Bootstrapping Econometric Models

by

Russell Davidson

Department of Economics and CIREQ
McGill University
Montréal, Québec, Canada
H3A 2T7

GREQAM
Centre de la Vieille Charité
2 Rue de la Charité
13236 Marseille cedex 02, France

Abstract

The bootstrap is a statistical technique used more and more widely in econometrics. While it is capable of yielding very reliable inference, some precautions should be taken in order to ensure this. Two "Golden Rules" are formulated that, if observed, help to obtain the best the bootstrap can offer. Bootstrapping always involves setting up a bootstrap data-generating process (DGP). The main types of bootstrap DGP in current use are discussed, with examples of their use in econometrics. The ways in which the bootstrap can be used to construct confidence sets differ somewhat from methods of hypothesis testing. The relation between the two sorts of problem is discussed.

Keywords: Bootstrap, hypothesis test, confidence set

JEL codes: C100, C120, C150

This work was supported by the Canada Research Chair program (Chair in Economics, McGill University) and by grants from the Social Sciences and Humanities Research Council of Canada, and the Fonds Québécois de Recherche sur la Société et la Culture. I am much indebted to James G. MacKinnon for many valuable discussions.

June 2007

1. Introduction

The bootstrap is a statistical technique that is most often implemented by simulation. Simulation is not an essential element of the bootstrap, although in practice only trivial uses do not require simulation. The basic idea of bootstrap testing is that, when a test statistic of interest has an unknown distribution under the null hypothesis under test, that distribution can be characterised by using information in the data set that is being analysed.

The simplest case arises when the statistic is *pivotal* for the null hypothesis. This means that the distribution of the statistic is the same whatever may be the data-generating process (DGP) of the statistic, provided only that this DGP satisfies the null hypothesis. If we denote the set of DGPs that satisfy the null by $\mathbb{M}$, then, when the statistic is pivotal, any procedure that gives the distribution under any DGP in $\mathbb{M}$ can be used to obtain information about the distribution. We can think of the set $\mathbb{M}$ as constituting a *model*, and of the null hypothesis as stating that this model is *well specified*, by which is meant that the true, unknown, DGP that generated the data under analysis belongs to $\mathbb{M}$.

The procedure most likely to be useful for finding the null distribution of the statistic is simulation. One generates many artificial data sets from whatever DGP in $\mathbb{M}$ makes for the simplest kind of simulation, and, for each of these data sets, usually called *bootstrap samples*, one computes a realisation of the statistic. The empirical distribution function (EDF) of these *bootstrap statistics* is then used as a simulation-based estimate of the unknown distribution.

If the distribution of a statistic under the null hypothesis is known, then statistical inference of various sorts becomes possible. The most commonly used types of inference are based on *critical values* or on *P values*. The former are defined as quantiles of the null distribution, determined as a function of the desired significance level of the test. The latter are marginal significance levels, that is, the levels at which the test is at the margin between rejection and non-rejection of the null hypothesis.

Specifically, if the null is to be rejected when the realised statistic is too large, then, for test at level α, the critical value is the $(1 - \alpha)$–quantile of the null distribution. For a realisation τ of the statistic, the associated P value in this case is $1 - F(\tau)$, where F is the cumulative distribution function (CDF) of the null distribution. For tests that reject for small values of the statistic the critical value is the α–quantile and the P value is $F(\tau)$. For two-tailed tests, two critical values are needed, a lower and an upper. The former is usually chosen as the $\alpha/2$–quantile and the latter as the $(1 - \alpha/2)$–quantile. The P value for a realisation τ is $2\min\big(F(\tau), 1 - F(\tau)\big)$.

There are other ways of constructing critical values for two-tailed tests. If one uses the β and γ–quantiles for the lower and upper critical values respectively, it is enough that $1 - \gamma + \beta = \alpha$ for the significance level to be equal to α. One might then choose β and γ so as to minimise the distance between the two critical values subject to that constraint.

Since the distribution of a statistic that is pivotal under the null hypothesis can be estimated by simulation, inference can be based on the quantiles or the CDF of the estimated distribution. In the limit with an infinite number of bootstrap samples, the simulation error vanishes, and we have exact inference, in the sense that the probability of rejection by a test of significance level α is exactly equal to α when the null is true. If inference is based on a P value, then, for any α between 0 and 1, the probability of obtaining a P value less than α under the null is exactly α.

It is unnecessary to go to the unattainable limit of an infinite number of bootstrap samples in order to have exact inference if one is prepared to restrict attention to certain significance levels. If the finite number of bootstrap samples used is denoted by B, then inference is exact if the level α is such that $\alpha(B+1)$ is an integer. To see this, note that the bootstrap statistics, which we denote as τ_j^*, $j = 1, \ldots, B$, along with the statistic τ obtained from the actual data, constitute a set of $B+1$ statistics which, under the null hypothesis, are independent and identically distributed (IID). Consequently, the number r of bootstrap statistics that are more extreme that τ, according to whatever rule has been chosen for defining rejection regions, is uniformly distributed on the set of integers $0, 1, \ldots, B$, each possible value of r having probability $1/(B+1)$. The bootstrap P value is the probability mass in the bootstrap distribution (that is, the empirical distribution of the B bootstrap statistics) in the region more extreme than τ, and that probability mass is just r/B.

The probability of finding a bootstrap P value less than α is thus $\Pr(r < \alpha B)$. Let $\lceil \alpha B \rceil$ be the smallest integer no smaller than αB. Then the number of possible values of r (strictly) less than αB is $\lceil \alpha B \rceil$. Consequently $\Pr(r < \alpha B) = \lceil \alpha B \rceil / (B+1)$. This probability is equal to α if and only if $\alpha(B+1) = \lceil \alpha B \rceil$. The requirement that $\alpha(B+1)$ should be an integer is clearly a necessary condition for this to be true. Conversely, suppose that $\alpha(B+1) = k$, k an integer. Then $\alpha B = k - \alpha$, and so $\lceil \alpha B \rceil = k$, since $0 < \alpha < 1$. Thus the probability that $r < \alpha B$ is $k/(B+1) = \alpha(B+1)/(B+1) = \alpha$.

The above property is the reason for which, in many studies, the number of bootstrap samples is set to a number like 99, 199, 399, or 999. The decimal system has led to our usual habit of wanting significance levels to be an integer percentage, and these numbers, plus 1, are evenly divisible by 100. In the present computer-dominated era, it would perhaps be more rational to set B equal to a multiple of 16, or 256 (in decimal notation!) minus 1.

Simulation-based testing using a pivotal statistic is in fact much older than bootstrapping. Such tests are called *Monte Carlo tests*, and were introduced back in the 1950s; see Dwass (1957), and also Dufour and Khalaf (2001) for a more recent discussion. At that time, it was not unheard of for a Monte Carlo test to be based on just 19 simulated samples, since that allows for exact inference at the 5% and 10% levels.

Exactly pivotal statistics occur rarely in econometric practice, although they are not completely unknown. It is much commoner to encounter approximately pivotal statistics, the distributions of which do depend on the particular DGP in $\mathbb{M}$ that generates them, but not very sensitively. To make sense of this vague definition, it is common to

construct an *asymptotic theory* for the model $\mathbb{M}$. What this means is that a mathematical construction is given that allows each DGP in $\mathbb{M}$ to generate samples of arbitrarily large size. Often it is quite obvious how to do this, as for instance if the observations in the sample are IID, but in other cases it may be a challenge to find a suitable asymptotic theory for the problem at hand. When the challenge is met, it must be the case that the limiting distribution of the statistic when the sample size tends to infinity is exactly the same for all DGPs in $\mathbb{M}$. A statistic for which an asymptotic theory satisfying this requirement can be found is called *asymptotically pivotal*.

Bootstrap testing is carried out in a way identical to what has been outlined above for Monte Carlo testing. A new problem presents itself, however. Since the distribution of a nonpivotal statistic in finite samples depends on the particular DGP in $\mathbb{M}$, we can no longer choose the DGP used to generate simulated samples arbitrarily. Just how to go about choosing the *bootstrap DGP* is discussed in the next section.

The main focus of this paper is on bootstrap testing, but the bootstrap can be used more generally. The basic principle is that, within the context of some set of DGPs, or model, the DGP that actually generated a given data set can be estimated from those data. Then any quantity, be it a scalar, vector, or matrix, that can be thought of as a function, or functional, of the DGP can be estimated as that function or functional of the estimated DGP. In this way, the bootstrap can be used to estimate bias and variance, quantiles, moments, and many other things. The bootstrap may not provide good estimates of such quantities in all circumstances, but, as we will see, in a testing situation it can provide more reliable inference than other conventional methods.

