$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

the covariance of u_t and u_{t-1} is

$$\mathrm{E}\big((\varepsilon_t + \alpha_1\varepsilon_{t-1})(\varepsilon_{t-1} + \alpha_1\varepsilon_{t-2})\big) = \alpha_1\sigma_\varepsilon^2,$$

and the covariance of u_t and u_{t-j} for $j > 1$ is 0. Therefore, the covariance matrix of the entire vector $\boldsymbol{u}$ is

$$\sigma_\varepsilon^2\boldsymbol{\Delta}(\alpha_1) \equiv \sigma_\varepsilon^2\begin{bmatrix} 1+\alpha_1^2 & \alpha_1 & 0 & \cdots & 0 & 0 \\ \alpha_1 & 1+\alpha_1^2 & \alpha_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_1 & 1+\alpha_1^2 \end{bmatrix}. \tag{9.33}$$

The autocorrelation matrix is the matrix (9.33) divided by $\sigma_\varepsilon^2(1+\alpha_1^2)$. It is evident that there is no correlation between disturbances which are more than one period apart. Moreover, the correlation between successive disturbances varies only between -0.5 and 0.5, the smallest and largest possible values of $\alpha_1/(1+\alpha_1^2)$, which are achieved when $\alpha_1 = -1$ and $\alpha_1 = 1$, respectively. Therefore, an MA(1) process cannot be appropriate when the observed correlation between successive residuals is large in absolute value, or when residuals that are not adjacent are correlated.

Just as AR(p) processes generalize the AR(1) process, higher-order moving-average processes generalize the MA(1) process. The **q^{th} order moving-average process**, or **MA(q) process**, may be written as

$$u_t = \varepsilon_t + \alpha_1\varepsilon_{t-1} + \alpha_2\varepsilon_{t-2} + \cdots + \alpha_q\varepsilon_{t-q}, \;\; \varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2). \tag{9.34}$$

Using lag-operator notation, the process (9.34) can also be written as

$$u_t = (1 + \alpha_1\mathrm{L} + \cdots + \alpha_q\mathrm{L}^q)\varepsilon_t \equiv \big(1 + \alpha(\mathrm{L})\big)\varepsilon_t, \;\; \varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2),$$

where $\alpha(\mathrm{L})$ is a polynomial in the lag operator.

Autoregressive processes, moving-average processes, and other related stochastic processes have many important applications in both econometrics and macroeconomics. Their properties have been studied extensively in the literature on **time-series methods**. A classic reference is Box and Jenkins (1976), which has been updated as Box, Jenkins, and Reinsel (1994). Books that are specifically aimed at economists include Granger and Newbold (1986), Harvey (1989), Hamilton (1994), and Hayashi (2000).

Copyright © 2021, Russell Davidson and James G. MacKinnon

9.7 Testing for Serial Correlation

Over the decades, an enormous amount of research has been devoted to the subject of specification tests for serial correlation in regression models. Even though a great many different tests have been proposed, many of them no longer of much interest, the subject is not really very complicated. As we show in this section, it is perfectly easy to test the null hypothesis that the disturbances of a regression model are serially uncorrelated against the alternative that they follow an autoregressive process of any specified order. Most of the tests that we will discuss are straightforward applications of testing procedures which were introduced in Chapter 5.

The null hypothesis of no serial correlation is the usual linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{9.35}$$

The alternative hypothesis can be written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2), \tag{9.36}$$

in which the disturbances follow an AR(1) process. Let $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$ be the OLS estimates of $\boldsymbol{\beta}$ and the OLS residuals respectively from (9.35). Let us use (9.36) to solve first for ε_t, and then for u_t, so as to get

$$\begin{aligned} \varepsilon_t &= u_t - \rho u_{t-1} \quad \text{and} \quad u_t = y_t - \boldsymbol{X}_t\boldsymbol{\beta}, \\ \varepsilon_t &= y_t - \boldsymbol{X}_t\boldsymbol{\beta} - \rho(y_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{\beta}), \quad \text{that is,} \\ y_t &= \rho y_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} - \rho\boldsymbol{X}_{t-1}\boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2). \end{aligned} \tag{9.37}$$

The last line above is a nonlinear regression with white-noise disturbances, but, as with testing for heteroskedasticity in Section 9.5, it can be linearized, after which the null hypothesis that $\rho = 0$ can then be tested in a conventional manner.

The partial derivative of the right-hand side of (9.37) with respect to ρ is $y_{t-1} - \boldsymbol{X}_{t-1}\boldsymbol{\beta}$, and the vector of partial derivatives with respect to the components of $\boldsymbol{\beta}$ is $\boldsymbol{X}_t - \rho\boldsymbol{X}_{t-1}$. We can now perform a first-order Taylor approximation around $\rho = 0$, as required by the null hypothesis, and $\tilde{\boldsymbol{\beta}}$, the estimates obtained under the null. (Recall Section 6.8 for Taylor's theorem.) The right-hand side of (9.37) can be approximated by

$$\boldsymbol{X}_t(\tilde{\boldsymbol{\beta}} + \boldsymbol{b}_{\boldsymbol{\beta}}) + (y_{t-1} - \boldsymbol{X}_{t-1}\tilde{\boldsymbol{\beta}})\rho + \varepsilon_t,$$

where we have written $\boldsymbol{b}_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}$. Then the nonlinear regression (9.37) can be approximated by a linearized version, as follows:

$$y_t = \boldsymbol{X}_t(\tilde{\boldsymbol{\beta}} + \boldsymbol{b}_{\boldsymbol{\beta}}) + (y_{t-1} - \boldsymbol{X}_{t-1}\tilde{\boldsymbol{\beta}})\rho + \varepsilon_t$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

This can be rewritten more simply as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + b_\rho \tilde{u}_{t-1} + \varepsilon_t,$$

or, in vector-matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + b_\rho \tilde{\boldsymbol{u}}_1 + \boldsymbol{\varepsilon}, \tag{9.38}$$

where $\tilde{\boldsymbol{u}}_1$ has typical element $\tilde{u}_{t-1}$. It is not hard to show that the OLS estimate of b_ρ and the t statistic are identical to those obtained from the regression

$$\tilde{\boldsymbol{u}} = \boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\beta}} + b_\rho \tilde{\boldsymbol{u}}_1 + \text{residuals}. \tag{9.39}$$

The null can be tested straightforwardly using the asymptotic t statistic t_ρ for $b_\rho = 0$ in either (9.38) or (9.39). The latter regression demonstrates that the t statistic is a function of the residuals $\tilde{\boldsymbol{u}}$ and the regressors $\boldsymbol{X}$ only.

Although the regression (9.38) looks perfectly simple, it is not quite clear how it should be implemented. There are two approaches: run both (9.35) and (9.38) over the entire sample period, or omit the first observation from both. In the former case, the unobserved value of $\tilde{u}_0$ must be replaced by 0 before the test regression is run. As Exercise 9.14 demonstrates, the different approaches result in test statistics that are numerically different, even though they all follow the same asymptotic distribution under the null hypothesis.

Our approach to tests based on linearized regressions can readily be used to test against higher-order autoregressive processes and even moving-average processes. For example, in order to test against an AR(p) process, we can simply run the test regression

$$\tilde{u}_t = \boldsymbol{X}_t\boldsymbol{b}_{\boldsymbol{\beta}} + b_{\rho_1}\tilde{u}_{t-1} + \ldots + b_{\rho_p}\tilde{u}_{t-p} + \text{residual} \tag{9.40}$$

and use an asymptotic F test of the null hypothesis that the coefficients of all the lagged residuals are zero; see Exercise 9.6. Of course, in order to run regression (9.40), we either need to drop the first p observations or replace the unobserved lagged values of $\tilde{u}_t$ with zeros.

If we wish to test against an MA(q) process, it turns out that we can proceed exactly as if we were testing against an AR(q) process. The reason is that an autoregressive process of any order is **locally equivalent** to a moving-average process of the same order. Intuitively, this means that, for large samples, an AR(q) process and an MA(q) process look the same in the neighborhood of the null hypothesis of no serial correlation. Since our tests based on linearized regressions use information on first derivatives only, it should not be surprising that those used for testing against both alternatives turn out to be identical; see Exercise 9.7.

The use of (9.39) for testing against AR(1) disturbances was first suggested by Durbin (1970). Breusch (1978) and Godfrey (1978a, 1978b) subsequently showed how to use similar testing regressions to test against AR(p) and MA(q) disturbances. For a more detailed treatment of these and related procedures, see Godfrey (1988).

Copyright © 2021, Russell Davidson and James G. MacKinnon

The Durbin-Watson Statistic

The best-known test statistic for serial correlation is the ***d* statistic** proposed by Durbin and Watson (1950, 1951) and commonly referred to as the **DW statistic**.

It is completely determined by the least squares residuals $\tilde{\boldsymbol{u}}$ of the model (9.35) under test:

$$\begin{aligned} d &= \frac{\sum_{t=2}^{n}(\tilde{u}_t - \tilde{u}_{t-1})^2}{\sum_{t=1}^{n}\tilde{u}_t^2} \\ &= \frac{n^{-1}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}} + n^{-1}\tilde{\boldsymbol{u}}_1^\top\tilde{\boldsymbol{u}}_1}{n^{-1}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}} - \frac{n^{-1}\tilde{u}_1^2 + 2n^{-1}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}_1}{n^{-1}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}}. \end{aligned} \tag{9.41}$$

If we ignore the difference between $n^{-1}\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}$ and $n^{-1}\tilde{\boldsymbol{u}}_1^\top\tilde{\boldsymbol{u}}_1$, and the term $n^{-1}\tilde{u}_1^2$, both of which clearly tend to zero as $n \to \infty$, it can be seen that the first term in the second line of (9.41) tends to 2 and the second term tends to $-2\tilde{\rho}$, where $\tilde{\rho} \equiv \tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}_1/\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}$ can be thought of as a crude estimator of ρ. Therefore, d is asymptotically equal to $2 - 2\tilde{\rho}$. In samples of reasonable size, then, a value of $d \cong 2$ corresponds to the absence of serial correlation in the residuals, while values of d less than 2 correspond to $\tilde{\rho} > 0$, and values greater than 2 correspond to $\tilde{\rho} < 0$. It is important to note that the DW statistic is not valid when there are lagged dependent variables among the regressors.

In Section 3.3 we saw that, for a correctly specified linear regression model, the residual vector $\tilde{\boldsymbol{u}}$ is equal to $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$. Therefore, even if the disturbances are serially independent, the residuals generally display a certain amount of serial correlation. This implies that the finite-sample distributions of all the test statistics we have discussed, including that of the DW statistic, depend on $\boldsymbol{X}$. In practice, applied workers generally make use of the fact that the critical values for d are known to fall between two bounding values, d_L and d_U, which depend only on the sample size, n, the number of regressors, k, and whether or not there is a constant term. These bounding critical values have been tabulated for many values of n and k; see Savin and White (1977).

