Bootstrap Testing in Nonlinear Models

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

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

russell@ehess.cnrs-mrs.fr

and

James G. MacKinnon

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

jgm@qed.econ.queensu.ca

Abstract

Bootstrap testing of nonlinear models normally requires at least one nonlinear estimation for every bootstrap sample. We show how to reduce computational costs by performing only a fixed, small number of Newton or quasi-Newton steps for each bootstrap sample. The number of steps is smaller for likelihood ratio tests than for other types of classical tests, and smaller for Newton's method than for quasi-Newton methods. The suggested procedures are applied to tests of slope coefficients in the tobit model and to tests of common factor restrictions. In both cases, bootstrap tests work well, and very few steps are needed.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Joel Horowitz, two referees, and seminar participants at the Universities of British Columbia, Guelph, Montreal, and Rochester, the ESRC Econometric Study Group Conference at Bristol University, and the European Meeting of the Econometric Society at Toulouse for comments on earlier versions.

Final Version, November, 1997

1. Introduction

The bootstrap provides a popular way to perform inference that is more reliable, in finite samples, than inference based on conventional asymptotic theory. In certain circumstances, the bootstrap will yield exact tests. Even when it does not, it will often yield tests that are very close to being exact. If n is the number of observations, the rejection probability for a bootstrap test should never, in regular cases, be in error by more than $O(n^{-1})$, and it will often be in error by only $O(n^{-3/2})$ or $O(n^{-2})$; see Davidson and MacKinnon (1996) for discussion and references. Thus there is good reason to believe that inferences based on bootstrap tests will generally be very accurate.

Although modern computer hardware has greatly reduced the cost of implementing the bootstrap, situations frequently arise in which each bootstrap replication involves nonlinear estimation, and this can often be costly. In this paper, we propose approximate bootstrap methods that achieve the accuracy of bootstrap inference without the need for repeated nonlinear estimation.

Most nonlinear estimation procedures involve a sequence of steps. Existing theoretical results, to be reviewed in the next section, tell us that these steps converge at rates that depend on the sample size, n. Since the bootstrap is normally accurate only up to some order described by a negative power of the sample size, it is enough to achieve the same order of accuracy in the computation of the bootstrap test statistic as is given by the bootstrap itself. Thus, if there are B bootstrap replications, it is only necessary to perform one nonlinear estimation, instead of $B+1$. The remaining B nonlinear estimations would be replaced either by mB Newton steps or by mB OLS estimations of artificial regressions, where m is a small integer. In most cases, one nonlinear estimation plus mB Newton steps or mB OLS regressions will be less costly than $B+1$ nonlinear estimations.

In the next section, we review some results from the theory of nonlinear estimation, which imply that a finite, usually small, number of steps can yield approximations accurate to the same order as the bootstrap itself. Then, in Section 3, we describe in detail how approximate bootstrap tests may be implemented for all of the classical testing procedures: Lagrange Multiplier, Likelihood Ratio, $C(\alpha)$, and Wald. In Sections 4 and 5, we examine the accuracy of these approximate tests in the context of the tobit model and tests of common factor restrictions, respectively.

2. Approximate Nonlinear Estimation

Most of the methods used for performing nonlinear estimation in econometrics are based on Newton's method, which is sometimes referred to as the Newton-Raphson method; for a discussion of these methods, see Quandt (1983). Because of its property of quadratic convergence, Newton's method is particularly effective in the neighborhood of the optimum of the criterion function, which may be a loglikelihood function, a sum of squared residuals, or a more general quadratic form, as in the case of GMM estimation.

Unfortunately, Newton's method, in its pure state, requires knowledge of the Hessian matrix, that is, the matrix of second derivatives of the criterion function,

and these may be difficult or troublesome to obtain. Consequently, other methods that use only first derivatives are in common use. These quasi-Newton methods may be thought of as linearizations of the models being estimated. They employ various approximations to the Hessian instead of the Hessian itself, and they are frequently implemented by means of artificial regressions.