This paper is complementary to a survey by Davidson and MacKinnon (2006) on bootstrap methods. Here, I focus on bootstrapping independent data, with no discussion of the many difficult problems that can arise when the observations in a sample are mutually dependent. Many of those problems are discussed in Politis (2003). For other useful surveys of the bootstrap, see Horowitz (2001) and Horowitz (2003).

2. The Golden Rules of Bootstrapping

If a test statistic τ is asymptotically pivotal for a given model $\mathbb{M}$, then its distribution should not vary too much as a function of the specific DGP, μ say, within that model. It is usually possible to show that 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.

Asymptotic refinements

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 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 $n^{-(\delta+\varepsilon)}$.

Arguments of the sort sketched in the previous paragraph are used to show that the bootstrap can, in favourable circumstances, benefit from *asymptotic refinements*. The form of the argument was given in a well-known paper of Beran (1988). No doubt wisely, Beran limits himself in this paper to the outline of the argument, with no discussion of formal regularity conditions. It remains true today that no really satisfying general theory of bootstrap testing has been found to embody rigorously the simple idea set forth by Beran. Rather, we have numerous piecemeal results that prove the existence of refinements in specific cases, along with other results that show that the bootstrap does not work in other specific cases. Perhaps the most important instance of negative results of this sort, often called *bootstrap failure*, applies to bootstrapping when the true DGP generates data with a heavy-tailed distribution; see Athreya (1987) for the case of infinite variance. Things are a good deal better for the *parametric bootstrap*, which we study in the next section.

A technique that has been used a good deal in work on asymptotic refinements for the bootstrap is *Edgeworth expansion* of distributions, usually distributions that become standard normal in the limit of infinite sample size. The standard reference to this line of work is Hall (1992), although there is no shortage of more recent work based on Edgeworth expansions. Whereas the technique can lead to useful theoretical insights, it is unfortunately not very useful as a quantitative explanation of the properties of bootstrap tests. In concrete cases, the true finite-sample distribution of a bootstrap P value, as estimated by simulation, can easily be further removed from an Edgeworth approximation to its distribution than from the asymptotic limiting distribution.

Rules for bootstrapping

All these theoretical caveats notwithstanding, experience has shown abundantly that bootstrap tests, in many circumstances of importance for applied econometrics, are much more reliable than tests based on asymptotic theories of one sort or another. In the remainder of this section, we will lay down some rules to follow when reliable bootstrap P values are desired.

The DGP used to generate bootstrap samples from which bootstrap statistics are computed is called the *bootstrap DGP*, and will be denoted as μ^*. 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 μ^* must belong to the model $\mathbb{M}$ 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. This point will become clearer when we discuss confidence sets. In such cases,

a common technique is to change 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.

Violations of Golden Rule 1 are nowadays vanishingly rare in econometric work, although they did occur in some early applications of the bootstrap in the econometric literature. One implication of the rule is that the null model $\mathbb{M}$ should be clearly defined before a bootstrap DGP is chosen. As will be explained in the next section, test statistics based on maximum likelihood estimation should be bootstrapped using a parametric bootstrap in order to satisfy Golden Rule 1. Resampling is appropriate only when the null model admits DGPs based on discrete distributions.

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 above.

Golden Rule 2:

> Unless the test statistic is pivotal for the null model $\mathbb{M}$, the bootstrap DGP should be as good an estimate of the true DGP as possible, under the assumption that the true DGP belongs to $\mathbb{M}$.

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.

Once the sort of bootstrap DGP has been chosen, the procedure for conducting a bootstrap test based on simulated bootstrap samples follows the following pattern.

(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 μ^*.

(iii) Generate B bootstrap samples using μ^*, and for each one compute a realisation of the bootstrap statistic, τ_j^*, $j = 1, \dots B$. 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_j^* that are more extreme than $\hat{\tau}$. For a statistic that rejects for large values, for instance, we have

$$P_{\text{bs}} = \frac{1}{B} \sum_{j=1}^{B} \text{I}(\tau_j^* > \hat{\tau}),$$

where I(·) 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 $P_{\text{bs}} < \alpha$.

3. The Parametric Bootstrap

If the model $\mathbb{M}$ that represents the null hypothesis can be estimated by maximum likelihood (ML), there is a one-one relation between the parameter space of the model and the DGPs that belong to it. For any fixed admissible set of parameters, the likelihood function evaluated at those parameters is a probability density. Thus there is one and only one DGP associated with the set of parameters. By implication, the only DGPs in $\mathbb{M}$ are those completely characterised by a set of parameters.

If the model $\mathbb{M}$ actually is estimated by ML, then the ML parameter estimates provide an asymptotically efficient estimate not only of the true parameters themselves, but also of the true DGP. Both golden rules are therefore satisfied if the bootstrap DGP is chosen as the DGP in $\mathbb{M}$ characterised by the ML parameter estimates. In this case we speak of a *parametric bootstrap*.

In microeconometrics, models like probit and logit are commonly estimated by ML. These are of course just the simplest of microeconometric models, but they are representative of all the others for which it is reasonable to suppose that the data can be described by a purely parametric model. We use the example of a binary choice model to illustrate the parametric bootstrap.

A binary choice model

Suppose that a binary dependent variable y_t, $t = 1, \ldots, n$, takes on only the values 0 and 1, with the probability that $y_t = 1$ being given by $F(\boldsymbol{X}_t\boldsymbol{\beta})$, where $\boldsymbol{X}_t$ is a $1 \times k$ vector of exogenous explanatory variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters, and F is a function that maps real numbers into the $[0, 1]$ interval. For probit, F is the CDF of the standard normal distribution; for logit, it is the CDF of the logistic distribution.

The contribution to the loglikelihood for the whole sample made by observation t is

$$\mathrm{I}(y_t = 1)\, \log F(\boldsymbol{X}_t\boldsymbol{\beta}) + \mathrm{I}(y_t = 0)\, \log\bigl(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\bigr),$$

where I(·) is again an indicator function. Suppose now that the parameter vector $\boldsymbol{\beta}$ can be partitioned into two subvectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and that, under the null hypothesis, $\boldsymbol{\beta}_2 = \mathbf{0}$. The *restricted* ML estimator, that is, the estimator of the subvector $\boldsymbol{\beta}_1$ only, with $\boldsymbol{\beta}_2$ set to zero, is then an asymptotically efficient estimator of the only parameters that exist under the null hypothesis. (It is assumed here that there is an asymptotic construction allowing for arbitarily large numbers of vectors $\boldsymbol{X}_t$ of explanatory variables, all with properties sufficiently similar to allow the application of the usual asymptotic theory of ML.)

Although asymptotic theory is used to convince us of the desirability of the ML estimator, the bootstrap itself is a purely finite-sample procedure. If we denote the restricted ML estimate as $\tilde{\boldsymbol{\beta}} \equiv [\tilde{\boldsymbol{\beta}}_1 \vdots \mathbf{0}]$, the bootstrap DGP can be represented as follows.

$$y_t^* = \begin{cases} 1 & \text{with probability } F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}}), \text{ and} \\ 0 & \text{with probability } 1 - F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}}). \end{cases}, \qquad t = 1, \dots, n. \tag{1}$$

Here the usual notational convention is followed, according to which variables generated by the bootstrap DGP are starred. Note that the explanatory variables $\boldsymbol{X}_t$ are *not* starred. Since they are assumed to be exogenous, it is not the business of the bootstrap DGP to regenerate them; rather they are thought of as fixed characteristics of the bootstrap DGP, and so are used unchanged in each bootstrap sample. Since the bootstrap samples are exactly the same size, n, as the original sample, there is no need to generate explanatory variables for any more observations than those actually observed.

It is easy to implement the formula (1) in order to generate bootstrap samples. A *random number* m_t is drawn, using a random number generator, as a drawing from the uniform U(0, 1) distribution. Then we generate y_t^* as $\mathrm{I}\big(m_t \leq F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})\big)$. Most matrix or econometric software can implement this as a vector relation, so that, after computing the n–vector with typical element $F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$, the vector $\boldsymbol{y}^*$ with typical element y_t^* can be generated by a single command.