The need for special tables, among other relevant considerations, mean that the Durbin-Watson statistic, despite its popularity, is not very satisfactory. Using it with these tables is relatively cumbersome and often yields inconclusive results. Moreover, the tables allow us to perform one-tailed tests against the alternative that $\rho > 0$ only. Since the alternative that $\rho < 0$ is often of interest as well, the inability to perform a two-tailed test, or a one-tailed test against this alternative, is a serious limitation. Although exact P values for both one-tailed and two-tailed tests, which depend on the $\boldsymbol{X}$ matrix, can be obtained by using appropriate software, many computer programs do not offer this capability. In addition, the DW statistic is not valid when the regressors include lagged dependent variables, and it cannot easily be generalized to test for higher-order processes. Fortunately, the development of simulation-based tests has made the DW statistic obsolete.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Monte Carlo and Bootstrap Tests for Serial Correlation

We discussed simulation-based tests, including Monte Carlo tests and bootstrap tests, at some length in Chapter 7. The techniques discussed there can readily be applied to the problem of testing for serial correlation in linear regression models.

The test statistics we have discussed, the t statistic from the testing regression (9.38) and d, are pivotal under the null hypothesis that $\rho = 0$ when the assumptions of the classical normal linear model are satisfied. This makes it possible to perform Monte Carlo tests that are exact in finite samples. Pivotalness follows from two properties that are shared by these statistics: first, they depend only on the residuals $\tilde{u}_t$ obtained by estimation under the null hypothesis and the exogenous explanatory variables $\boldsymbol{X}$. The distribution of the residuals depends on $\boldsymbol{X}$, but this matrix is given and the same for all DGPs in a classical normal linear model. The distribution does not depend on the parameter vector $\boldsymbol{\beta}$ of the regression function, because, if $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, then $\boldsymbol{M_X y} = \boldsymbol{M_X u}$ whatever the value of the vector $\boldsymbol{\beta}$.

The second property of the statistics is **scale invariance**. By this, we mean that multiplying the dependent variable by an arbitrary scalar λ leaves the statistic unchanged. In a linear regression model, multiplying the dependent variable by λ causes the residuals to be multiplied by λ. But the t statistic from (9.38) and the DW statistic d are clearly unchanged if all the residuals are multiplied by the same constant, and so they are scale invariant. Since the residuals $\tilde{\boldsymbol{u}}$ are equal to $\boldsymbol{M_X u}$, it follows that multiplying σ by an arbitrary λ multiplies the residuals by λ. Consequently, the distributions of the statistics are independent of σ^2 as well as of $\boldsymbol{\beta}$. This implies that, for the classical normal linear model, both statistics are pivotal.

We now outline how to perform Monte Carlo tests for serial correlation in the context of the classical normal linear model. Let us call the test statistic we are using τ and its realized value $\hat{\tau}$. If we want to test for AR(1) disturbances, the best choice for the statistic τ is the t statistic t_ρ from (9.38), but it could also be the DW statistic. If we want to test for AR(p) disturbances, the best choice for τ would be the F statistic from (9.40).

The first step, evidently, is to compute $\hat{\tau}$. The next step is to generate B sets of simulated residuals and use each of them to compute a simulated test statistic, say τ_j^*, for $j = 1, \ldots, B$. Because the parameters do not matter, we can simply draw B vectors $\boldsymbol{u}_j^*$ from the $\mathrm{N}(\boldsymbol{0}, \boldsymbol{\mathrm{I}})$ distribution and regress each of them on $\boldsymbol{X}$ to generate the simulated residuals $\boldsymbol{M_X u}_j^*$, which are then used to compute τ_j^*. This can be done very inexpensively. The final step is to calculate an estimated P value. This can be done for either a one-tailed or a two-tailed test; see Section 7.3.

Whenever the regression function contains lagged dependent variables, or whenever the distribution of the disturbances is unknown, none of the standard test statistics for serial correlation is pivotal. Nevertheless, it is still possible to obtain very accurate inferences, even in quite small samples, by using

Copyright © 2021, Russell Davidson and James G. MacKinnon

bootstrap tests. The procedure is essentially unchanged from the Monte Carlo test. We still generate B simulated test statistics and use them to compute a P value. For best results, the test statistic used should be asymptotically valid for the model that is being tested. In particular, we should avoid d whenever there are lagged dependent variables.

It is extremely important to generate the bootstrap samples in such a way that they are compatible with the model under test. Ways of generating bootstrap samples for regression models were discussed in Section 7.4. When the model includes lagged dependent variables, we need to generate $\boldsymbol{y}_j^*$ rather than just $\boldsymbol{u}_j^*$. For this, we need estimates of the parameters of the regression function. If the model includes lagged dependent variables, we must generate the bootstrap samples recursively, as in (7.09). Unless we are going to assume that the disturbances are normally distributed, we should draw the bootstrap disturbances from the EDF of the residuals for the model under test, after they have been appropriately rescaled. Recall that there is more than one way to do this. The simplest approach is just to multiply each residual by $(n/(n-k))^{1/2}$, as in (7.11).

We strongly recommend the use of simulation-based tests for serial correlation, rather than asymptotic tests. Monte Carlo tests are appropriate only in the context of the classical normal linear model, but bootstrap tests are appropriate under much weaker assumptions. It is generally a good idea to test for both AR(1) disturbances and higher-order ones, at least fourth-order in the case of quarterly data, and at least twelfth-order in the case of monthly data.

Heteroskedasticity-Robust Tests

The tests for serial correlation that we have discussed are based on the assumption that the disturbances are homoskedastic. When this crucial assumption is violated, the asymptotic distributions of all the test statistics differ from whatever distributions they are supposed to follow asymptotically.

Suppose we wish to test the linear regression model (9.35), in which the disturbances are serially uncorrelated, against the alternative that the disturbances follow an AR(p) process. Under the assumption of homoskedasticity, we could simply run the testing regression (9.40) and use an asymptotic F test. If we let $\boldsymbol{Z}$ denote an $n \times p$ matrix with typical element $Z_{ti} = \tilde{u}_{t-i}$, where any missing lagged residuals are replaced by zeros, this regression can be written as

$$\tilde{\boldsymbol{u}} = \boldsymbol{Xb} + \boldsymbol{Zc} + \text{residuals}. \tag{9.42}$$

The ordinary F test for $\boldsymbol{c} = \boldsymbol{0}$ in (9.42) is not robust to heteroskedasticity, but it is straightforward to compute a robust Wald test using an HCCME.

Although this heteroskedasticity-robust test is asymptotically valid, it is not exact in finite samples. We can expect to obtain more reliable results by using bootstrap P values instead of asymptotic ones. This requires the use

Copyright © 2021, Russell Davidson and James G. MacKinnon

of a bootstrap DGP adapted to heteroskedasticity; the best choice is the wild bootstrap.

Other Tests Based on OLS Residuals

The tests for serial correlation that we have discussed in this section are by no means the only scale-invariant tests based on least squares residuals that are regularly encountered in econometrics. Many tests for heteroskedasticity, skewness, kurtosis, and other deviations from the NID assumption also have these properties. For example, consider tests for heteroskedasticity based on regression (9.23). Nothing in that regression depends on $\boldsymbol{y}$ except for the squared residuals that constitute the regressand. Further, it is clear that both the F statistic for the hypothesis that $\boldsymbol{b}_\gamma = \boldsymbol{0}$ and n times the centered R^2 are scale invariant. Therefore, for a classical normal linear model with $\boldsymbol{X}$ and $\boldsymbol{Z}$ fixed, these statistics are pivotal. Consequently, Monte Carlo tests based on them, in which we draw the disturbances from the $\mathrm{N}(0,1)$ distribution, are exact in finite samples.

When the normality assumption is not appropriate, we have two options. If some other distribution that is known up to a scale parameter is thought to be appropriate, we can draw the disturbances from it instead of from the $\mathrm{N}(0,1)$ distribution. Then, if the assumed distribution really is the true one, we obtain an exact test. Alternatively, we can perform a bootstrap test in which the disturbances are obtained either by resampling the rescaled residuals or using the wild bootstrap. This is also appropriate when there are lagged dependent variables among the regressors. The bootstrap test is not exact, but it should still perform well in finite samples no matter how the disturbances actually happen to be distributed.

9.8 Estimating Models with Autoregressive Disturbances

If we decide that the disturbances of a regression model are serially correlated, either on the basis of theoretical considerations or as a result of specification testing, and we are confident that the regression function itself is not misspecified, the next step is to estimate a modified model which takes account of the serial correlation. The simplest such model is (9.36), which is the original regression model modified by having the disturbances follow an AR(1) process. For ease of reference, we rewrite (9.36) here:

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2). \tag{9.43}$$

In many cases, as we will discuss in the next section, the best approach may actually be to specify a more complicated, dynamic, model for which the disturbances are not serially correlated. In this section, however, we ignore this important issue and simply discuss how to estimate the model (9.43) under various assumptions.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Estimation by Feasible GLS

We have seen that, if the u_t follow a stationary AR(1) process, that is, if $|\rho| < 1$ and $\mathrm{Var}(u_1) = \sigma_u^2 = \sigma_\varepsilon^2/(1-\rho^2)$, then the covariance matrix of the entire vector $\boldsymbol{u}$ is the $n \times n$ matrix $\boldsymbol{\Omega}(\rho)$ given in equation (9.27). In order to compute GLS estimates, we need to find a matrix $\boldsymbol{\Psi}$ with the property that $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ is proportional to $\boldsymbol{\Omega}^{-1}$. This property is satisfied whenever the covariance matrix of the vector $\boldsymbol{\Psi}^\top\boldsymbol{u}$ is proportional to the identity matrix, which it must be if we choose $\boldsymbol{\Psi}$ in such a way that $\boldsymbol{\Psi}^\top\boldsymbol{u} = \boldsymbol{\varepsilon}$.

For $t = 2, \ldots, n$, we know from (9.24) that

$$\varepsilon_t = u_t - \rho u_{t-1}, \tag{9.44}$$

and this allows us to construct the rows of $\boldsymbol{\Psi}^\top$ except for the first row. The t^{th} row has 1 in the t^{th} position, $-\rho$ in the $(t-1)^{\text{st}}$ position, and 0s everywhere else.

For the first row of $\boldsymbol{\Psi}^\top$, however, we need to be a little more careful. Under the hypothesis of stationarity of $\boldsymbol{u}$, the variance of u_1 is σ_u^2. Further, since the ε_t are innovations, u_1 is uncorrelated with the ε_t for $t = 2, \ldots, n$. Thus, if we define ε_1 by the formula

$$\varepsilon_1 = (\sigma_\varepsilon/\sigma_u)u_1 = (1-\rho^2)^{1/2}u_1, \tag{9.45}$$

it can be seen that the n–vector $\boldsymbol{\varepsilon}$, with the first component ε_1 defined by (9.45) and the remaining components ε_t defined by (9.44), has a covariance matrix equal to $\sigma_\varepsilon^2\mathbf{I}$.