Robinson (1988)[1] provides a number of results concerning the rates of convergence of Newton's method and quasi-Newton methods in the stochastic context. Suppose that we wish to maximize or minimize a random criterion function $Q^n(\boldsymbol{\theta})$, computed using a sample of n observations, with respect to the $k-$vector of parameters $\boldsymbol{\theta}$. It is assumed that $Q^n(\boldsymbol{\theta}) = O_p(1)$ as $n \to \infty$. Let the vector which maximizes Q^n be denoted $\hat{\boldsymbol{\theta}}$. If we denote the starting point for a Newton step, that is, one iteration of Newton's method, by $\boldsymbol{\theta}$, then the step leads to the point

$$\acute{\boldsymbol{\theta}} \equiv \boldsymbol{\theta} - \left(\boldsymbol{H}^n(\boldsymbol{\theta})\right)^{-1} \boldsymbol{g}^n(\boldsymbol{\theta}),$$

where the $k-$vector $\boldsymbol{g}^n(\boldsymbol{\theta})$ denotes the gradient and the $k \times k$ matrix $\boldsymbol{H}^n(\boldsymbol{\theta})$ denotes the Hessian of Q^n at $\boldsymbol{\theta}$. Suppose the starting point for Newton's method, which we denote by $\boldsymbol{\theta}_{(0)}$, is such that $\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}} = O_p(n^{-1/2})$. Then, if $\boldsymbol{\theta}_{(i)}$ is the result of i iterations of Newton's method, one of Robinson's results shows that

$$\boldsymbol{\theta}_{(i)} - \hat{\boldsymbol{\theta}} = O_p(n^{-2^{i-1}}). \tag{1}$$

This result simply expresses the quadratic convergence of Newton's method in the context of random functions. The regularity needed for (1) is the existence of third partial derivatives of $Q^n(\boldsymbol{\theta})$ that are $O_p(1)$ in a neighborhood of $\hat{\boldsymbol{\theta}}$.

When a quasi-Newton method is used, the Hessian is generally approximated with an error that is of order $n^{-1/2}$ in probability. In this case, under the same regularity conditions, Robinson shows that, if $\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}} = O_p(n^{-1/2})$ as before,

$$\boldsymbol{\theta}_{(i)} - \hat{\boldsymbol{\theta}} = O_p(n^{-(i+1)/2}). \tag{2}$$

Thus using a quasi-Newton method means that convergence is no longer quadratic. For large values of i, the difference between (1) and (2) can be very great. Even so, (2) implies that one gains an order of $n^{-1/2}$ at each step of the procedure.

It is quite easy to see how the result (2) arises when estimation is based on an artificial regression. As above, the criterion function $Q^n(\boldsymbol{\theta})$ is $O_p(1)$ as $n \to \infty$, and $\boldsymbol{\theta}$ is a $k-$vector. The artificial regression may be written as

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}. \tag{3}$$

Here the regressand, $\boldsymbol{r}(\boldsymbol{\theta})$, is a column vector, and the matrix of regressors, $\boldsymbol{R}(\boldsymbol{\theta})$, is a matrix with k columns. The length of the vector $\boldsymbol{r}(\boldsymbol{\theta})$ will often be n, but sometimes it will be an integer multiple of n. "Residuals" is used here as a neutral term to avoid any implication that (3) is a statistical model.

---

[1] We are grateful to Joel Horowitz for drawing our attention to this paper

Our discussion generalizes the theory of artificial regressions, which was developed for models estimated by maximum likelihood in Davidson and MacKinnon (1990), to the case of M-estimation. We suppose that the 'true" parameter vector is $\boldsymbol{\theta}_0$. By this it is meant that $\boldsymbol{\theta}_0$ maximizes or minimizes $\operatorname{plim}_{n\to\infty} Q^n(\boldsymbol{\theta})$. The artificial variables $\boldsymbol{r}(\boldsymbol{\theta})$ and $\boldsymbol{R}(\boldsymbol{\theta})$ must satisfy the following conditions:

(i) $n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta}) = \pm\boldsymbol{g}^n(\boldsymbol{\theta})$

(ii) if $\boldsymbol{\theta}_{(0)} - \boldsymbol{\theta}_0 = O_p(n^{-1/2})$, then $n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta}_{(0)})\boldsymbol{R}(\boldsymbol{\theta}_{(0)}) \pm \boldsymbol{H}^n(\boldsymbol{\theta}_0) = O_p(n^{-1/2})$.

The choice of sign depends on whether Q^n is maximized (+), or minimized (−).

The one-step estimator $\boldsymbol{\theta}_{(1)}$ defined by the artificial regression (3) with $\boldsymbol{\theta}_{(0)}$ as starting point is obtained by running (3) with the variables evaluated at $\boldsymbol{\theta}_{(0)}$, obtaining the artificial OLS estimates $\boldsymbol{b}_{(0)}$, and setting $\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{(0)} + \boldsymbol{b}_{(0)}$. Denoting $\boldsymbol{R}(\boldsymbol{\theta}_{(0)})$ and $\boldsymbol{r}(\boldsymbol{\theta}_{(0)})$ by $\boldsymbol{R}_{(0)}$ and $\boldsymbol{r}_{(0)}$ respectively, we have

$$\boldsymbol{b}_{(0)} = \big(n^{-1}\boldsymbol{R}_{(0)}{}^\top\boldsymbol{R}_{(0)}\big)^{-1} n^{-1}\boldsymbol{R}_{(0)}{}^\top\boldsymbol{r}_{(0)}, \tag{4}$$

from which it is clear from (i) and (ii) above that the artificial regression implements a quasi-Newton step.