Recursive simulation

In dynamic models, the implementation of the bootstrap DGP may require *recursive simulation*. Let us now take as an example the very simple autoregressive time-series model

$$y_t = \alpha + \rho y_{t-1} + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2), \quad t = 2, \dots, n. \tag{2}$$

The notation indicates that the u_t are independent and identically distributed as $\mathrm{N}(0, \sigma^2)$. Thus the dependent variable y_t is now continuous, unlike the binary dependent variable above. The model parameters are α, ρ, and σ^2. However, even if the values of these parameters are specified, (2) is still not a complete characterisation of a DGP. Because (2) is a recurrence relation, it needs a starting value, or initialisation, before it yields a unique solution. Thus, although it is not a parameter in the usual sense, the first observation, y_1, must also be specified in order to complete the model.

ML estimation of the model (2) is the same as estimation by ordinary least squares (OLS) omitting the first observation. If (2) represents the null hypothesis, then we would indeed estimate α, ρ, and σ by OLS. If the null hypothesis specifies the value of any one of those parameters, requiring for instance that $\rho = \rho_0$, then we would use OLS to estimate the model in which this restriction is imposed:

$$y_t - \rho_0 y_{t-1} = \alpha + u_t,$$

with the same specification of the disturbances u_t as in (2).

The bootstrap DGP is then the DGP contained in the null hypothesis that is characterised by the restricted parameter estimates, and by some suitable choice of the starting value, y_1^*. One way to choose y_1^* is just to set it y_1, the value in the original sample. In most cases, this is the best choice. It restricts the model (2) by fixing the initial value. A bootstrap sample can now be generated recursively, starting with y_2^*. For all $t = 2, \ldots, n$, we have

$$y_t^* = \tilde{\alpha} + \tilde{\rho} y_{t-1}^* + \tilde{\sigma} v_t^*, \quad v_t^* \sim \text{NID}(0, 1). \tag{3}$$

Often, one wants to restrict the possible values of ρ to values strictly between -1 and 1. This restriction makes the series y_t asymptotically stationary, by which we mean that, if we generate a very long sample from the recurrence (2), then towards the end of the sample, the distribution of y_t becomes independent of t, as also the joint distribution of any pair of observations, y_t and y_{t+s}, say. Sometimes it make sense to require that the series y_t should be stationary, and not just asymptotically stationary, so that the distribution of every observation y_t, including the first, is always the same. It is then possible to include the information about the first observation into the ML procedure, and so get a more efficient estimate that incorporates the extra information. For the bootstrap DGP, y_1^* should now be a random drawing from the stationary distribution.

The bootstrap discrepancy

Unlike a Monte Carlo test based on an exactly pivotal statistic, a bootstrap test does not in general yield exact inference. This means that there is a difference between the actual probability of rejection and the nominal significance level of the test. We can define the *bootstrap discrepancy* as this difference, as a function of the true DGP and the nominal level. In order to study the bootstrap discrepancy, we suppose, without loss of generality, that the test statistic, denoted τ, is already in approximate P value form. Rejection at level α is thus the event $\tau < \alpha$.

We introduce two functions of the nominal level α of the test and the DGP μ. The first of these is the *rejection probability function*, or RPF. The value of this function is the true rejection probability under μ of a test at level α, and for some fixed finite sample size n. It is defined as

$$R(\alpha, \mu) \equiv \Pr_\mu(\tau < \alpha). \tag{4}$$

Throughout, we assume that, for all $\mu \in \mathbb{M}$, the distribution of τ has support $[0, 1]$ and is absolutely continuous with respect to the uniform distribution on that interval.

For given μ, $R(\alpha, \mu)$ is just the CDF of τ evaluated at α. The inverse of the RPF is the *critical value function*, or CVF, which is defined implicitly by the equation

$$\Pr_\mu\big(\tau < Q(\alpha, \mu)\big) = \alpha. \tag{5}$$

It is clear from (5) that $Q(\alpha, \mu)$ is the α–quantile of the distribution of τ under μ. In addition, the definitions (4) and (5) imply that

$$R\big(Q(\alpha, \mu), \mu\big) = Q\big(R(\alpha, \mu), \mu\big) = \alpha \tag{6}$$

for all α and μ.

In what follows, we will abstract from simulation randomness, and assume that the distribution of τ under the bootstrap DGP is known exactly. The bootstrap critical value for τ at level α is $Q(\alpha, \mu^*)$; recall that μ^* denotes the bootstrap DGP. This is a random variable which would be nonrandom and equal to α if τ were exactly pivotal. If τ is approximately (for example, asymptotically) pivotal, realisations of $Q(\alpha, \mu^*)$ should be close to α. This is true whether or not the true DGP belongs to the null hypothesis, since the bootstrap DGP μ^* does so, according to the first Golden Rule. The bootstrap discrepancy under a DGP $\mu \in \mathbb{M}$ arises from the possibility that, in a finite sample, $Q(\alpha, \mu^*) \neq Q(\alpha, \mu)$.

Rejection by the bootstrap test is the event $\tau < Q(\alpha, \mu^*)$. Applying the increasing transformation $R(\cdot, \mu^*)$ to both sides and using (6), we see that the bootstrap test rejects whenever

$$R(\tau, \mu^*) < R\big(Q(\alpha, \mu^*), \mu^*\big) = \alpha.$$

Thus the bootstrap P value is just $R(\tau, \mu^*)$. This can be interpreted as a bootstrap test statistic. The probability under μ that the bootstrap test rejects at nominal level α is

$$\Pr{}_\mu\big(\tau < Q(\alpha, \mu^*)\big) = \Pr{}_\mu\big(R(\tau, \mu^*) < \alpha\big).$$

We define two random variables that are deterministic functions of the two random elements, τ and μ^*, needed for computing the bootstrap P value $R(\tau, \mu^*)$. The first of these random variables is distributed as $\mathrm{U}(0, 1)$ under μ; it is

$$p \equiv R(\tau, \mu). \tag{7}$$

The uniform distribution of p follows from the fact that $R(\cdot, \mu)$ is the CDF of τ under μ and the assumption that the distribution of τ is absolutely continuous on the unit interval for all $\mu \in \mathbb{M}$. The second random variable is

$$r \equiv R(Q(\alpha, \mu^*), \mu). \tag{8}$$

We may rewrite the event which leads to rejection by the bootstrap test at level α as $R(\tau, \mu) < R(Q(\alpha, \mu^*), \mu)$, by acting on both sides of the inequality $\tau < Q(\alpha, \mu^*)$ by the increasing function $R(\cdot, \mu)$. With the definitions (7) and (8), this event becomes simply $p < r$. Let the CDF of r under μ conditional on the random variable p be denoted as $F(r \,|\, p)$. Then the probability under μ of rejection by the bootstrap test at level α is

$$\begin{aligned} \mathrm{E}(\mathrm{I}(p < r)) &= \mathrm{E}\big(\mathrm{E}(\mathrm{I}(p < r) \,|\, p)\big) = \mathrm{E}\big(\mathrm{E}(\mathrm{I}(r > p) \,|\, p)\big) \\ &= \mathrm{E}\big(1 - F(p \,|\, p)\big) = 1 - \int_0^1 F(p \,|\, p)\, \mathrm{d}p, \end{aligned} \tag{9}$$

since the marginal distribution of p is $\mathrm{U}(0, 1)$.

A useful expression for the bootstrap discrepancy is obtained by defining the random variable $q \equiv r - \alpha$. The CDF of q conditional on p is then $F(\alpha + q \,|\, p) \equiv G(q \,|\, p)$. The RP (9) minus α is

$$1 - \alpha - \int_0^1 G(p - \alpha \,|\, p)\, \mathrm{d}p.$$

Changing the integration variable from p to $x = p - \alpha$ gives for the bootstrap discrepancy

$$\begin{aligned} & 1 - \alpha - \int_{-\alpha}^{1-\alpha} G(x \,|\, \alpha + x)\, \mathrm{d}x \\ &= 1 - \alpha - \Big[x\, G(x \,|\, \alpha + x) \Big]_{-\alpha}^{1-\alpha} + \int_{-\alpha}^{1-\alpha} x\, \mathrm{d}G(x \,|\, \alpha + x) \\ &= \int_{-\alpha}^{1-\alpha} x\, \mathrm{d}G(x \,|\, \alpha + x), \end{aligned} \tag{10}$$

because $G(-\alpha \,|\, 0) = F(0 \,|\, 0) = 0$ and $G(1 - \alpha \,|\, 1) = F(1 \,|\, 1) = 1$.

To a very high degree of approximation, (10) can often be replaced by

$$\int_{-\infty}^{\infty} x\, \mathrm{d}G(x \,|\, \alpha), \tag{11}$$

that is, the expectation of q conditional on p being at the margin of rejection at level α. In cases in which p and q are independent or nearly so, it may even be a good approximation to replace (11) by the unconditional expectation of q.