Putting together (9.44) and (9.45), we conclude that $\boldsymbol{\Psi}^\top$ should be defined as an $n \times n$ matrix with all diagonal elements equal to 1 except for the first, which is equal to $(1-\rho^2)^{1/2}$, and all other elements equal to 0 except for the ones on the diagonal immediately below the principal diagonal, which are equal to $-\rho$. In terms of $\boldsymbol{\Psi}$ rather than of $\boldsymbol{\Psi}^\top$, we have:

$$\boldsymbol{\Psi}(\rho) = \begin{bmatrix} (1-\rho^2)^{1/2} & -\rho & 0 & \cdots & 0 & 0 \\ 0 & 1 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\rho \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \tag{9.46}$$

where the notation $\boldsymbol{\Psi}(\rho)$ emphasizes that the matrix depends on the usually unknown parameter ρ. The matrix $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ is proportional to the inverse of the autocovariance matrix that appears in equation (9.27). The calculations needed to show that this is so are outlined in Exercises 9.9 and 9.10.

It is essential that the AR(1) parameter ρ is either known or is consistently estimable. If we know ρ, we can obtain GLS estimates. If we do not know it

Copyright © 2021, Russell Davidson and James G. MacKinnon

but can estimate it consistently, we can obtain feasible GLS estimates. For the case in which the explanatory variables are all exogenous, the simplest way to estimate ρ consistently is to use the estimator $\tilde{\rho}$ from the regression

$$\tilde{u}_t = b_\rho \tilde{u}_{t-1} + \text{residual}, \quad t = 1, \dots, n, \tag{9.47}$$

where, as above, the $\tilde{u}_t$ are the residuals from regression (9.35). In order to be able to keep the first observation, we assume that $\tilde{u}_0 = 0$. This regression yields an estimate of b_ρ, which we will call $\tilde{\rho}$ because it is an estimate of ρ based on the residuals under the null. Explicitly, we have

$$\tilde{\rho} = \frac{\sum_{t=1}^n \tilde{u}_t \tilde{u}_{t-1}}{\sum_{t=1}^n \tilde{u}_{t-1}^2}, \tag{9.48}$$

It turns out that, if the explanatory variables $\boldsymbol{X}$ in (9.35) are all exogenous, then $\tilde{\rho}$ is a consistent estimator of the parameter ρ in model (9.36), or, equivalently, (9.37), where it is not assumed that $\rho = 0$. This slightly surprising result depends crucially on the assumption of exogenous regressors. If one of the variables in $\boldsymbol{X}$ is a lagged dependent variable, the result no longer holds.

Whatever estimate of ρ is used must satisfy the stationarity condition that $|\rho| < 1$, without which the process would not be stationary, and the transformation for the first observation would involve taking the square root of a negative number. Unfortunately, the estimator $\tilde{\rho}$ is not guaranteed to satisfy the stationarity condition, although, in practice, it is very likely to do so when the model is correctly specified, even if the true value of ρ is quite large in absolute value.

Whether ρ is known or estimated, the next step in GLS estimation is to form the vector $\boldsymbol{\Psi}^\top \boldsymbol{y}$ and the matrix $\boldsymbol{\Psi}^\top \boldsymbol{X}$. It is easy to do this without having to store the $n \times n$ matrix $\boldsymbol{\Psi}$ in computer memory. The first element of $\boldsymbol{\Psi}^\top \boldsymbol{y}$ is $(1 - \rho^2)^{1/2} y_1$, and the remaining elements have the form $y_t - \rho y_{t-1}$. Each column of $\boldsymbol{\Psi}^\top \boldsymbol{X}$ has precisely the same form as $\boldsymbol{\Psi}^\top \boldsymbol{y}$ and can be calculated in precisely the same way.

The final step is to run an OLS regression of $\boldsymbol{\Psi}^\top \boldsymbol{y}$ on $\boldsymbol{\Psi}^\top \boldsymbol{X}$. This regression yields the (feasible) GLS estimates

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{y} \tag{9.49}$$

along with the estimated covariance matrix

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = s^2 (\boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{\Psi}^\top \boldsymbol{X})^{-1}, \tag{9.50}$$

where s^2 is the usual OLS estimate of the variance of the disturbances. Of course, the estimator (9.49) is formally identical to (9.04), since (9.49) is valid for any $\boldsymbol{\Psi}$ matrix.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The main weakness of GLS as used above arises whenever one or more of the explanatory variables are lagged dependent variables, or, more generally, predetermined but not exogenous variables. Even with a consistent estimator of ρ, one of the conditions for the applicability of feasible GLS, condition (9.18), does not hold when any elements of $\boldsymbol{X}_t$ are not exogenous. Fortunately, there is not much temptation to use GLS if the non-exogenous explanatory variables are lagged variables, because lagged variables are not observed for the first observation. In all events, the conclusion is simple: We should avoid GLS if the explanatory variables are not all exogenous.

In Section 9.4, we mentioned the possibility of using an **iterated feasible GLS** procedure. We can now see precisely how such a procedure would work for this model. In the first step, we obtain the OLS parameter vector $\tilde{\boldsymbol{\beta}}$. In the second step, the formula (9.48) is evaluated at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ to obtain $\tilde{\rho}$, a consistent estimate of ρ. In the third step, we use (9.49) to obtain the feasible GLS estimate $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$. At this point, we go back to the second step and use $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$ to update the residuals $\tilde{u}_t$ which can then be used in (9.48) for an updated estimate of ρ, which we subsequently use in (9.49) for the next estimate of $\boldsymbol{\beta}$. The iterative procedure may then be continued until convergence, assuming that it does converge.

Although the iterated feasible GLS estimator generally performs well, it does have one weakness: If $\hat{\rho}$ denotes the iterated estimate of ρ, there is no way to ensure that $|\hat{\rho}| < 1$. In the unlikely but not impossible event that $|\hat{\rho}| \geq 1$, the estimated covariance matrix (9.50) is not valid. In such cases, one can use maximum likelihood estimation (not discussed in this book; see the textbook ETM, Chapter 9), which shares the good properties of iterated feasible GLS while also ensuring that the estimate of ρ satisfies the stationarity condition.

The iterated feasible GLS procedure considered above has much in common with a very old, but still widely-used, algorithm for estimating models with stationary AR(1) disturbances. This algorithm, which is called **iterated Cochrane-Orcutt**, was originally proposed in a classic paper by Cochrane and Orcutt (1949). It works in exactly the same way as iterated feasible GLS, except that it omits the first observation. The properties of this algorithm are explored in Exercises 9.15-16.

9.9 Specification Testing and Serial Correlation

Models estimated using time-series data frequently appear to have disturbances which are serially correlated. However, as we will see, many types of misspecification can create the *appearance* of serial correlation. Therefore, finding evidence of serial correlation does not mean that it is necessarily appropriate to model the disturbances as following some sort of autoregressive or moving-average process. If the regression function of the original model is misspecified in any way, then a model like (9.37), which has been modified to

Copyright © 2021, Russell Davidson and James G. MacKinnon

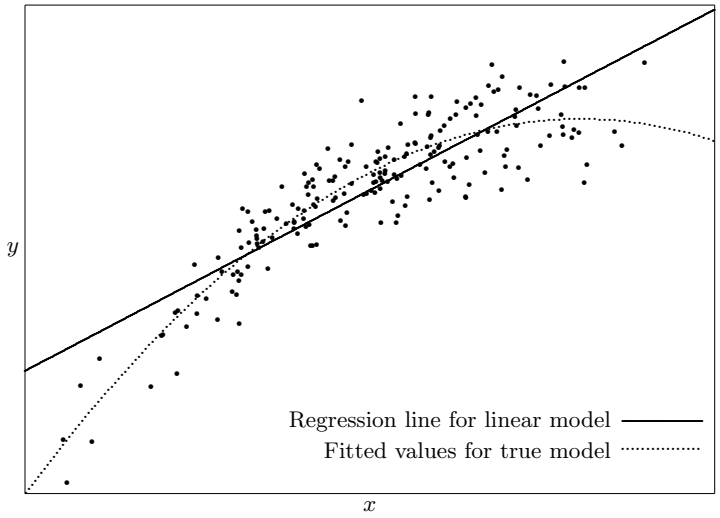


Figure 9.1 The appearance of serial correlation

incorporate AR(1) disturbances, is probably also misspecified. It is therefore extremely important to test the specification of any regression model that has been "corrected" for serial correlation.

The Appearance of Serial Correlation

There are several types of misspecification of the regression function that can incorrectly create the appearance of serial correlation. For instance, it may be that the true regression function is nonlinear in one or more of the regressors while the estimated one is linear. In that case, depending on how the data are ordered, the residuals from a linear regression model may well appear to be serially correlated. All that is needed is for the independent variables on which the dependent variable depends nonlinearly to be correlated with time.

As a concrete example, consider Figure 9.1, which shows 200 hypothetical observations on a regressor x and a regressand y, together with an OLS regression line and the fitted values from the true, nonlinear model. For the linear model, the residuals are always negative for the smallest and largest values of x, and they tend to be positive for the intermediate values. As a consequence, they appear to be serially correlated: If the observations are ordered according to the value of x, the estimate $\tilde{\rho}$ obtained by regressing the OLS residuals on themselves lagged once is 0.298, and the t statistic for $\rho = 0$ is 4.462. Thus, if the data are ordered in this way, there appears to be strong evidence of serial correlation. But this evidence is misleading. Either plotting the residuals against x or including x^2 as an additional regressor quickly reveals the true nature of the misspecification.

Copyright © 2021, Russell Davidson and James G. MacKinnon

The true regression function in this example contains a term in x^2. Since the linear model omits this term, it is underspecified, in the sense discussed in Section 4.8. Any sort of underspecification has the potential to create the appearance of serial correlation if the incorrectly omitted variables are themselves serially correlated. Therefore, whenever we find evidence of serial correlation, our first reaction should be to think carefully about the specification of the regression function. Perhaps one or more additional independent variables should be included among the regressors. Perhaps powers, cross-products, or lags of some of the existing independent variables need to be included. Or perhaps the regression function should be made dynamic by including one or more lags of the dependent variable.

9.10 Models for Panel Data

Many data sets are measured across two dimensions. One dimension is time, and the other is usually called the cross-section dimension. For example, we may have 40 annual observations on 25 countries, or 100 quarterly observations on 50 states, or 6 annual observations on 3100 individuals. Data of this type are often referred to as **panel data**. The disturbances for a model using panel data are likely to display certain types of dependence, which should be taken into account when we estimate such a model.

For simplicity, we restrict our attention to the linear regression model

$$y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + u_{it}, \quad i = 1, \dots, m, \;\; t = 1, \dots, T, \tag{9.51}$$

where $\boldsymbol{X}_{it}$ is a $1 \times k$ vector of observations on explanatory variables. There are assumed to be m cross-sectional units and T time periods, for a total of $n = mT$ observations. If each u_{it} has expectation zero conditional on its corresponding $\boldsymbol{X}_{it}$, we can estimate equation (9.51) by ordinary least squares. But the OLS estimator is not efficient if the u_{it} are not IID, and the IID assumption is rarely realistic with panel data.