Now write $\boldsymbol{H}_{(0)} = \boldsymbol{H}(\boldsymbol{\theta}_{(0)})$ and $\boldsymbol{g}_{(0)} = \boldsymbol{g}(\boldsymbol{\theta}_{(0)})$. Then, since $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$ by the first-order conditions for an optimum, a short Taylor expansion gives

$$\boldsymbol{g}_{(0)} = -\big(\boldsymbol{H}_{(0)} + O_p(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}})\big)(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}).$$

By (i) and (ii) above, whichever sign is used, this can be written as

$$n^{-1/2}\boldsymbol{R}_{(0)}{}^\top\boldsymbol{r}_{(0)} = \big(n^{-1}\boldsymbol{R}_{(0)}{}^\top\boldsymbol{R}_{(0)} + O_p(n^{-1/2})\big)\, n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{(0)}). \tag{5}$$

But (4) can be written as

$$n^{-1/2}\boldsymbol{R}_{(0)}{}^\top\boldsymbol{r}_{(0)} = \big(n^{-1}\boldsymbol{R}_{(0)}{}^\top\boldsymbol{R}_{(0)}\big)\, n^{1/2}\boldsymbol{b}_{(0)}. \tag{6}$$

Subtracting (5) from (6), and recalling that $n^{-1}\boldsymbol{R}_{(0)}{}^\top\boldsymbol{R}_{(0)} = O_p(1)$, we find that

$$\boldsymbol{\theta}_{(1)} - \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{(0)} + \boldsymbol{b}_{(0)} - \hat{\boldsymbol{\theta}} = O_p\big(n^{-1/2}(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}})\big) = O_p(n^{-1}).$$

Thus we see that one step, from $\boldsymbol{\theta}_{(0)}$ to $\boldsymbol{\theta}_{(1)}$, has led to a gain of order $n^{-1/2}$. Repeating the argument with $\boldsymbol{\theta}_{(0)}$ replaced by $\boldsymbol{\theta}_{(i)}$ at step i shows that every step yields a gain of order $n^{-1/2}$. This is a special case of the result (2) that was proved by Robinson.

For the approximate bootstrap procedures to be considered in the next section, we actually need something slightly different from the closeness in probability to some order of the approximants $\boldsymbol{\theta}_{(i)}$ to $\hat{\boldsymbol{\theta}}$. The accuracy of bootstrap probabilities, or P values, is measured by the probability that a statistic, which could be computed as a function either of $\hat{\boldsymbol{\theta}}$ or of $\boldsymbol{\theta}_{(i)}$, is in some specified rejection region. Thus, what we wish to ensure is that the probabilities computed with $\hat{\boldsymbol{\theta}}$ and the probabilities computed with $\boldsymbol{\theta}_{(i)}$ be close to the desired order.

This problem is briefly considered by Robinson, in the context of what he calls "higher-order efficiency comparisons". In Theorem 7 of Robinson (1988), it is shown that, if $\boldsymbol{\theta}_{(i)} - \hat{\boldsymbol{\theta}} = O_p(n^{-(i+1)/2})$, a sufficient condition for the rejection probabilities of statistics based on $\boldsymbol{\theta}_{(i)}$ and $\hat{\boldsymbol{\theta}}$ to differ only at order $n^{-(i+1)/2}$ is that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ admit a density uniformly bounded for large n, and that

$$\lim_{n\to\infty} n^{i/2} \Pr\Big(n^{(i+1)/2} \log n \big\|\boldsymbol{\theta}_{(i)} - \hat{\boldsymbol{\theta}}\big\| \geq 1\Big) = 0.$$

This condition can be guaranteed, as Robinson shows, if $n^{(i+1)/2}(\boldsymbol{\theta}_{(i)} - \hat{\boldsymbol{\theta}})$ has uniformly bounded moments of sufficiently high order, but also under other sorts of conditions. We will not investigate this matter further here. We simply assume that rejection probabilities differ at the same order as the order in probability of the difference between the statistics themselves.