The random variable r is the probability that a statistic generated by the DGP μ is less than the α–quantile of the bootstrap distribution, conditional on that distribution. The expectation of r minus α can thus be interpreted as the bias in rejection probability when the latter is estimated by the bootstrap. The actual bootstrap discrepancy, which is a nonrandom quantity, is the expectation of $q = r - \alpha$ conditional on being at the margin of rejection. The approximation (11) sets the margin at the α–quantile of τ under μ, while the exact expression (10) takes account of the fact that the margin is in fact determined by the bootstrap DGP.

If the statistic τ is asymptotically pivotal, the random variable q tends to zero under the null as the sample size n tends to infinity. This follows because, for an asymptotically pivotal statistic, the limiting value of $R(\alpha, \mu)$ for given α is the same for all $\mu \in \mathbb{M}$, and similarly for $Q(\alpha, \mu)$. Let the limiting functions of α alone be denoted by $R^\infty(\alpha)$ and $Q^\infty(\alpha)$. Under the assumption of an absolutely continuous distribution, the functions R^∞ and Q^∞ are inverse functions (recall (6)), and so, as $n \to \infty$, $r = R(Q(\alpha, \mu^*), \mu)$ tends to $R^\infty(Q^\infty(\alpha)) = \alpha$, and so $q = r - \alpha$ tends to zero in distribution, and so also in probability.

Suppose now that the random variables q and p are independent. Then the conditional CDF $G(\cdot \,|\, \cdot)$ is just the unconditional CDF of q, and the bootstrap discrepancy

(10) is the unconditional expectation of q. The unconditional expectation of a random variable that tends to 0 can tend to 0 more quickly than the variable itself, and more quickly than the expectation conditional on another variable correlated with it. Independence of q and p does not often arise in practice, but approximate (asymptotic) independence occurs regularly when the parametric bootstrap is used along with ML estimation of the null hypothesis. It is a standard result of the asymptotic theory of maximum likelihood that the ML parameter estimates of a model are asymptotically independent of the classical test statistics used to test the null hypothesis that the model is well specified against some parametric alternative. In such cases, the bootstrap discrepancy tends to zero faster than if inefficient parameter estimates are used to define the bootstrap DGP. This argument, which lends support to Golden Rule 2, is developed in Davidson and MacKinnon (1999).

4. Resampling

The analysis of the previous section relies on the absolute continuity of the distribution of the test statistic for all $\mu \in \mathbb{M}$. Even when a parametric bootstrap is used, absolute continuity does not always pertain. For instance, the dependent variable of a binary choice model is a discrete random variable, and so too are any test statistics that are functions of it, unless continuity arises for some other reason, which is not the case with the test statistics in common use for binary choice models. However, since the discrete set of values a test statistic can take on rapidly becomes very rich as sample size increases, it is reasonable to suppose that the theory of the previous section remains a good approximation for realistic sample sizes.

Basic resampling

Another important circumstance in which absolute continuity fails is when the bootstrap DGP makes use of *resampling*. Resampling was a key aspect of the original conception of the bootstrap, as set out in Efron's (1979) pioneering paper. Resampling is valuable when it is undesirable to constrain a model so tightly that all of its possibilities are encompassed by the variation of a finite set of parameters. A classic instance is a regression model where one does not wish to impose the normality of the disturbances. To take a concrete example, let us look again at the autoregressive model (2), relaxing the condition on the disturbances so as to require only IID disturbances with expectation 0 and variance σ^2.

The bootstrap DGP (3) satisfies Golden Rule 1, because the normal distribution is plainly allowed when all we specify are the first two moments. But Golden Rule 2 incites us to seek as good an estimate as possible of the unknown distribution of the disturbances. If the disturbances were observed, then the best nonparametric estimate of their distribution would be their EDF. The unobserved disturbances can be estimated, or proxied, by the residuals from estimating the null model. If we denote the empirical distribution of these residuals by $\hat{F}$, then (3) could be replaced by

$$y_t^* = \tilde{\alpha} + \tilde{\rho} y_{t-1}^* + u_t^*, \quad u_t^* \sim \text{IID}(\hat{F}), \quad t = 2, \dots, n.$$

where the notation indicates that the bootstrap disturbances, the u_t^*, are IID drawings from the empirical distribution characterised by the EDF $\hat{F}$.

The term *resampling* comes from the fact that the easiest way to generate the u_t^* is to sample from the residuals at random with replacement. The residuals are thought of as sampling the true DGP, and so this operation is called "resampling". For each $t = 2, \ldots, n$, one can draw a random number m_t from the U(0, 1) distribution, and then obtain u_t^* by the operations:

$$s = \lfloor 2 + (n-1)m_t \rfloor, \quad u_t^* = \tilde{u}_s,$$

where the notation $\lfloor x \rfloor$ means the greatest integer not greater than x. For n_t close to 0, $s = 2$; for n_t close to 1, $s = n$, and we can see that s is uniformly distributed over the integers $2, \ldots, n$. Setting u_t^* equal to the (restricted) residual $\tilde{u}_s$ therefore implements the required resampling operation.

More sophisticated resampling

But is the empirical distribution of the residuals really the best possible estimate of the distribution of the disturbances? Not always. Consider an even simpler model than (2), one with no constant term:

$$y_t = \rho y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \tag{12}$$

When this is estimated by OLS, or, if the null hypothesis fixes the value of ρ, in which case the "residuals" are just the observed values $y_t - \rho_0 y_{t-1}$, the residuals do not in general sum to zero, precisely because there is no constant term. But the model (12) requires that the expectation of the disturbance distribution should be zero, whereas the expectation of the empirical distribution of the residuals is their mean. Thus using this empirical distribution violates Golden Rule 1.

This is easily fixed by replacing the residuals by the deviations from their mean, and then resampling these centred residuals. But now what about Golden Rule 2? The variance of the centred residuals is the sum of their squares divided by n:

$$V = \frac{1}{n} \sum_{t=1}^{n} (\tilde{u}_t^2 - \bar{u})^2,$$

where $\bar{u}$ is the mean of the uncentred residuals. But the unbiased estimator of the variance of the disturbances is

$$s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (\tilde{u}_t^2 - \bar{u})^2.$$

More generally, in any regression model that uses up k degrees of freedom in estimating regression parameters, the unbiased variance estimate is the sum of squared residuals

divided by $n - k$. What this suggests is that what we want to resample is a set of *rescaled* residuals, which here would be the $\sqrt{n/(n-k)}\tilde{u}_t$. The variance of the empirical distribution of these rescaled residuals is then equal to the unbiased variance estimate.

Of course, some problems are scale-invariant. Indeed, test statistics that are ratios are scale invariant for both models (2) and (12) under the stationarity assumption. For models like these, therefore, there is no point in rescaling, since bootstrap statistics computed with the same set of random numbers are unchanged by scaling. This property is akin to pivotalness, in that varying some, but not all, of the parameters of the null model leaves the distribution of the test statistic invariant. In such cases, it is unnecessary to go to the trouble of estimating parameters that have no effect on the distribution of the statistic τ.

A poverty index

Centring and scaling are simple operations that alter the first two moments of a distribution. In some circumstances, we may wish to affect the values of more complicated functionals of a distribution. Suppose for instance that we wish to perform inference about a poverty index. An IID sample of individual incomes is available, drawn at random from the population under study, and the null hypothesis is that a particular poverty index has a particular given value. For concreteness, let us consider one of the FGT indices, defined as follows; see Foster, Greer, and Thorbecke (1984).

$$\Delta^\alpha(z) = \int_0^z (z-y)^{\alpha-1}\,\mathrm{d}F(y).$$

Here z is interpreted as a poverty line, and F is the CDF of income. As the parameter α increases, the index puts progressively greater weight on large values of the *poverty gap*, that is, the difference $z - y$ between the the poverty line and the income y of a poor individual. We assume that the poverty line z and the parameter α are fixed at some prespecified values. The obvious estimator of of $\Delta^\alpha(z)$ is just

$$\hat{\Delta}^\alpha(z) = \int_0^z (z-y)^{\alpha-1}\,\mathrm{d}\hat{F}(y),$$

where $\hat{F}$ is the EDF of income in the sample. For sample size n, we have explicitly that

$$\hat{\Delta}^\alpha(z) = \frac{1}{n}\sum_{i=1}^{n}(z-y_i)_+^{\alpha-1}, \tag{13}$$

where y_i is income for observation i, and $(x)_+$ denotes $\max(0, x)$.