If certain shocks affect the same cross-sectional unit at all points in time, the disturbances u_{it} and u_{is} must be correlated for all $t \neq s$. Similarly, if certain shocks affect all cross-sectional units at the same point in time, the disturbances u_{it} and u_{jt} must be correlated for all $i \neq j$. In consequence, if we use OLS, not only do we obtain inefficient parameter estimates, but we also obtain an inconsistent estimate of their covariance matrix; recall the discussion of Section 6.4. If the expectation of u_{it} conditional on $\boldsymbol{X}_{it}$ is *not* zero, then OLS actually yields inconsistent parameter estimates. This happens, for example, when $\boldsymbol{X}_{it}$ contains lagged dependent variables and the u_{it} are serially correlated.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Error-Components Models

The two most popular approaches for dealing with panel data are both based on what are called **error-components models**. The idea is to specify the disturbance u_{it} in (9.51) as consisting of two or three separate shocks, each of which is assumed to be independent of the others. A fairly general specification is

$$u_{it} = e_t + v_i + \varepsilon_{it}. \tag{9.52}$$

Here e_t affects all observations for time period t, v_i affects all observations for cross-sectional unit i, and ε_{it} affects only observation it. It is generally assumed that the e_t are independent across t, the v_i are independent across i, and the ε_{it} are independent across all i and t. Classic papers on error-components models include Balestra and Nerlove (1966), Fuller and Battese (1974), and Mundlak (1978).

In order to estimate an error-components model, the e_t and v_i can be regarded as being either fixed or random, in a sense that we will explain. If the e_t and v_i are thought of as **fixed effects**, then they are treated as parameters to be estimated. It turns out that they can then be estimated by OLS using dummy variables. If they are thought of as **random effects**, then we must figure out the covariance matrix of the u_{it} as functions of the variances of the e_t, v_i, and ε_{it}, and use feasible GLS. Each of these approaches can be appropriate in some circumstances but may be inappropriate in others.

In what follows, we simplify the error-components specification (9.52) by eliminating the e_t. Thus we assume that there are shocks specific to each cross-sectional unit, or group, but no time-specific shocks. This assumption is often made in empirical work, and it considerably simplifies the algebra. In addition, we assume that the $\boldsymbol{X}_{it}$ are exogenous. The presence of lagged dependent variables in panel data models raises a number of issues that we do not wish to discuss here; see Arellano and Bond (1991) and Arellano and Bover (1995).

Fixed-Effects Estimation

Fixed-effects estimation was discussed on Chapter 3. We recall that discussion here for convenience. The model that underlies fixed-effects estimation, based on equation (9.51) and the simplified version of equation (9.52), can be written as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma_\varepsilon^2 \mathbf{I}_n, \tag{9.53}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are n–vectors with typical elements y_{it} and ε_{it}, respectively, and $\boldsymbol{D}$ is an $n \times m$ matrix of indicator (dummy) variables. Column j of $\boldsymbol{D}$ indicates the cross-sectional unit j, $j = 1, \ldots, m$: the element in the row corresponding to observation it, for $i = 1, \ldots, m$ and $t = 1, \ldots, T$, is equal to 1 if $i = j$ and equal to 0 otherwise.[3] The m–vector $\boldsymbol{\eta}$ has typical element v_i,

[3] If the data are ordered so that all the observations in the first group appear first, followed by all the observations in the second group, and so on, the row corresponding to observation it is row $T(i-1)+t$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

and so it follows that the n–vector $\boldsymbol{D}\boldsymbol{\eta}$ has element v_i in the row corresponding to observation it. Note that there is exactly one element of $\boldsymbol{D}$ equal to 1 in each row, which implies that the n–vector $\boldsymbol{\iota}$ with each element equal to 1 is a linear combination of the columns of $\boldsymbol{D}$. Consequently, in order to avoid collinear regressors, the matrix $\boldsymbol{X}$ should not contain a constant.

The vector $\boldsymbol{\eta}$ plays the role of a parameter vector, and it is in this sense that the v_i are called *fixed* effects. They could in fact be random; the essential thing is that they must be uncorrelated with the disturbances ε_{it}. They may, however, be correlated with the explanatory variables in the matrix $\boldsymbol{X}$. Whether or not this is the case, the model (9.53), interpreted conditionally on $\boldsymbol{\eta}$, implies that the following functions of data and parameters are zero functions:

$$\boldsymbol{X}_{it}^\top(y_{it} - \boldsymbol{X}_{it}\boldsymbol{\beta} - v_i) \quad \text{and} \quad y_{it} - \boldsymbol{X}_{it}\boldsymbol{\beta} - v_i.$$

The **fixed-effects estimator**, which is the OLS estimator of $\boldsymbol{\beta}$ in equation (9.53), is based on the estimating equations implied by these estimating functions. Because of the way it is computed, this estimator is sometimes called the **least squares dummy variables**, or **LSDV**, estimator.

Let $\boldsymbol{M}_{\boldsymbol{D}}$ denote the projection matrix $\mathbf{I} - \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top$. Then, by the FWL Theorem, we know that the OLS estimator of $\boldsymbol{\beta}$ in (9.53) can be obtained by regressing $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{y}$, the residuals from a regression of $\boldsymbol{y}$ on $\boldsymbol{D}$, on $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{X}$, the matrix of residuals from regressing each of the columns of $\boldsymbol{X}$ on $\boldsymbol{D}$. The fixed-effects estimator is therefore

$$\hat{\boldsymbol{\beta}}_{\text{FE}} = (\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{y}. \tag{9.54}$$

For any n–vector $\boldsymbol{x}$, let $\bar{x}_i$ denote the **group mean** $T^{-1}\sum_{t=1}^T x_{it}$. Then it is easy to check that element it of the vector $\boldsymbol{M}_{\boldsymbol{D}}\boldsymbol{x}$ is equal to $x_{it} - \bar{x}_i$, the deviation from the group mean. Since all the variables in (9.54) are premultiplied by $\boldsymbol{M}_{\boldsymbol{D}}$, it follows that this estimator makes use only of the information in the variation around the mean for each of the m groups. For this reason, it is often called the **within-groups estimator**. Because $\boldsymbol{X}$ and $\boldsymbol{D}$ are exogenous, this estimator is unbiased. Moreover, since the conditions of the Gauss-Markov theorem are satisfied, we can conclude that the fixed-effects estimator is BLUE.

The fixed-effects estimator (9.54) has advantages and disadvantages. It is easy to compute, even when m is very large, because it is never necessary to make direct use of the $n \times n$ matrix $\boldsymbol{M}_{\boldsymbol{D}}$. All that is needed is to compute the m group means for each variable. In addition, the estimates $\hat{\boldsymbol{\eta}}$ of the fixed effects may well be of interest in their own right. However, the estimator cannot be used with an explanatory variable that takes on the same value for all the observations in each group, because such a column would be collinear with the columns of $\boldsymbol{D}$. More generally, if the explanatory variables in the matrix $\boldsymbol{X}$ are well explained by the dummy variables in $\boldsymbol{D}$, the parameter vector $\boldsymbol{\beta}$ is not estimated at all precisely. It is of course possible to estimate a constant, simply by taking the mean of the estimates $\hat{\boldsymbol{\eta}}$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Random-Effects Estimation

It is possible to improve on the efficiency of the fixed-effects estimator if one is willing to impose restrictions on the model (9.53). For that model, all we require is that the matrix $\boldsymbol{X}$ of explanatory variables and the cross-sectional shocks v_i should both be uncorrelated with the idiosyncratic shocks ε_{it}, but this does not rule out the possibility of a correlation between the variables in $\boldsymbol{X}$ and the v_i. The restrictions imposed for random-effects estimation require that $\mathrm{E}(v_i \,|\, \boldsymbol{X}) = 0$ for all $i = 1, \ldots, m$.

This assumption is by no means always plausible. For example, in a panel of observations on individual workers, an observed variable like the hourly wage rate may well be correlated with an unobserved variable like ability, which implicitly enters into the individual-specific disturbance v_i. However, if the assumption is satisfied, it follows that

$$\mathrm{E}(u_{it} \,|\, \boldsymbol{X}) = \mathrm{E}(v_i + \varepsilon_{it} \,|\, \boldsymbol{X}) = 0, \tag{9.55}$$

since v_i and ε_{it} are then both uncorrelated with $\boldsymbol{X}$. Condition (9.55) is precisely the condition which ensures that OLS estimation of the model (9.51) yields unbiased estimates.

However, OLS estimation of equation (9.51) is not in general efficient, because the u_{it} are not IID. We can calculate the covariance matrix of the u_{it} if we assume that the v_i are IID random variables with expectation zero and variance σ_v^2. This assumption accounts for the term "random" effects. From (9.52), setting $e_t = 0$ and using the assumption that the shocks are independent, we can easily see that

$$\begin{aligned} \mathrm{Var}(u_{it}) &= \sigma_v^2 + \sigma_\varepsilon^2, \\ \mathrm{Cov}(u_{it}\, u_{is}) &= \sigma_v^2, \text{ and} \\ \mathrm{Cov}(u_{it}\, u_{js}) &= 0 \text{ for all } i \neq j. \end{aligned}$$

These define the elements of the $n \times n$ covariance matrix $\boldsymbol{\Omega}$, which we need for GLS estimation. If the data are ordered by the cross-sectional units in m blocks of T observations each, this matrix has the form

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix},$$

where

$$\boldsymbol{\Sigma} \equiv \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_v^2\, \boldsymbol{\iota}\boldsymbol{\iota}^\top \tag{9.56}$$

is the $T \times T$ matrix with $\sigma_v^2 + \sigma_\varepsilon^2$ in every position on the principal diagonal and σ_v^2 everywhere else. Here $\boldsymbol{\iota}$ is a T–vector of 1s.

Copyright © 2021, Russell Davidson and James G. MacKinnon

To obtain GLS estimates of $\boldsymbol{\beta}$, we would need to know the values of σ_ε^2 and σ_v^2, or, at least, the value of their ratio, since, as we saw in Section 9.3, GLS estimation requires only that $\boldsymbol{\Omega}$ should be specified up to a factor. To obtain feasible GLS estimates, we need a consistent estimate of that ratio. However, the reader may have noticed that we have made no use in this section so far of asymptotic concepts, such as that of a consistent estimate. This is because, in order to obtain definite results, we must specify what happens to both m and T when $n = mT$ tends to infinity.

Consider the fixed-effects model (9.53). If m remains fixed as $T \to \infty$, then the number of regressors also remains fixed as $n \to \infty$, and standard asymptotic theory applies. But if T remains fixed as $m \to \infty$, then the number of parameters to be estimated tends to infinity, and the m–vector $\hat{\boldsymbol{\eta}}$ of estimates of the fixed effects is not consistent, because each estimated effect depends only on the finite number T of observations. It is nevertheless possible to show that, even in this case, $\hat{\boldsymbol{\beta}}$ remains consistent; see Exercise 9.20.