3. Approximate Bootstrapping

Consider a parametrized model for which the parameter vector $\boldsymbol{\theta}$ can be partitioned as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 \vdots \boldsymbol{\theta}_2]$, where $\boldsymbol{\theta}_1$ is a k_1-vector, $\boldsymbol{\theta}_2$ is a k_2-vector, and $k_1 + k_2 = k$. Suppose, without loss of generality, that the null hypothesis we wish to test is that $\boldsymbol{\theta}_2 = \mathbf{0}$. To test it, a wide variety of tests is available — see, for instance, Davidson and MacKinnon (1993) and Newey and McFadden (1994). Here we will consider likelihood ratio (LR) tests, Lagrange multiplier (LM) tests, $C(\alpha)$ tests, and Wald tests, all of which are available for models estimated by maximum likelihood, and most of which are available more generally. Most of these tests have numerous variants, distinguished by the use of different estimators of the asymptotic covariance matrix, and, perhaps, by different parametrizations of the model. All of these tests may be bootstrapped in order to improve their finite-sample properties.

While there are several procedures that may be used for bootstrapping test statistics, our preferred procedure is to compute a bootstrap P value corresponding to the observed value of a test statistic. Let this value be $\hat{\tau}$, and suppose for simplicity that we want to reject the null when $\hat{\tau}$ is sufficiently large. For the class of models that we are discussing here, the bootstrap procedure works as follows:

1. Compute the test statistic $\hat{\tau}$ and a vector of parameter estimates $\tilde{\boldsymbol{\theta}} \equiv [\tilde{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$ that satisfy the null hypothesis.
2. Using a bootstrap DGP based on the parameter vector $\tilde{\boldsymbol{\theta}}$, generate B bootstrap samples, each of size n. Use each bootstrap sample to compute a bootstrap test statistic, say τ_j^*, for $j = 1, \ldots, B$, in the same way as $\hat{\tau}$ was computed from the real data.
3. Calculate the estimated bootstrap P value $\hat{p}^*$ as the proportion of bootstrap samples for which τ_j^* exceeds $\hat{\tau}$. If a formal test at level α is desired, reject the null hypothesis whenever $\hat{p}^* < \alpha$.

At step 2, various different bootstrap DGPs can be used in various contexts. A parametric bootstrap is usually appropriate if we are using ML estimation, because then a DGP is completely characterized by a parameter vector, and so we can

simply draw from the DGP characterized by $\tilde{\boldsymbol{\theta}}$. With other estimation methods, such as least squares or GMM, a nonparametric bootstrap involving resampling will usually be used, but the bootstrap DGP will generally depend on $\tilde{\boldsymbol{\theta}}$ to some extent.

The bootstrap procedure we have just described is remarkably simple, and it often works remarkably well; see Davidson and MacKinnon (1996) and Sections 4 and 5 below for evidence on this point. Of course, other procedures are also perfectly possible and also work well; see, for example, Horowitz (1994).

The easiest theoretical analysis of the bootstrap assumes that B is infinite. In practice, when B is finite, the performance of the bootstrap will fall short of its theoretical performance, for two reasons. First, the estimated bootstrap P value $\hat{p}^*$ will not equal the true bootstrap P value p^*. Second, there will be some loss of power. Both of these issues have been dealt with in the literature on Monte Carlo testing; see Davidson and MacKinnon (1997) for discussion and references. If B is to be chosen as a fixed number, which is the easiest but not the most computationally efficient approach, it should be chosen so that $\alpha(B+1)$ is an integer for any level α that may be of interest, and so that it is not too small.

The three-step procedure laid out above requires the computation of B bootstrap test statistics, $\tau_j^*,\ j = 1, \ldots, B$. If nonlinear estimation is involved, this may be costly. In the remainder of this section, we show how computational cost may be reduced by using approximations to the τ_j^* based on a small number of steps of Newton's method or a quasi-Newton method. We discuss LM, LR, $C(\alpha)$, and Wald tests in that order.