Since according to (13) $\hat{\Delta}^\alpha(z)$ is just the mean of a set of IID variables, its variance can be estimated by

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}(z-y_i)_+^{2\alpha-2} - \left(\frac{1}{n}\sum_{i=1}^{n}(z-y_i)_+^{\alpha-1}\right)^2. \tag{14}$$

A suitable test statistic for the hypothesis that $\Delta^\alpha(z) = \Delta_0$ is then

$$t = \frac{\hat{\Delta}^\alpha(z) - \Delta_0}{\hat{V}^{1/2}}.$$

Under the null t is distributed approximately as $\mathrm{N}(0,1)$, and an approximate P value for a two-tailed test is $\tau = 2\Phi(-|t|)$, where $\Phi(\cdot)$ is the standard normal CDF.

With probability 1, the estimate $\hat{\Delta}^\alpha(z)$ is not equal to Δ_0. If the statistic t is bootstrapped using ordinary resampling of the data in the original sample, this fact means that we violate Golden Rule 1. The simplest way around this difficulty, as mentioned after the statement of Golden Rule 1, is to change the null hypothesis tested by the bootstrap statistics, testing rather what is true under the resampling DGP, namely $\Delta^\alpha(z) = \hat{\Delta}^\alpha(z)$. Thus each bootstrap statistic takes the form

$$t^* = \frac{(\Delta^\alpha(z))^* - \hat{\Delta}^\alpha(z)}{(V^*)^{1/2}}.$$

Here $(\Delta^\alpha(z))^*$ is the estimate (13) computed using the bootstrap sample, and V^* is the variance estimator (14) computed using the bootstrap sample. Golden Rule 1 is saved by the trick of changing the null hypothesis for the bootstrap samples, but Golden Rule 2 would be better satisfied if we could somehow impose the real null hypothesis on the bootstrap DGP.

Weighted resampling

A way to impose the null hypothesis with a resampling bootstrap is to resample with unequal weights. Ordinary resampling assigns a weight of n^{-1} to each observation, but if different weights are assigned to different observations, it is possible to impose various sorts of restrictions. This approach is suggested by Brown and Newey (2002).

A nonparametric technique that shares many properties with parametric maximum likelihood is *empirical likelihood*; see Owen (2001). In the case of an IID sample, the empirical likelihood is a function of a set of nonnegative probabilities p_i, $i = 1, \ldots, n$, such that $\sum_{i=1}^n p_i = 1$. The empirical loglikelihood, easier to manipulate than the empirical likelihood itself, is given as

$$\ell(\boldsymbol{p}) = \sum_{i=1}^n \log p_i. \tag{15}$$

Here $\boldsymbol{p}$ denotes the $n-$vector of the probabilities p_i. The idea now is to maximise (15) subject to the constraint that the FGT index for the reweighted sample is equal to Δ_0. Specifically, $\ell(\boldsymbol{p})$ is maximised subject to the constraint

$$\sum_{i=1}^n p_i (z - y_i)_+^{\alpha-1} = \Delta_0. \tag{16}$$

With very small sample sizes, it is possible that this constrained maximisation problem has no solution with nonnegative probabilities. In such a case, the *empirical likelihood ratio* statistic would be set equal to ∞, and the null hypothesis rejected out of hand, with no need for bootstrapping.

In the more common case in which the problem can be solved, the bootstrap DGP resamples the original sample with observation i resampled with probability p_i rather than n^{-1}. The use of empirical likelihood for the determination of the p_i means that these probabilities have various optimality properties relative to any other set satisfying (16). Golden Rule 2 is satisfied.

The best algorithm for weighted resampling appears to be little known in the econometrics community. It is described in Knuth (1998). Briefly, for a set of probabilities p_i, $i = 1, \ldots, n$, two tables of n elements each are set up, containing the values q_i, with $0 < q_i \leq 1$, and y_i, where y_i is an integer in the set $1, \ldots, n$. In order to obtain the index j of the observation to be resampled, a random number m_i from $\mathrm{U}(0, 1)$ is used as follows.

$$k_i = \lceil nm_i \rceil, \quad r_i = k_i - nm_i, \quad j = \begin{cases} k_i & \text{if } r_i \leq q_i, \\ y_i & \text{otherwise.} \end{cases}$$

For details, readers are referred to Knuth's treatise.

5. Other Bootstrap Methods

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. But if the disturbances of a regression are heteroskedastic, with an unknown pattern of heteroskedasticity, there is nothing that is even approximately IID. 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); see Eicker (1963) and White (1980). 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 belongs to a null hypothesis that, since it allows heteroskedasticity, must also allow the special case of homoskedasticity. But Golden Rule 2 poses a more severe challenge.

The 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 bootstraps 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 $1, \ldots, n$, 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.

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 $\boldsymbol{X}_t$.

Suppose that the regression model itself is written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad t = 1, \ldots, n, \tag{17}$$

with $\boldsymbol{X}_t$ a $1 \times k$ vector and $\boldsymbol{\beta}$ a $k \times 1$ vector of parameters. The disturbances u_t are allowed to be heteroskedastic, but must have an expectation of 0 conditional on the explanatory variables. Thus $\mathrm{E}(y_t|\boldsymbol{X}_t) = \boldsymbol{X}_t\boldsymbol{\beta}_0$ if $\boldsymbol{\beta}_0$ is the parameter vector for the true DGP. Let us consider a null hypothesis according to which a subvector of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_2$ say, is zero. This null hypothesis is not satisfied by the pairs bootstrap DGP. In order to respect Golden Rule 1, therefore, we must modify either the null hypothesis to be tested in the bootstrap samples, or the bootstrap DGP itself.

In the empirical joint distribution of the pairs $(y_t, \boldsymbol{X}_t)$, the expectation of the first element y conditional on the second element $\boldsymbol{X}$ is defined only if $\boldsymbol{X} = \boldsymbol{X}_t$ for some $t = 1, \ldots, n$. Then $\mathrm{E}(y|\boldsymbol{X} = \boldsymbol{X}_t) = y_t$. This result does not help determine what the true value of $\boldsymbol{\beta}$, or of $\boldsymbol{\beta}_2$, might be for the bootstrap DGP. Given this, what is usually done is to use the OLS estimate $\hat{\boldsymbol{\beta}}_2$ as true for the bootstrap DGP, and so to test the hypothesis that $\boldsymbol{\beta}_2 = \hat{\boldsymbol{\beta}}_2$ when computing the bootstrap statistics.

In Flachaire (1999), the bootstrap DGP is changed. 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 $1, \ldots, n$, y_t^* is generated by

$$y_t^* = \boldsymbol{X}_{s1}\tilde{\boldsymbol{\beta}}_1 + \hat{u}_s, \tag{18}$$

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 bootstrap DGP (18) now satisfies the null hypothesis as originally formulated.

The 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 (2001).

For the regression model (17), the wild bootstrap DGP takes the form

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + s_t^*\tilde{u}_t \tag{19}$$

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.

The 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 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.

For a long time, the most commonly used distribution for the s_t^* was the following two-point distribution,

$$s_t^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}), \end{cases}$$

which was suggested by Mammen (1993). A simpler two-point distribution is the *Rademacher distribution*

$$s_t^* = \begin{cases} -1 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}. \end{cases} \tag{20}$$

Davidson and Flachaire (2001) propose this simpler distribution, which leaves the absolute value of each residual unchanged in the bootstrap DGP, while assigning it an arbitrary sign. They show by means of simulation experiments that the choice (20) often leads to more reliable bootstrap inference than other choices.

Vector autoregressions

A potential problem with Freedman's pairs bootstrap is that it treats all variables, endogenous and exogenous, in the same way. Some models, however, have more than one endogenous variable, and so, except in a few cases in which we can legitimately condition on some of them, the bootstrap DGP has to be able to generate all of the endogenous variables simultaneously. This is not at all difficult for models such as vector autoregressive (VAR) models. A typical VAR model can be written as

$$\boldsymbol{Y}_t = \sum_{i=1}^{p} \boldsymbol{Y}_{t-i}\boldsymbol{\Pi}_i + \boldsymbol{X}_t\boldsymbol{B} + \boldsymbol{U}_t, \quad t = p+1, \dots, n. \tag{21}$$

Here $\boldsymbol{Y}_t$ and $\boldsymbol{U}_t$ are $1 \times m$ vectors, the $\boldsymbol{\Pi}_i$ are all $m \times m$ matrices, $\boldsymbol{X}_t$ is a $1 \times k$ vector, and $\boldsymbol{B}$ is a $k \times m$ matrix. The m elements of $\boldsymbol{Y}_t$ are the endogenous variables for observation t. The elements of $\boldsymbol{X}_t$ are exogenous explanatory variables – although some VAR models dispense with exogenous variables, so that $k = 0$ in such cases. The elements of the matrices $\boldsymbol{\Pi}_i$, $i = 1, \ldots, p$ and those of $\boldsymbol{B}$ are the parameters of the model. The vectors $\boldsymbol{U}_t$ have expectation zero, and are usually assumed to be mutually independent, although correlated among themselves; the independent elements of the *contemporaneous covariance matrix* $\boldsymbol{\Sigma}$, of dimension $m \times m$, are also parameters of the model.