It is always possible to find a consistent estimate of σ_ε^2 by estimating the fixed-effects model (9.53), because, no matter how m and T may behave as $n \to \infty$, there are n residuals. Thus, if we divide the SSR from (9.53) by $n - m - k$, we obtain an unbiased and consistent estimate of σ_ε^2, since the disturbances for this model are just the ε_{it}. But the natural estimator of σ_v^2, namely, the sample variance of the m elements of $\hat{\boldsymbol{\eta}}$, is not consistent unless $m \to \infty$. In practice, therefore, it is probably undesirable to use the random-effects estimator when m is small.

There is another way to estimate σ_v^2 consistently if $m \to \infty$ as $n \to \infty$. One starts by running the regression

$$\boldsymbol{P_D y} = \boldsymbol{P_D X \beta} + \text{residuals}, \tag{9.57}$$

where $\boldsymbol{P_D} \equiv \mathbf{I} - \boldsymbol{M_D}$, so as to obtain the **between-groups estimator**

$$\hat{\boldsymbol{\beta}}_{\mathrm{BG}} = (\boldsymbol{X}^\top \boldsymbol{P_D X})^{-1} \boldsymbol{X}^\top \boldsymbol{P_D y}. \tag{9.58}$$

Although regression (9.57) appears to have $n = mT$ observations, it really has only m, because the regressand and all the regressors are the same for every observation in each group. The estimator bears the name "between-groups" because it uses only the variation among the group means. If $m < k$, note that the estimator (9.58) does not even exist, since the matrix $\boldsymbol{X}^\top \boldsymbol{P_D X}$ can have rank at most m.

If the restrictions of the random-effects model are not satisfied, the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{BG}}$, if it exists, is in general biased and inconsistent. To see this, observe that, for unbiased estimating equations, we require that

$$\mathrm{E}\big((\boldsymbol{P_D X})_{it}^\top (y_{it} - \boldsymbol{X}_{it}\boldsymbol{\beta})\big) = \mathbf{0}, \tag{9.59}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $(\boldsymbol{P_D X})_{it}$ is the row labelled it of the $n \times k$ matrix $\boldsymbol{P_D X}$. Since $y_{it} - \boldsymbol{X}_{it}\boldsymbol{\beta} = v_i + \varepsilon_{it}$, and since ε_{it} is independent of everything else in condition (9.59), this condition is equivalent to the absence of correlation between the v_i and the elements of the matrix $\boldsymbol{X}$.

As readers are asked to show in Exercise 9.21, the variance of the disturbances in regression (9.57) is $\sigma_v^2 + \sigma_\varepsilon^2/T$. Therefore, if we run it as a regression with m observations, divide the SSR by $m - k$, and then subtract $1/T$ times our estimate of σ_ε^2, we obtain a consistent, but not necessarily positive, estimate of σ_v^2. If the estimate turns out to be negative, we probably should not be estimating an error-components model.

As we will see in the next paragraph, both the OLS estimator of model (9.51) and the feasible GLS estimator of the random-effects model are matrix-weighted averages of the within-groups, or fixed-effects, estimator (9.54) and the between-groups estimator (9.58). For the former to be consistent, we need only the assumptions of the fixed-effects model, but for the latter we need in addition the restrictions of the random-effects model. Thus both the OLS estimator of (9.51) and the feasible GLS estimator are consistent only if the between-groups estimator is consistent.

For the OLS estimator of (9.51),

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\boldsymbol{X}^\top \boldsymbol{M_D y} + \boldsymbol{X}^\top \boldsymbol{P_D y}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M_D X} \hat{\boldsymbol{\beta}}_{\mathrm{FE}} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P_D X} \hat{\boldsymbol{\beta}}_{\mathrm{BG}},\end{aligned}$$

which shows that the estimator is indeed a matrix-weighted average of $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{BG}}$. As readers are asked to show in Exercise 9.22, the GLS estimator of the random-effects model can be obtained by running the OLS regression

$$(\mathbf{I} - \lambda \boldsymbol{P_D}) \boldsymbol{y} = (\mathbf{I} - \lambda \boldsymbol{P_D}) \boldsymbol{X} \boldsymbol{\beta} + \text{residuals}, \tag{9.60}$$

where the scalar λ is defined by

$$\lambda \equiv 1 - \left(\frac{T \sigma_v^2}{\sigma_\varepsilon^2} + 1 \right)^{-1/2}. \tag{9.61}$$

For feasible GLS, we need to replace σ_ε^2 and σ_v^2 by the consistent estimators that were discussed earlier in this subsection.

Equation (9.60) implies that the random-effects GLS estimator is a matrix-weighted average of the OLS estimator for equation (9.51) and the between-groups estimator, and thus also of $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{BG}}$. The GLS estimator is identical to the OLS estimator when $\lambda = 0$, which happens when $\sigma_v^2 = 0$, and equal to the within-groups, or fixed-effects, estimator when $\lambda = 1$, which happens when $\sigma_\varepsilon^2 = 0$. Except in these two special cases, the GLS estimator

Copyright © 2021, Russell Davidson and James G. MacKinnon

is more efficient, in the context of the random-effects model, than either the OLS estimator or the fixed-effects estimator. But equation (9.60) also implies that the random-effects estimator is inconsistent whenever the between-groups estimator is inconsistent.

Unbalanced Panels

Up to this point, we have assumed that we are dealing with a **balanced panel**, that is, a data set for which there are precisely T observations for each cross-sectional unit. However, it is quite common to encounter **unbalanced panels**, for which the number of observations is not the same for every cross-sectional unit. The fixed-effects estimator can be used with unbalanced panels without any real change. It is still based on regression (9.53), and the only change is that the matrix of dummy variables $\boldsymbol{D}$ no longer has the same number of 1s in each column. The random-effects estimator can also be used with unbalanced panels, but it needs to be modified slightly.

Let us assume that the data are grouped by cross-sectional units. Let T_i denote the number of observations associated with unit i, and partition $\boldsymbol{y}$ and $\boldsymbol{X}$ as follows:

$$\boldsymbol{y} = [\boldsymbol{y}_1 \vdots \boldsymbol{y}_2 \vdots \cdots \vdots \boldsymbol{y}_m], \quad \boldsymbol{X} = [\boldsymbol{X}_1 \vdots \boldsymbol{X}_2 \vdots \cdots \vdots \boldsymbol{X}_m],$$

where $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ denote the T_i rows of $\boldsymbol{y}$ and $\boldsymbol{X}$ that correspond to the i^{th} unit. By analogy with (9.61), make the definition

$$\lambda_i \equiv 1 - \left(\frac{T_i \sigma_v^2}{\sigma_\varepsilon^2} + 1 \right)^{-1/2}.$$

Let $\bar{\boldsymbol{y}}_i$ denote a T_i–vector, each element of which is the mean of the elements of $\boldsymbol{y}_i$. Similarly, let $\bar{\boldsymbol{X}}_i$ denote a $T_i \times k$ matrix, each element of which is the mean of the corresponding column of $\boldsymbol{X}_i$. Then the random-effects estimator can be computed by running the linear regression

$$\begin{bmatrix} \boldsymbol{y}_1 - \lambda_1 \bar{\boldsymbol{y}}_1 \\ \boldsymbol{y}_2 - \lambda_2 \bar{\boldsymbol{y}}_2 \\ \vdots \\ \boldsymbol{y}_m - \lambda_m \bar{\boldsymbol{y}}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 - \lambda_1 \bar{\boldsymbol{X}}_1 \\ \boldsymbol{X}_2 - \lambda_2 \bar{\boldsymbol{X}}_2 \\ \vdots \\ \boldsymbol{X}_m - \lambda_m \bar{\boldsymbol{X}}_m \end{bmatrix} \boldsymbol{\beta} + \text{residuals}. \tag{9.62}$$

Note that $\boldsymbol{P_D y}$ is just $[\bar{\boldsymbol{y}}_1 \vdots \bar{\boldsymbol{y}}_2 \vdots \cdots \vdots \bar{\boldsymbol{y}}_m]$, and similarly for $\boldsymbol{P_D X}$. Therefore, since all the λ_i are equal to λ when the panel is balanced, regression (9.62) reduces to regression (9.60) in that special case.

Group Effects and Individual Data

Error-components models are also relevant for regressions on cross-section data with no time dimension, but where the observations naturally belong to

Copyright © 2021, Russell Davidson and James G. MacKinnon

groups. For example, each observation might correspond to a household living in a certain state, and each group would then consist of all the households living in a particular state. In such cases, it is plausible that the disturbances for individuals within the same group are correlated. An error-components model that combines a group-specific disturbance v_i, with variance σ_v^2, and an individual-specific disturbance ε_{it}, with variance σ_ε^2, is a natural way to model this sort of correlation. Such a model implies that the correlation between the disturbances for observations in the same group is $\rho \equiv \sigma_v^2/(\sigma_v^2 + \sigma_\varepsilon^2)$ and the correlation between the disturbances for observations in different groups is zero.

A fixed-effects model is often unsatisfactory for dealing with group effects. In many cases, some explanatory variables are observed only at the group level, so that they have no within-group variation. Such variables are perfectly collinear with the group dummies used in estimating a fixed-effects model, making it impossible to identify the parameters associated with them. On the other hand, they are identified by a random-effects model for an unbalanced panel, because this model takes account of between-group variation. This can be seen from equation (9.62): Collinearity of the transformed group-level variables on the right-hand side occurs only if the explanatory variables are collinear to begin with. The estimates of σ_ε^2 and σ_v^2 needed to compute the λ_i may be obtained in various ways, some of which were discussed in the subsection on random-effects estimation. As we remarked there, these work well only if the number of groups m is not too small.

If it is thought that the within-group correlation ρ is small, it may be tempting to ignore it and use OLS estimation, with the usual OLS covariance matrix. This can be a serious mistake unless ρ is actually zero, since the OLS standard errors can be drastic underestimates even with small values of ρ, as Kloek (1981) and Moulton (1986, 1990) have pointed out; recall Section 6.6 on clustered data. The problem is particularly severe when the number of observations per group is large, as readers are asked to show in Exercise 9.23. The correlation of the disturbances within groups means that the effective sample size is much smaller than the actual sample size when there are many observations per group.

In this section, we have presented just a few of the most basic ideas concerning estimation with panel data. Of course, GLS is not the only method that can be used to estimate models for data of this type. For more detailed treatments of various models for panel data, see, among others, Chamberlain (1984), Hsiao (1986, 2001), Ruud (2000, Chapter 24), Baltagi (2001), Arellano and Honoré (2001), Greene (2002, Chapter 14), and Wooldridge (2002).

9.11 Final Remarks

Several important concepts were introduced in the first four sections of this chapter, which dealt with the basic theory of generalized least squares estimation. The concept of an efficient estimator, which we introduced in Section 9.2, is important for many more models than just linear regression. The key idea of feasible GLS estimation is that an unknown covariance matrix may in some circumstances be replaced by a consistent estimate of that matrix without changing the asymptotic properties of the resulting estimator.