Among the hypotheses that can be tested in the context of a model like (21) are tests for *Granger causality*; see Granger (1969) or Davidson and MacKinnon (2004) for a textbook treatment. The null hypothesis of these tests is Granger *non-causality*, and it imposes zero restrictions on subsets of the elements of the $\boldsymbol{\Pi}_i$. Unrestricted, the model (21) can be efficiently estimated by least squares applied to each equation separately, with the covariance matrix $\boldsymbol{\Sigma}$ estimated by the empirical covariance matrix of the residuals. Subject to restrictions, the model is usually estimated by maximum likelihood under the assumption that the disturbances are jointly normally distributed.

Bootstrap DGPs can be set up for models that impose varying levels of restrictions. In all cases, the $\boldsymbol{\Pi}_i$ matrices, the $\boldsymbol{\Sigma}$ matrix, and the $\boldsymbol{B}$ matrix, if present, should be set equal to their restricted estimates. In all cases, as well, bootstrap samples should be conditioned on the first p observations from the original sample, unless stationarity is assumed, in which case the first p observations of each bootstrap sample should be drawn from the stationary distribution of p contiguous $m-$vectors $\boldsymbol{Y}_t, \ldots, \boldsymbol{Y}_{t+p-1}$. If normal disturbances are assumed, the bootstrap disturbances can be generated as IID drawings from the multivariate $\mathrm{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}})$ distribution – one obtains by Cholesky decomposition an $m \times m$ matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^\top = \tilde{\boldsymbol{\Sigma}}$, and generates $\boldsymbol{U}_t^*$ as $\boldsymbol{A}\boldsymbol{V}_t^*$, where the m elements of $\boldsymbol{V}_t^*$ are IID standard normal. If it is undesirable to assume normality, then the *vectors* of restricted residuals $\tilde{\boldsymbol{U}}_t$ can be resampled. If it is undesirable even to assume that the $\boldsymbol{U}_t$ are IID, a wild bootstrap can be used in which each of the vectors $\tilde{\boldsymbol{U}}_t$ is multiplied by a scalar s_t^*, with the s_t^* IID drawings from a distribution with expectation 0 and variance 1.

Simultaneous equations

Things are a little more complicated with a *simultaneous-equations model*, in which the endogenous variables for a given observation are determined as the solution of a set of simultaneous equations that also involve exogenous explanatory variables. Lags of the endogenous variables can also appear as explanatory variables; they are said to be *predetermined*. If they are present, the bootstrap DGP must rely on recursive simulation.

A simultaneous-equations model can be written as

$$\boldsymbol{Y}_t\boldsymbol{\Gamma} = \boldsymbol{W}_t\boldsymbol{B} + \boldsymbol{U}_t, \tag{22}$$

with $\boldsymbol{Y}_t$ and $\boldsymbol{U}_t$ $1 \times m$ vectors, $\boldsymbol{W}_t$ a $1 \times k$ vector or exogenous or predetermined explanatory variables, $\boldsymbol{\Gamma}$ an $m \times m$ matrix, and $\boldsymbol{B}$ a $k \times m$ matrix. The elements of $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$, along with the independent elements of the contemporaneous covariance matrix $\boldsymbol{\Sigma}$, are the parameters of the model. In order for the endogenous variables $\boldsymbol{Y}_t$ to be defined by (22), $\boldsymbol{\Gamma}$ must be nonsingular.

The set of equations (22) is called the *structural form* of the model. The *reduced form* is obtained by solving the equations of the structural form to get

$$\boldsymbol{Y}_t = \boldsymbol{W}_t \boldsymbol{B} \boldsymbol{\Gamma}^{-1} + \boldsymbol{V}_t, \tag{23}$$

where the contemporaneous covariance matrix of the $\boldsymbol{V}_t$ is $(\boldsymbol{\Gamma}^\top)^{-1} \boldsymbol{\Sigma} \boldsymbol{\Gamma}^{-1}$. The reduced form can be estimated unrestricted, using least squares on each equation of the set of equations

$$\boldsymbol{Y}_t = \boldsymbol{W}_t \boldsymbol{\Pi} + \boldsymbol{V}_t$$

separately, with $\boldsymbol{\Pi}$ a $k \times m$ matrix of parameters. Often, however, the structural form is *overidentified*, meaning that restrictions are imposed on the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$. This is always the case if the null hypothesis imposes such restrictions. Many techniques exist for the restricted estimation of either one of the equivalent models (22) or (23). When conventional asymptotic theory is used, asymptotic efficiency is achieved by two techniques, *three-stage least squares* (3SLS), and full-information maximum likelihood (FIML). These standard techniques are presented in most econometrics textbooks.

Nonlinear models

Bootstrap DGPs should in all cases use efficient restricted estimates of the parameters, obtained by 3SLS or FIML, with a slight preference for FIML, which has higher-order optimality properties not shared by 3SLS. Bootstrap disturbances can be generated from the multivariate normal distribution, or by resampling vectors of restricted residuals, or by a wild bootstrap procedure. See Davidson and MacKinnon (2006b) for a detailed discussion.

Bootstrapping is often seen as a very computationally intensive procedure, although with the hardware and software available at the time of writing, this is seldom a serious problem in applied work. Models that require nonlinear estimation can be an exception to this statement, because the algorithms used in nonlinear estimation may fail to converge after a small number of iterations. If this happens while estimating a model with real data, the problem is not related to bootstrapping, but arises rather from the relation between the model and the data. The problem for bootstrapping occurs when an estimation procedure that works with the original data does not work with one or more of the bootstrap samples.

In principle nonlinear estimation should be easier in the bootstrap context than otherwise. One knows the true bootstrap DGP, and can use the true parameters for that DGP as the starting point for the iterative procedure used to implement the nonlinear estimation. In those cases in which it is necessary to estimate two models, one

restricted, the other unrestricted, one can use the estimates from the restricted model, say, as the starting point for the unrestricted estimation, thus making use of properties specific to a particular bootstrap sample.

When any nonlinear procedure is repeated thousands of times, it seems that anything that can go wrong will go wrong at least once. Most of the time, the arguments of the previous paragraph apply, but not always. Any iterative procedure can go into an infinite loop if it does not converge, with all sorts of undesirable consequences. It is therefore good practice to set a quite modest upper limit to the number of iterations permitted for each bootstrap sample.

In many cases, an upper limit of just 3 or 4 iterations can be justified theoretically. Asymptotic theory can usually provide a *rate of convergence*, with respect to the sample size, of the bootstrap discrepancy to zero. It can also provide the rate of convergence of Newton's method, or a quasi-Newton method, used by the estimation algorithm. If the bootstrap discrepancy goes to zero as $n^{-3/2}$ say, then there is little point in seeking numerical accuracy with a better rate of convergence. With most quasi-Newton methods, the Gauss-Newton algorithm for instance, each iteration reduces the distance between the current parameters of the algorithm and those to which the algorithm will converge (assuming that it does converge) by a factor of $n^{-1/2}$. Normally, we can initialise the algorithm with parameters that differ from those at convergence by an amount of order $n^{-1/2}$. After three iterations, the difference is of order only n^{-2}, a lower order than that of the bootstrap discrepancy. The same order of accuracy is thus achieved on average as would be attainable if the iterations continued until convergence by some stricter criterion. Since bootstrap inference is based on an average over the bootstrap repetitions, this is enough for most purposes.

More details of this idea for limiting the number of iterations can be found in Davidson and MacKinnon (1999), where it is also pointed out that numerical methods for computing likelihood ratio statistics converge even faster than those for Wald or Lagrange mulitplier statistics. A detailed treatment of the asymptotic theory behind the idea can be found in Andrews (2002).

6. Confidence Sets

It is probably fair to say that, in the statistical literature on the bootstrap, the greatest effort has gone into developing bootstrap methods for constructing confidence intervals, or confidence sets more generally. Inference based on confidence sets is in principle equivalent to inference based on hypothesis tests, but in practice there can be obstacles to this theoretical equivalence. As a general rule, conventional methods of bootstrap hypothesis testing perform better than conventional methods of constructing bootstrap confidence sets.

Percentile methods

Consider the case of inference about a scalar parameter θ. Suppose that, for each possible value of the parameter, there is a test statistic $\tau(\theta)$ of which the distribution is known, approximately at least, when θ is the true parameter. A confidence interval at *confidence level* $1 - \alpha$, $0 < \alpha < 1$, is the set of θ for which the hypothesis that θ is the true parameter is not rejected at significance level α by the test based on the statistic $\tau(\theta)$. Let C denote the confidence interval thus generated. Then, if the true parameter is θ, we have, perhaps only approximately, that

$$\Pr(\theta \in C) = 1 - \Pr\big(\tau(\theta) \in \mathrm{Rej}(\alpha)\big) = 1 - \alpha, \tag{24}$$

where $\mathrm{Rej}(\alpha)$ is the rejection region for the test at significance level α. We use this notation so as to be able to cover the cases of one-sided or two-sided confidence intervals, arising from one-tailed or two-tailed tests respectively.

Conversely, if for each confidence level $1 - \alpha$, we have a confidence interval C_α, a test at significance level α rejects the hypothesis that θ is the true parameter if and only if $\theta \notin C_\alpha$. A P value for this hypothesis can be defined by the relation

$$P(\theta) = \max\{\alpha \,|\, \theta \in C_\alpha\}.$$

A very straightforward way of getting a confidence interval for θ is called the *percentile method*. A model $\mathbb{M}$ is assumed, and, for each DGP $\mu \in \mathbb{M}$, there is an associated parameter $\theta(\mu)$, the "true" parameter for the DGP μ. Since there are usually other parameters besides θ, they must be estimated along with θ in order to set up a bootstrap DGP. It is not necessary for this discussion to distinguish the various possible bootstrap DGPs; in any case they all make use of the estimated parameters. The first step is then to obtain estimates of all the parameters, which should be as efficient as possible for the model $\mathbb{M}$.

Consider first the case of an *equal-tailed* confidence interval. Let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ be the $\alpha/2$ and $(1 - \alpha/2)$–quantiles of the distribution of $\hat{\theta} - \theta$, where θ is the true parameter. Then we see that

$$\Pr(q_{\alpha/2} \le \hat{\theta} - \theta \le q_{1-\alpha/2}) = \alpha.$$

The inequalities above are equivalent to

$$\hat{\theta} - q_{1-\alpha/2} \le \theta \le \hat{\theta} - q_{\alpha/2},$$

and from this it is clear that the confidence interval with lower bound $\hat{\theta} - q_{1-\alpha/2}$ and upper bound $\hat{\theta} - q_{\alpha/2}$ contains the true θ with probability α.

The next step is to generate a set of bootstrap samples, and to compute the parameter estimate, θ^* say, for each of them. Since the true value of θ for the bootstrap DGP is $\hat{\theta}$, we can use the distribution of $\theta^* - \hat{\theta}$ as an estimate of the distribution of $\hat{\theta} - \theta$.

In particular, the $\alpha/2$ and $(1-\alpha/2)$–quantiles of the distribution of $\theta^* - \hat{\theta}$, $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$ say, give the percentile confidence interval

$$C^*_\alpha = [\hat{\theta} - q^*_{1-\alpha/2},\ \hat{\theta} - q^*_{\alpha/2}].$$

For a one-sided confidence interval that is open to the right, we use $[\hat{\theta} - q^*_{1-\alpha},\ \infty[$, and for one that is open to the left $]-\infty,\ \hat{\theta} - q^*_\alpha]$. Note the somewhat counter-intuitive fact that the *upper* quantile of the distribution determines the *lower* limit of the confidence interval, and *vice versa*.

The percentile interval is very far from being the best bootstrap confidence interval. The first reason is that, in almost all interesting cases, the random variable $\hat{\theta} - \theta$ is not even approximately pivotal. Indeed, conventional asymptotics give a limiting distribution of $\mathrm{N}(0, \sigma^2_\theta)$, for some asymptotic variance σ^2_θ. Unless σ^2_θ is constant for all DGPs in $\mathbb{M}$, it follows that $\hat{\theta} - \theta$ is not asymptotically pivotal.

For this reason, a more popular bootstrap confidence interval is the *percentile–t* interval. Now we suppose that we can estimate the variance of $\hat{\theta}$, and so base the confidence interval on the *studentised* quantity $(\hat{\theta} - \theta)/\hat{\sigma}_\theta$, which in many circumstances is asymptotically standard normal, and hence asymptotically pivotal. Let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ be the relevant quantiles of the distribution of $(\hat{\theta} - \theta)/\hat{\sigma}_\theta$, when the true parameter is θ. Then

$$\Pr\left(q_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\hat{\sigma}_\theta} \leq q_{1-\alpha/2}\right) = \alpha.$$

If the quantiles are estimated by the quantiles of the distribution of $(\theta^* - \hat{\theta})/\sigma^*_\theta$, where σ^*_θ is the square root of the variance estimate computed using the bootstrap sample, we obtain the percentile–t confidence interval

$$C^*_\alpha = [\hat{\theta} - \hat{\sigma}_\theta q^*_{1-\alpha/2},\ \hat{\theta} - \hat{\sigma}_\theta q^*_{\alpha/2}]. \tag{25}$$

In many cases, the performance of the percentile–t interval is much better than that of the percentile interval. For a more complete discussion of bootstrap confidence intervals of this sort, see Hall (1992).

Equal-tailed confidence intervals are not the only ones than can be constructed using the percentile or percentile–t methods. Recall that critical values for tests at level α can be based on the β and γ–quantiles for the lower and upper critical values provided that $1 - \gamma + \beta = \alpha$. A bootstrap distribution is rarely symmetric about its central point (unless it is deliberately so constructed). The β and γ that minimise the distance between the β–quantile and the γ–quantile under the constraint $1 - \gamma + \beta = \alpha$ are then not $\alpha/2$ and $1 - \alpha/2$ in general. Using the β and γ obtained in this way leads to the *shortest* confidence interval at confidence level $1 - \alpha$.

The construction of the percentile–t interval follows the rule by which the confidence set C of (24) is constructed. The statistic $\tau(\theta)$ becomes $(\hat{\theta} - \theta)/\hat{\sigma}_\theta$, and the rejection region $\mathrm{Rej}(\alpha)$ is determined from the bootstrap distribution of $(\theta^* - \hat{\theta})/\sigma^*_\theta$. Golden

Rule 1 is respected, because the bootstrap statistic tests a hypothesis that is true of the bootstrap DGP, namely that $\theta = \hat{\theta}$.

The confidence interval takes the simple form (25) only because the test statistic is a simple function of θ. This simplicity may come at a cost, however. The statistic $(\hat{\theta}-\theta)/\hat{\sigma}_\theta$ is a Wald statistic, and it is known that Wald statistics may have undesirable properties. The worst of these is that such statistics are not invariant to nonlinear reparametrisations. For instance, if we define a new parameter ϕ by the relation $\phi = h(\theta)$, where h is a monotonically increasing nonlinear function, then a confidence interval based on the Wald statistic $(\hat{\phi}-\phi)/\hat{\sigma}_\phi$ is different from that based on $(\hat{\theta}-\theta)/\hat{\sigma}_\theta$. Similarly, for a given null hypothesis $\theta = \theta_0$, or, equivalently, $\phi = \phi_0 = h(\theta_0)$, a test, bootstrap or otherwise, based on one statistic may reject while the other does not. See Gregory and Veall (1985) and Lafontaine and White (1986) for analysis of this phenomenon.

Confidence intervals based on better statistics

A confidence set need not be based on a Wald test. Suppose that $\tau(\theta)$ is a likelihood ratio statistic, or a Lagrange multiplier statistic, that tests the hypothesis that θ is the true parameter value. These statistics can be made invariant under reparametrisation. They are often approximately distributed as chi-squared, and so reject for large values of the statistic. A confidence set C_α with nominal confidence level $1-\alpha$ is characterised as usual:

$$C_\alpha = \{\theta \,|\, \tau(\theta) > q_{1-\alpha}\} \tag{26}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$–quantile of whatever nominal distribution is used to determine rejection – the chi-squared distribution with appropriate degrees of freedom for a test based on asymptotics, or a distribution obtained by bootstrapping. Boundary points of the confidence set (26) then satisfy the equation $\tau(\theta) = q_{1-\alpha}$.