The remainder of the chapter dealt with the treatment of heteroskedasticity and serial correlation in linear regression models, and with error-components models for panel data. Although this material is of considerable practical importance, most of the techniques we discussed, although sometimes complicated in detail, are conceptually straightforward applications of feasible GLS estimation and the methods for testing hypotheses that were introduced in Chapters 5 and 8.

Copyright © 2021, Russell Davidson and James G. MacKinnon

9.12 Exercises

9.1 Using the fact that $\mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top \mid \boldsymbol{X}) = \boldsymbol{\Omega}$ for regression (9.01), show directly, without appeal to standard OLS results, that the covariance matrix of the GLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is given by the rightmost expression of (9.05).

9.2 Show that the matrix (9.11), reproduced here for easy reference,

$$\boldsymbol{X}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{X} - \boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{\Omega} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{X},$$

is positive semidefinite. As in Section 7.2, this may be done by showing that this matrix can be expressed in the form $\boldsymbol{Z}^\top \boldsymbol{M} \boldsymbol{Z}$, for some $n \times k$ matrix $\boldsymbol{Z}$ and some $n \times n$ orthogonal projection matrix $\boldsymbol{M}$. It is helpful to express $\boldsymbol{\Omega}^{-1}$ as $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, as in equation (9.02).

9.3 Using the data in the file **earnings.data**, run the regression

$$y_t = \beta_1 d_{1t} + \beta_2 d_{2t} + \beta_3 d_{3t} + u_t,$$

which was previously estimated in Exercise 3.23. Recall that the d_{it} are dummy variables. Then test the null hypothesis that $\mathrm{E}(u_t^2) = \sigma^2$ against the alternative that

$$\mathrm{E}(u_t^2) = \gamma_1 d_{1t} + \gamma_2 d_{2t} + \gamma_3 d_{3t}.$$

Report P values for F and nR^2 tests.

9.4 If u_t follows the stationary AR(1) process

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{IID}(0, \sigma_\varepsilon^2), \quad |\rho| < 1,$$

show that $\mathrm{Cov}(u_t, u_{t-j}) = \mathrm{Cov}(u_t, u_{t+j}) = \rho^j \sigma_\varepsilon^2/(1-\rho^2)$. Then use this result to show that the correlation between u_t and u_{t-j} is just ρ^j.

9.5 Consider the nonlinear regression model $y_t = x_t(\boldsymbol{\beta}) + u_t$. Derive the GNR for testing the null hypothesis that the u_t are serially uncorrelated against the alternative that they follow an AR(1) process.

9.6 Show how to test the null hypothesis that the disturbances of the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ are serially uncorrelated against the alternative

Copyright © 2021, Russell Davidson and James G. MacKinnon

that they follow an AR(4) process by means of a linearized regression. Derive the test regression from first principles.

9.7 Consider the following three models, where u_t is assumed to be IID$(0, \sigma^2)$:

$$\begin{aligned} H_0\colon &\quad y_t = \beta + u_t \\ H_1\colon &\quad y_t = \beta + \rho(y_{t-1} - \beta) + u_t \\ H_2\colon &\quad y_t = \beta + u_t + \alpha u_{t-1} \end{aligned}$$

Explain how to test H_0 against H_1 by using a linearized regression. Then show that exactly the same test statistic is also appropriate for testing H_0 against H_2.

9.8 Write the trace in the right-hand side of equation xxx explicitly in terms of $\boldsymbol{P_X}$ rather than $\boldsymbol{M_X}$, and show that the terms containing one or more factors of $\boldsymbol{P_X}$ all vanish asymptotically.

9.9 By direct matrix multiplication, show that, if $\boldsymbol{\Psi}$ is given by (9.46), then $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$ is equal to the matrix

$$\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

Show further, by direct calculation, that this matrix is proportional to the inverse of the matrix $\boldsymbol{\Omega}$ given in equation (9.27).

9.10 Show that equation (9.25), relating $\boldsymbol{u}$ to $\boldsymbol{\varepsilon}$, can be modified to take account of the definition (9.45) of ε_1, with the result that

$$u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \cdots + \frac{\rho^{t-1}}{(1-\rho^2)^{1/2}}\varepsilon_1. \tag{9.63}$$

The relation $\boldsymbol{\Psi}^\top\boldsymbol{u} = \boldsymbol{\varepsilon}$ implies that $\boldsymbol{u} = (\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\varepsilon}$. Use the result (9.63) to show that $\boldsymbol{\Psi}^{-1}$ can be written as

$$\begin{bmatrix} \theta & \rho\theta & \rho^2\theta & \cdots & \rho^{n-1}\theta \\ 0 & 1 & \rho & \cdots & \rho^{n-2} \\ 0 & 0 & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

where $\theta \equiv (1-\rho^2)^{-1/2}$. Verify by direct calculation that this matrix is the inverse of the $\boldsymbol{\Psi}$ given by (9.46).

9.11 Consider a square, symmetric, nonsingular matrix partitioned as follows

$$\boldsymbol{H} \equiv \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}^\top \\ \boldsymbol{C} & \boldsymbol{B} \end{bmatrix}, \tag{9.64}$$

Copyright © 2021, Russell Davidson and James G. MacKinnon

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are also square symmetric nonsingular matrices. By using the rules for multiplying partitioned matrices (see Section 1.4), show that $\boldsymbol{H}^{-1}$ can be expressed in partitioned form as

$$\boldsymbol{H}^{-1} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{E}^\top \\ \boldsymbol{E} & \boldsymbol{F} \end{bmatrix},$$

where

$$\begin{aligned} \boldsymbol{D} &= (\boldsymbol{A} - \boldsymbol{C}^\top\boldsymbol{B}^{-1}\boldsymbol{C})^{-1}, \\ \boldsymbol{E} &= -\boldsymbol{B}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{C}^\top\boldsymbol{B}^{-1}\boldsymbol{C})^{-1} = -(\boldsymbol{B} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}^\top)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}, \text{ and} \\ \boldsymbol{F} &= (\boldsymbol{B} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{C}^\top)^{-1}. \end{aligned}$$

9.12 Suppose that the matrix $\boldsymbol{H}$ of the previous question is positive definite. It therefore follows (see Section 3.4) that there exists a square matrix $\boldsymbol{X}$ such that $\boldsymbol{H} = \boldsymbol{X}^\top\boldsymbol{X}$. Partition $\boldsymbol{X}$ as $[\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, so that

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2 \end{bmatrix},$$

where the blocks of the matrix on the right-hand side are the same as the blocks in (9.64). Show that the top left block $\boldsymbol{D}$ of $\boldsymbol{H}^{-1}$ can be expressed as $(\boldsymbol{X}_1^\top\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}$, where $\boldsymbol{M}_2 \equiv \mathbf{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top$. Use this result to show that $\boldsymbol{D} - \boldsymbol{A}^{-1} = (\boldsymbol{X}_1^\top\boldsymbol{M}_2\boldsymbol{X}_1)^{-1} - (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}$ is a positive semidefinite matrix.

⋆9.13 Consider testing for first-order serial correlation of the disturbances in the regression model

$$\boldsymbol{y} = \beta\boldsymbol{y}_1 + \boldsymbol{u}, \quad |\beta| < 1, \tag{9.65}$$

where $\boldsymbol{y}_1$ is the vector with typical element y_{t-1}, by use of the statistics t_{GNR} and t_{SR} defined in xxx and xxx, respectively. Show first that the vector denoted as $\boldsymbol{M_X}\tilde{\boldsymbol{u}}_1$ in xxx and xxx is equal to $-\tilde{\beta}\boldsymbol{M_X}\boldsymbol{y}_2$, where $\boldsymbol{y}_2$ is the vector with typical element y_{t-2}, and $\tilde{\beta}$ is the OLS estimate of β from (9.65). Then show that, as $n \to \infty$, t_{GNR} tends to the random variable $\tau \equiv \sigma_u^{-2}\,\text{plim}\, n^{-1/2}(\beta\boldsymbol{y}_1 - \boldsymbol{y}_2)^\top\boldsymbol{u}$, whereas t_{SR} tends to the same random variable times β. Show finally that t_{GNR}, but not t_{SR}, provides an asymptotically correct test, by showing that the random variable τ is asymptotically distributed as N$(0, 1)$.

9.14 The file **money.data** contains seasonally adjusted quarterly data for the logarithm of the real money supply, m_t, real GDP, y_t, and the 3-month Treasury Bill rate, r_t, for Canada for the period 1967:1 to 1998:4. A conventional demand for money function is

$$m_t = \beta_1 + \beta_2 r_t + \beta_3 y_t + \beta_4 m_{t-1} + u_t. \tag{9.66}$$

Estimate this model over the period 1968:1 to 1998:4, and then test it for AR(1) disturbances using two different testing regressions that differ in their treatment of the first observation.

Copyright © 2021, Russell Davidson and James G. MacKinnon

⋆**9.15** The algorithm called **iterated Cochrane-Orcutt**, alluded to in Section 8.8, is just iterated feasible GLS without the first observation. This algorithm is begun by running the regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ by OLS, preferably omitting observation 1, in order to obtain the first estimate of $\boldsymbol{\beta}$. The residuals from this equation are then used to estimate ρ according to equation xxx. What is the next step in this procedure? Complete the description of iterated Cochrane-Orcutt as iterated feasible GLS, showing how each step of the procedure can be carried out using an OLS regression.

Show that, when the algorithm converges, conditions xxx for NLS estimation are satisfied. Also show that, unlike iterated feasible GLS including observation 1, this algorithm *must* eventually converge, although perhaps only to a local, rather than the global, minimum of $\mathrm{SSR}(\boldsymbol{\beta}, \rho)$.

9.16 Estimate this model using the iterated Cochrane-Orcutt algorithm, using a sequence of OLS regressions, and see how many iterations are needed to achieve the same estimates as those achieved by NLS. Compare this number with the number of iterations used by NLS itself.

Repeat the exercise with a starting value of 0.5 for ρ instead of the value of 0 that is conventionally used.

9.17 Test the hypothesis that the disturbances of the linear regression model (9.66) are serially uncorrelated against the alternatives that they follow the simple AR(4) process $u_t = \rho_4 u_{t-4} + \varepsilon_t$ and that they follow a general AR(4) process.

9.18 Consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{9.67}$$

where there are n observations, and k_0, k_1, and k_2 denote the numbers of parameters in $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$, respectively. Let H_0 denote the hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{0}$ and $\boldsymbol{\beta}_2 = \boldsymbol{0}$, H_1 denote the hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, and H_2 denote the model (9.67) with no restrictions.

Show that the F statistics for testing H_0 against H_1 and for testing H_1 against H_2 are asymptotically independent of each other.

9.19 This question uses data on daily returns for the period 1989–1998 for shares of Mobil Corporation from the file **daily-crsp.data**. These data are made available by courtesy of the Center for Research in Security Prices (CRSP); see the comments at the bottom of the file. Regress these returns on a constant and themselves lagged once, twice, three, and four times, dropping the first four observations. Then test the null hypothesis that all coefficients except the constant term are equal to zero, as they should be if market prices fully reflect all available information. Be sure to report a P values for the test.

9.20 Consider the fixed-effects model (9.53). Show that, under mild regularity conditions, which you should specify, the OLS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ tends in probability to the true parameter vector $\boldsymbol{\beta}_0$ as m, the number of cross-sectional units, tends to infinity, while T, the number of time periods, remains fixed.

9.21 Suppose that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v} + \boldsymbol{\varepsilon}, \tag{9.68}$$

where there are $n = mT$ observations, $\boldsymbol{y}$ is an n–vector with typical element y_{it}, $\boldsymbol{X}$ is an $n \times k$ matrix with typical row $\boldsymbol{X}_{it}$, $\boldsymbol{\varepsilon}$ is an n–vector with typical element ε_{it}, and $\boldsymbol{v}$ is an n–vector with v_i repeated in the positions that correspond to y_{i1} through y_{iT}. Let the v_i have variance σ_v^2 and the ε_{it} have variance σ_ε^2. Given these assumptions, show that the variance of the disturbances in regression (9.57) is $\sigma_v^2 + \sigma_\varepsilon^2/T$.

Copyright © 2021, Russell Davidson and James G. MacKinnon

⋆**9.22** Show that, for $\boldsymbol{\Sigma}$ defined in (9.56),

$$\boldsymbol{\Sigma}^{-1/2} = \frac{1}{\sigma_\varepsilon}(\mathbf{I}_T - \lambda\boldsymbol{P}_{\boldsymbol{\iota}}),$$

where $\boldsymbol{P}_{\boldsymbol{\iota}} \equiv \boldsymbol{\iota}(\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^\top = (1/T)\boldsymbol{\iota}\boldsymbol{\iota}^\top$, and

$$\lambda = 1 - \left(\frac{T\sigma_v^2}{\sigma_\varepsilon^2} + 1\right)^{-1/2}.$$

Then use this result to show that the GLS estimates of $\boldsymbol{\beta}$ may be obtained by running regression (9.60). What is the covariance matrix of the GLS estimator?

⋆**9.23** Suppose that, in the error-components model (9.68), none of the columns of $\boldsymbol{X}$ displays any within-group variation. Recall that, for this model, the data are balanced, with m groups and T observations per group. Show that the OLS and GLS estimators are identical in this special case. Then write down the true covariance matrix of both these estimators. How is this covariance matrix related to the usual one for OLS that would be computed by a regression package under classical assumptions? What happens to this relationship as T and ρ, the correlation of the disturbances within groups, change?

Copyright © 2021, Russell Davidson and James G. MacKinnon

References

Amemiya, T. (1973a). "Generalized least squares with an estimated autocovariance matrix," *Econometrica*, **41**, 723–32.

Andrews, D. W. K. (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, **59**, 817–58.

Andrews, D. W. K. (2004). "The block-block bootstrap: Improved asymptotic refinements", *Econometrica*, **72**, 673–700.

Andrews, D. W. K. (2005). "Cross-section regression with common shocks," *Econometrica*, **73**, 1551–85.

Anglin, P. A., and R. Gençay (1996). "Semiparametric estimation of a hedonic price function," *Journal of Applied Econometrics*, **11**, 633–48.

Angrist, J. D., and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, Princeton University Press.

Arellano, M., and S. Bond (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, **58**, 277–97.

Arellano, M., and O. Bover (1995). "Another look at instrumental variable estimation of error-components models," *Journal of Econometrics*, **68**, 29–51.

Arellano, M., and B. Honoré (2001). "Panel data models: Some recent developments," Ch. 53 in *Handbook of Econometrics*, Vol. 5, ed. J. J. Heckman and E. E. Leamer, Amsterdam, North-Holland.

Athey, S. and Imbens, G. W. (2006). "Identification and Inference in Nonlinear difference-in-differences models", *Econometrica*, **74**, 431–97.

Bahadur, R. R. and L. J. Savage (1956). "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems", *Annals of Statistics*, 27, 1115–22.

Balestra, P., and M. Nerlove (1966). "Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas," *Econometrica*, **34**, 585–612.

Baltagi, B. (2001). *Econometric Analysis of Panel Data*, second edition, New York, John Wiley & Sons.

Basmann, R. L. (1957). "A generalized classical method of linear estimation of coefficients in a structural equation," *Econometrica*, **25**, 77–83.

Bell, R. M., and D. F. McCaffrey (2002). "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology* 28, 169–81.

Benjamini, Y., and Y. Hochberg (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, Series B, **57**, 299–300.

Benjamini, Y., and D. Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, **29**, 1165–88.

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, **83**, 687–97.

Berkowitz, J., and L. Kilian (2000). "Recent developments in bootstrapping time series", *Econometric Reviews*, **19**, 1–48.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). "How much should we trust differences-in-differences estimates?," *Quarterly Journal of Economics*, **119**, 249–75.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, **165**, 137–51.

Billingsley, P. (1995). *Probability and Measure*, third edition, New York, John Wiley & Sons.

Box, G. E. P., and G. M. Jenkins (1976). *Time Series Analysis, Forecasting and Control*, revised edition, San Francisco, Holden Day.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis, Forecasting and Control*, third edition, Englewood Cliffs, N.J., Prentice-Hall.

Breusch, T. S. (1978). "Testing for autocorrelation in dynamic linear models," *Australian Economic Papers*, **17**, 334–55.

Bryant, P. (1984). "Geometry, statistics, probability: Variations on a common theme," *The American Statistician*, **38**, 38–48.

Bühlmann, P. (1997). "Sieve bootstrap for time series", *Bernoulli*, **3**, 123–48.

Bühlmann, P. (2002). "Bootstraps for time series", *Statistical Science*, **17**, 52–72.

Cameron, A. C., and D. L. Miller (2015). "A practitioner's guide to cluster robust inference," *Journal of Human Resources*, **50**, 317–72.

Card, D., and A. B. Krueger (1994). "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review*, **84**, 772–93.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2015. "Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity," University of California, Santa Barbara, working paper.

Chamberlain, G. (1984). "Panel data," Ch. 22 in *Handbook of Econometrics*, Vol. 2, ed. Z. Griliches and M. D. Intriligator, Amsterdam, North-Holland, 1247–318.

Chang, Y., and J. Y. Park (2003). "A sieve bootstrap for the test of a unit root", *Journal of Time Series Analysis*, **24**, 379–400.

Chesher, A. (1989). "Hájek inequalities, measures of leverage and the size of heteroskedasticity robust tests," *Econometrica*, **57**, 971–77.

Chesher, A., and G. Austin (1991). "The finite-sample distributions of heteroskedasticity robust Wald statistics," *Journal of Econometrics*, **47**, 153–73.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Choi, E., and P. Hall, (2000). "Bootstrap confidence regions computed from autoregressions of arbitrary order", *Journal of the Royal Statistical Society* Series B, **62**, 461–77.

Chow, G. C. (1960). "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, **28**, 591–605.

Clements, M. P., and D. F. Hendry (2002). *A Companion to Economic Forecasting*, Oxford, Blackwell Publishers.

Cochrane, D., and G. H. Orcutt (1949). "Application of least squares regression to relationships containing autocorrelated error terms," *Journal of the American Statistical Association*, **44**, 32–61.

Dagenais, M. G., and D. L. Dagenais (1997). "Higher moment estimators for linear regression models with errors in the variables," *Journal of Econometrics*, **76**, 193–221.

Danilov, D., and J. R. Magnus (2004). "On the harm that ignoring pretesting can cause," *Journal of Econometrics*, **122**, 27–46.

Das Gupta, S., and M. D Perlman (1974). "Power of the noncentral F test: Effect of additional variates on Hotelling's T^2 test," *Journal of the American Statistical Association*, **69**, 174–80.

Davidson, R., E. Flachaire (2008). "The wild bootstrap, tamed at last", *Journal of Econometrics*, **146**, 162–9

Davidson, R., and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, Oxford University Press.

Davidson, R., and J. G. MacKinnon (1999a). "Bootstrap testing in nonlinear models," *International Economic Review*, **40**, 487–508.

Davidson, R., and J. G. MacKinnon (1999b). "The size distortion of bootstrap tests," *Econometric Theory*, **15**, 361–76.

Davidson, R., and J. G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?," *Econometric Reviews*, **19**, 55–68.

Davidson, R., and J. G. MacKinnon (2002a). "Bootstrap J tests of nonnested linear regression models," *Journal of Econometrics*, **109**, 167–93.

Davidson, R., and J. G. MacKinnon (2004). *Econometric Theory and Methods*, New York, Oxford University Press. (ETM)

Davidson, R., and J. G. MacKinnon (2010). "Wild Bootstrap Tests for IV Regression", *Journal of Business and Economic Statistics*, **28**, 128–44.

Davidson, J. E. H. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford, Oxford University Press.

Davison, A. C., and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*, Cambridge, Cambridge University Press.

Dennett, D. C. (2003). *Freedom Evolves*, the Penguin Group.

Deutsch, D. (1997). *The Fabric of Reality*, Penguin Books.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Devroye, L. (1986). *Non-Uniform Random Variate Generation*, Springer-Verlag, New York. But see: **http://luc.devroye.org/rnbookindex.html**

DiCiccio, T. J., and B. Efron (1996). "Bootstrap confidence intervals" (with discussion), *Statistical Science*, **11**, 189–228.

Donald, S. G., and K. Lang (2007). "Inference with difference-in-differences and other panel data." *Review of Economics and Statistics*, **89**, 221–33.

Dufour, J.-M. (1982). "Generalized Chow tests for structural change: A coordinate-free approach," *International Economic Review*, **23**, 565–75.

Dufour, J.-M. (1997). "Some impossibility theorems in econometrics with application to structural and dynamic models," *Econometrica*, **65**, 1365–88.

Dufour, J.-M., and L. Khalaf (2001). "Monte Carlo test methods in econometrics," Ch. 23 in *A Companion to Theoretical Econometrics*, ed. B. Baltagi, Oxford, Blackwell Publishers, 494–519.

Durbin, J. (1954). "Errors in variables," *Review of the International Statistical Institute*, **22**, 23–32.

Durbin, J. (1970). "Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables," *Econometrica*, **38**, 410–21.

Durbin, J., and G. S. Watson (1950). "Testing for serial correlation in least squares regression I," *Biometrika*, **37**, 409–28.

Dwass, M., 1957. "Modified randomization tests for nonparametric hypotheses," *Annals of Mathematical Statistics*, **28**, 181–7.

Efron, B. (1979). "Bootstrapping methods: Another look at the jackknife," *Annals of Statistics*, **7**, 1–26.

Efron, B., and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*, New York, Chapman and Hall.

Eicker, F. (1963). "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, **34**, 447–56.

Eicker, F. (1967). "Limit theorems for regressions with unequal and dependent errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. L. M. Le Cam and J. Neyman, Berkeley, University of California, **1**, 59–82.