In general, it may be awkward, or even impossible, to obtain an analytic form for $\tau(\theta)$. Should this be so, the equation $\tau(\theta) = q_{1-\alpha}$ may have to be solved by numerical methods. In regular cases, this equation has exactly two solutions, the lower and upper endpoints of the confidence interval. In less well-behaved cases, the equation may define an unbounded interval, or even a union of disjoint intervals, any of which may be unbounded. Confidence sets that are not bounded intervals are documented by Dufour (1997).

Simultaneous confidence regions for more than one parameter can also be defined by (26), by reinterpreting θ as a vector. Wald statistics give rise to ellipsoidal regions that are easy to characterise. But other sorts of statistic can lead to confidence regions with more complicated shapes. There is no difference in principle between confidence sets based on asymptotics and those based on the bootstrap. In all cases, a nominal distribution provides a quantile or quantiles, and these characterise the confidence region.

A Wald statistic is, by definition, based on the estimation of the *alternative* hypothesis. This fact does not sit well with Golden Rule 2. But even if we use a Lagrange multiplier

statistic, based on estimation of the null hypothesis, it can be argued that Golden Rule 2 is still not satisfied. One problem is that, in order to construct a confidence set, it is in principle necessary to consider an infinity of null hypotheses; in (26), θ can range over an open interval that is often the entire real line. In practice, provided one is sure that a confidence set is a single, connected, interval, then it is enough to locate the two values of θ that satisfy $\tau(\theta) = q_{1-\alpha}$.

Respecting Golden Rule 2

Where Golden Rule 2 is not respected is in the assumption that the distribution of $\tau(\theta)$, under a DGP for which θ is the true parameter, is the same for all θ. If this happens to be the case, the statistic is called pivotal, and there is no further problem. But if the statistic is only approximately pivotal, its distribution when the true θ is an endpoint of a confidence interval is not the same as when the true parameter is the point estimate $\hat{\theta}$. The true parameter for the bootstrap DGP, however, is $\hat{\theta}$.

For Golden Rule 2 to be fully respected, the equation that should be solved for endpoints of the confidence interval is

$$\tau(\theta) = q_{1-\alpha}(\theta), \tag{27}$$

where $q_{1-\alpha}(\theta)$ is the $(1-\alpha)$–quantile of the distribution of $\tau(\theta)$ when θ is the true parameter. If θ is the only parameter, then it is possible, although usually not easy, to solve (27) by numerical methods based on simulation. In general, though, things are even more complicated. If, besides θ, there are other parameters, that we can call nuisance parameters in this context, then according to Golden Rule 2, we should use the best estimate possible of these parameters under the null for the bootstrap DGP. So, for each value of θ considered in a search for the solution to (27), we should re-estimate these nuisance parameters under the constraint that θ is the true parameter, and then base the bootstrap DGP on θ and these restricted estimates. This principle underlies the so-called *grid bootstrap* proposed by Hansen (1999). It is, not surprisingly, very computationally intensive, but Hansen shows that it yields satisfactory results for an autoregressive model where other bootstrap confidence intervals give unreliable inference.

7. Concluding Remarks

The bootstrap is a statistical technique capable of giving reliable inference for a wide variety of econometric models. In this article, I focus on bootstrap-based inference. 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.

In this article, only a sketch is given of the numerous uses of the bootstrap in econometrics. Nothing is said about the thorny problems of bootstrapping dependent data,

or about the difficulties posed by heavy-tailed distributions. Both of these are currently active fields of research, and it is to be hoped that we will understand more about them in the near future.

It seems clear that 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 characterisation as possible of the distribution, under the null hypothesis under test, of the test statistics on which inference is based.

Some time has elapsed since Beran (1988) pointed out that the bootstrap gives more reliable inference when it is used in conjunction with approximately pivotal quantities. In practice, statistics that are supposedly approximately pivotal can have distributions that depend heavily on nuisance parameters. In other contexts, no approximately pivotal quantities can readily be found. The bootstrap can still "work" in such cases, but it cannot be expected to be as reliable as in better circumstances. Observing the Golden Rules proposed here can improve reliability even in these cases.

Selected Bibliography

Andrews, D. W. K. (2002). "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators". *Econometrica*, 70, 119–162.

Athreya, K. B. (1987). "Bootstrap of the mean in the infinite variance case", *Annals of Statistics*, 15, 724–731.

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements", *Journal of the American Statistical Association*, 83, 687–697.

Brown, B. W. and W. Newey (2002). "Generalized method of moments, efficient bootstrapping, and improved inference", *Journal of Business and Economic Statistics*, 20, 507–517.

Davidson, R. and E. Flachaire (2001). "The wild bootstrap, tamed at last", GREQAM Document de Travail 99A32, revised, Marseille, France.

Davidson, R. and J. G. MacKinnon (1999). "Bootstrap testing in nonlinear models". *International Economic Review 40*, 487–508.

Davidson R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*, Oxford University Press.

Davidson R. and J. G. MacKinnon (2006). "Bootstrap methods in econometrics", Chapter 23 of *Palgrave Handbook of Econometrics*, Volume 1, *Econometric Theory*, eds T. C. Mills and K. Patterson, Palgrave-Macmillan, London.

Davidson R. and J. G. MacKinnon (2006b). "Bootstrap inference in a linear equation estimated by instrumental variables", Discussion Paper 1024, Queen's University, Kingston, Ontario.

Dufour, J.-M., (1997). "Some impossibility theorems in econometrics with applications to structural and dynamic models", *Econometrica*, 65, 1365–1387.

Dufour, J.-M., and L. Khalaf (2001). "Monte Carlo test methods in econometrics", Ch. 23 in *A Companion to Econometric Theory*, ed. B. Baltagi, Oxford, Blackwell Publishers, 494–519.

Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses", *Annals of Mathematical Statistics*, 28, 181–187.

Efron, B. (1979). "Bootstrap methods: Another look at the jackknife", *Annals of Statistics*, 7, 1–26.

Eicker, F. (1963). "Asymptotic normality and consistency of the least squares estimators for families of linear regressions", *The Annals of Mathematical Statistics*, 34, 447–456.

Flachaire, E. (1999). "A better way to bootstrap pairs", *Economics Letters*, 64, 257–262.

Foster, J.E., J. Greer and E. Thorbecke (1984). "A class of decomposable poverty measures", *Econometrica*, 52, 761–776.

Freedman, D. A. (1981). "Bootstrapping regression models", *Annals of Statistics*, 9, 1218–1228.

Godambe, V. P. (1960). "An optimum property of regular maximum likelihood estimation", *Annals of Mathematical Statistics*, 31, 1208–11.

Godambe, V. P., and M. E. Thompson (1978). "Some aspects of the theory of estimating equations", *Journal of Statistical Planning and Inference*, 2, 95–104.

Granger, C. W. J. (1969). "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica*, 37, 424–38.

Gregory, A. W., and M. R. Veall (1985). "On formulating Wald tests for nonlinear restrictions", *Econometrica*, 53, 1465–68.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Hansen, B. E. (1999). "The grid bootstrap and the autoregressive model", *Review of Economics and Statistics*, 81, 594–607.

Horowitz, J. L. (2001). "The bootstrap", Ch. 52 in *Handbook of Econometrics* Vol. 5, eds. J. J. Heckman and E. E. Leamer, North-Holland, Amsterdam, 3159–3228.

Horowitz, J. L. (2003). "The bootstrap in econometrics", *Statistical Science*, 18, 211–218.

Hu, F. and J. D. Kalbfleisch (2000). "The estimating function bootstrap", *Canadian Journal of Statistics*, 28, 449–481.

Knuth, D. E. (1998). *The Art of Computer Programming*, Vol 2, *Seminumerical Algorithms*, 3rd edition, Addison-Wesley.

Lafontaine, F., and K. J. White (1986). "Obtaining any Wald statistic you want", *Economics Letters*, 21, 35–40.

Liu, R. Y. (1988). "Bootstrap procedures under some non-I.I.D. models", *Annals of Statistics*, 16, 1696–1708.

Mammen, E. (1993). "Bootstrap and wild bootstrap for high dimensional linear models", *Annals of Statistics*, 21, 255–285.

Owen, A. B. (2001). *Empirical Likelihood*, Chapman and Hall.

Politis, D. N. (2003) "The impact of bootstrap methods on time series analysis", *Statistical Science*, 18, 219–230.

White, H. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", *Econometrica*, 48, 817–838.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis", *Annals of Statistics* 14, 1261–1295.