Elliott, G., and A. Timmermann (2008). "Economic forecasting," *Journal of Economic Literature*, **46**, 3–56.

Elliott, G., and A. Timmermann (2016). *Economic Forecasting*, Princeton, Princeton University Press.

Fisher, F. M. (1970). "Tests of equality between sets of coefficients in two linear regressions: An expository note," *Econometrica*, **38**, 361–66.

Flachaire, E., 1999. A better way to bootstrap pairs. *Economics Letters* **64**, 257–262.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Freedman, D. (1981). "Bootstrapping regression models," *Annals of Statistics*, **9**, 1218–28.

Friedman, M. (1957). *A Theory of the Consumption Function*, Princeton, Princeton University Press.

Frisch, R. (1934). "Statistical Confluence Analysis by Means of Complete Regression Systems," Publikasjon nr. 5. Universitetets Økonomiske Institutt, Oslo.

Frisch, R., and F. V. Waugh (1933). "Partial time regressions as compared with individual trends," *Econometrica*, **1**, 387–401.

Fuller, W. A., and G. E. Battese (1974). "Estimation of linear models with crossed-error structure," *Journal of Econometrics*, **2**, 67–78.

Gallant, A. R. (1997). *An Introduction to Econometric Theory*, Princeton, Princeton University Press.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods*, New York, Springer-Verlag.

Ghysels, E., and D. R. Osborn (2001). *The Econometric Analysis of Seasonal Time Series*, Cambridge, Cambridge University Press.

Godfrey, L. G. (1978a). "Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables," *Econometrica*, **46**, 1293–301.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics*, Cambridge, Cambridge University Press.

Godfrey, L. G. (1998). "Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap," *Journal of Econometrics*, **84**, 59–74.

Gonçalves, S., and L. Kilian (2004). "Bootstrapping autoregressions with heteroskedasticity of unknown form," *Journal of Econometrics*, **123**, 89–120.

Granger, C. W. J., and P. Newbold (1974). "Spurious regressions in econometrics," *Journal of Econometrics*, **2**, 111–20.

Greene, W. H. (2002). *Econometric Analysis*, fifth edition, New York, Prentice-Hall.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York, Springer-Verlag.

Hall, P., J. L. Horowitz, and B.-Y. Jing (1995). "On blocking rules for the bootstrap with dependent data", *Biometrika* **82**, 561–74.

Hall, P., and S. R. Wilson (1991). "Two guidelines for bootstrap hypothesis testing," *Biometrics*, **47**, 757–62.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton, Princeton University Press.

Hansen, L. P. (1982). "Large sample properties of generalized method of moments estimators," *Econometrica*, **50**, 1029–54.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge, Cambridge University Press.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Härdle, W., J. L. Horowitz, and J. P. Kreiss (2003). "Bootstrap methods for time series", *International Statistical Review* **71**, 435-459.

Hausman, J. A. (1978). "Specification tests in econometrics," *Econometrica*, **46**, 1251–72.

Hausman, J. A., and M. W. Watson (1985). "Errors-in-variables and seasonal adjustment procedures," *Journal of the American Statistical Association*, **80**, 531–40.

Hayashi, F. (2000). *Econometrics*, Princeton, Princeton University Press.

Heckman, J. J. (2001). "Econometrics and Empirical Economics", *Journal of Econometrics*, 100, 3-5.

Hendry, D. F. (1995). *Dynamic Econometrics*, Oxford, Oxford University Press.

Herr, D. G. (1980). "On the history of the use of geometry in the general linear model," *The American Statistician*, **34**, 43–7.

Hinkley, D. V. (1977). "Jackknifing in unbalanced situations," *Technometrics*, **19**, 285–92.

Hochberg, Y. (1988). "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, **75**, 800–2.

Hogg, R. V., J. McKean, and A. T. Craig (2007). *Introduction to Mathematical Statistics*, seventh edition, New York, Pearson.

Horowitz, J. L. (1994). "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, **61**, 395–411.

Horowitz, J. L. (2001). "The Bootstrap," Ch. 52 in *Handbook of Econometrics*, Vol. 5, ed. J. J. Heckman and E. E. Leamer, Amsterdam, North-Holland.

Horowitz, J. L. (2003). "The bootstrap in econometrics", *Statistical Science*, **18**, 211–8.

Hsiao, C. (1986). *Analysis of Panel Data*, Cambridge, Cambridge University Press.

Hsiao, C. (2001). "Panel data models," Ch. 16 in *A Companion to Theoretical Econometrics*, ed. B. Baltagi, Oxford, Blackwell Publishers, 349–65.

Huber, P. J. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. L. M. Le Cam and J. Neyman, Berkeley, University of California, **1**, 221–33.

Hylleberg, S. (1986). *Seasonality in Regression*, New York, Academic Press.

Imbens, G. W., and Michal Kolesár (2016). "Robust standard errors in small samples: Some practical advice," *Review of Economics and Statistics*, forthcoming.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*, Palgrave Macmillan.

Kinal, T. W. (1980). "The existence of moments of k-class estimators," *Econometrica*, **48**, 241–49.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Kloek, T. (1981). "OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated," *Econometrica*, **49**, 205–7.

Knuth, Donald E. (1998). *The Art of Computer Programming*, Vol. 2, *Seminumerical Algorithms*, third edition, Reading, Mass., Addison-Wesley.

Künsch, H. R. (1989). "The jackknife and the bootstrap for general stationary observations", *Annals of Statistics*, **17**, 1217–41.

Lahiri, S. N. (1999). "Theoretical comparisons of block bootstrap methods", *Annals of Statistics* **27**, 386–404.

Leamer, E. E. (1987). "Errors in variables in linear systems," *Econometrica*, **55**, 893–909.

L'Ecuyer, P. (2012). "Random number generation," in *Handbook of Computational Statistics: Concepts and Methods*, ed. J. E. Gentle, W. K. Härdle, and Yuichi Mori, New York, Springer, 35–71.

Li, H., and G. S. Maddala (1996). "Bootstrapping time series models" (with discussion), *Econometric Reviews*, **15**, 115–95.

Liang, K.-Y., and S. L. Zeger (1986). "Longitudinal data analysis using generalized linear models." *Biometrika* **73**, 13–22.

Liu, R. Y., 1988. "Bootstrap procedures under some non-I.I.D. models," *Annals of Statistics*, **16**, 1696–708.

Long, J. S., and L. H. Ervin (2000). "Using heteroscedasticity consistent standard errors in the linear regression model," *The American Statistician*, **54**, 217–24.

Lovell, M. C. (1963). "Seasonal adjustment of economic time series and multiple regression analysis," *Journal of the American Statistical Association*, **58**, 993–1010.

MacKinnon, J. G. (2006). "Bootstrap Methods in Econometrics", *The Economic Record*, The Economic Society of Australia, vol. 82(s1),2–18

MacKinnon, J. G. (2012). "Thirty years of heteroskedasticity-robust inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. Xiaohong Chen and Norman R. Swanson, New York, Springer, 437–461.

MacKinnon, J. G. (2015). "Wild cluster bootstrap confidence intervals," *L'Actualité Économique*, **91**, 11–33

MacKinnon, J. G., and M. D. Webb (2016). "Wild bootstrap inference for wildly different cluster sizes," *Journal of Applied Econometrics*, forthcoming.

MacKinnon, J. G., and H. White (1985). "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, **29**, 305–25.

Magnus, J. R. and J. Durbin (1999). "Estimation of regression coefficients on interest when other regression coefficients are of no interest," *Econometrica*, **67**, 639–643.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Mammen, E., 1993. "Bootstrap and wild bootstrap for high dimensional linear models", *Annals of Statistics*, **21**, 255–85.

Matsumoto, M. and T. Nishimura (1998). "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator", *ACM Transactions on Modeling and Computer Simulation* **8** (1): 3–330.

Mittelhammer, R. (2013). *Mathematical Statistics for Economics and Business*, second edition, New York, Springer.

Morgan, M. S. (1990). *The History of Econometric Ideas*, Cambridge, Cambridge University Press.

Moulton, B. R. (1986). "Random group effects and the precision of regression estimates," *Journal of Econometrics*, **32**, 385–97.

Moulton, B. R. (1990). "An illustration of a pitfall in estimating the effects of aggregate variables on micro units," *Review of Economics and Statistics*, **72**, 334–8.

Mundlak, Y. (1978). "On the pooling of time series and cross sectional data," *Econometrica*, **46**, 69–86.

Newey, W. K., and K. D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, **55**, 703–8.

Newey, W. K., and K. D. West (1994). "Automatic lag selection in covariance matrix estimation," *Review of Economic Studies*, **61**, 631–53.

Park, J. Y. (2002). "An invariance principle for sieve bootstrap in time series", *Econometric Theory*, **18**, 469–90.

Park, J. Y. (2003). "Bootstrap unit root tests", *Econometrica* **71**, 1845–95.

Pesaran, M. H. (2015). *Time Series and Panel Data Econometrics*, Oxford, Oxford University Press.

Politis, D. N. (2003). "The impact of bootstrap methods on time series analysis", *Statistical Science*, **18**, 219–30.

Politis, D. N., and J. P. Romano (1994). "The stationary bootstrap", *Journal of the American Statistical Association*, **89**, 1303–13.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2007). *Numerical Recipes: The Art of Scientific Computing*, third edition, Cambridge, Cambridge University Press.

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*, New York, Oxford University Press.

Sargan, J. D. (1958). "The estimation of economic relationships using instrumental variables," *Econometrica*, **26**, 393–415.

Savin, N. E., and K. J. White (1977). "The Durbin-Watson test for serial correlation with extreme sample sizes or many regressors," *Econometrica*, **45**, 1989–96.

Shao, J. (2007). *Mathematical Statistics*, second edition, New York, Springer.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Schervish, M. J. (1996). *Theory of Statistics*, New York, Springer.

Seber, G. A. F. (1980). *The Linear Hypothesis: A General Theory*, second edition, London, Charles Griffin.

Simes, R. J. (1986). "An improved Bonferroni procedure for multiple tests of significance," *Biometrika*, **73**, 751–4.

Snedecor, G. W. (1934). *Calculation and Interpretation of Analysis of Variance and Covariance*, Ames, Iowa, Collegiate Press.

Theil, H. (1953). "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Cambridge, Cambridge University Press.

White, H. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, **48**, 817–38.

White, H. (2000). *Asymptotic Theory for Econometricians*, revised edition, Orlando, Academic Press.

White, H., and I. Domowitz (1984). "Nonlinear regression with dependent observations," *Econometrica*, **52**, 143–61.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Mass., MIT Press.

Wu, C. F. J. (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis", *Annals of Statistics*, **14**, 1261–95.

Wu, D.-M. (1973). "Alternative tests of independence between stochastic regressors and disturbances," *Econometrica*, **41**, 733–50.

Young, A. (2015). "Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections," working paper, London School of Economics.

Copyright © 2021, Russell Davidson and James G. MacKinnon

Author Index

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Subject Index

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon

Copyright © 2021, Russell Davidson and James G. MacKinnon