Bootstrapping LM tests

For a classical LM test, the criterion function is the loglikelihood function, and the test statistic is based on the estimates obtained by maximizing it subject to the restrictions of the null hypothesis, that is, $\boldsymbol{\theta}_2 = \mathbf{0}$. However, the test statistic is expressed in terms of the gradient and Hessian of the loglikelihood of the full unrestricted model. One form of the statistic is

$$LM = -\boldsymbol{g}^\top(\tilde{\boldsymbol{\theta}})\boldsymbol{H}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{g}(\tilde{\boldsymbol{\theta}}), \tag{7}$$

where $\boldsymbol{g}(\tilde{\boldsymbol{\theta}})$ and $\boldsymbol{H}(\tilde{\boldsymbol{\theta}})$ are, respectively, the gradient and Hessian of the unrestricted loglikelihood, evaluated at the restricted estimates $\tilde{\boldsymbol{\theta}}$. Note that the $\boldsymbol{g}$ and $\boldsymbol{H}$ of Section 2 would be the quantities in (7) divided by n, since, in order to have a criterion function that is $O_p(1)$ as $n \to \infty$, we must divide the loglikelihood by n. Here and subsequently, we use the more intuitive notation of (7).

Even in the ML context, there are numerous variants of (7) which replace $-\boldsymbol{H}(\tilde{\boldsymbol{\theta}})$ by some other consistent estimator of the information matrix. For instance, if an artificial regression satisfying (i) and (ii) of the preceding section were used, the test statistic would be

$$\boldsymbol{r}^\top(\tilde{\boldsymbol{\theta}})\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\big(\boldsymbol{R}^\top(\tilde{\boldsymbol{\theta}})\boldsymbol{R}(\tilde{\boldsymbol{\theta}})\big)^{-1}\boldsymbol{R}^\top(\tilde{\boldsymbol{\theta}})\boldsymbol{r}(\tilde{\boldsymbol{\theta}}). \tag{8}$$

More generally, $\boldsymbol{g}$ can be the gradient of whatever criterion function is used to estimate $\boldsymbol{\theta}$, and $\boldsymbol{H}$ or $-\boldsymbol{H}$ will be some appropriate estimate of the covariance matrix

of this gradient. The details of the construction of the LM statistic have no bearing on the approximate bootstrap procedure, although some variants of the statistic will have better finite-sample properties than others, even after bootstrapping.

Bootstrapping the LM statistic is conceptually straightforward, but in step 2 it involves estimating the model B additional times under the null hypothesis. We propose to replace the nonlinear estimation by a predetermined, finite, usually small, number of Newton or quasi-Newton steps, starting from the estimates given by the real data. The justification of such a scheme is as follows. It has been known for some time that bootstrap tests will not be exact, on account of the difference between the bootstrap DGP, say $\tilde{\mu}$, and the true unknown DGP, say μ_0, that actually generated the data. A key paper is Beran (1988); see Hall (1992) for a full treatment and references to earlier work. The rejection probability for the bootstrap test at any given level α will be distorted by an amount that depends on the joint distribution of the test statistic and the (random) level-α critical value for that statistic under the bootstrap DGP $\tilde{\mu}$; see Davidson and MacKinnon (1996) for a full discussion. It is usually possible to determine an integer l such that the rejection probability for the bootstrap test at nominal level α differs from α by an amount that is $O(n^{-l/2})$; typically, $l = 3$ or $l = 4$. This being so, the same order of accuracy will be achieved even if there is an error that is $O_p(n^{-l/2})$ in the computation of the bootstrap P values.

The "true" value of the parameters for the bootstrap DGP is $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$. If we denote the fully nonlinear estimates from a bootstrap sample by $\tilde{\boldsymbol{\theta}}_1^*$, then, by construction, we have that $\tilde{\boldsymbol{\theta}}_1^* - \tilde{\boldsymbol{\theta}}_1 = O_p(n^{-1/2})$. Thus $\tilde{\boldsymbol{\theta}}_1$ is a suitable starting point for Newton's method or a quasi-Newton method applied to the restricted model. Notice that the gradient and Hessian involve derivatives with respect to the components of $\boldsymbol{\theta}_1$ only. If the exact Hessian is used, then, by the result (1), the successive estimates $\tilde{\boldsymbol{\theta}}_{1(i)}^*$, $i = 0, 1, 2, \ldots$, satisfy

$$\tilde{\boldsymbol{\theta}}_{1(i)}^* - \tilde{\boldsymbol{\theta}}_1^* = O_p(n^{-2^{i-1}}). \tag{9}$$

If an approximate Hessian is used, then, by (2), they instead satisfy

$$\tilde{\boldsymbol{\theta}}_{1(i)}^* - \tilde{\boldsymbol{\theta}}_1^* = O_p(n^{-(i+1)/2}). \tag{10}$$

The successive approximations to the LM statistic are defined by

$$LM(\tilde{\boldsymbol{\theta}}_{(i)}^*) \equiv -\boldsymbol{g}^\top(\tilde{\boldsymbol{\theta}}_{(i)}^*)\boldsymbol{H}^{-1}(\tilde{\boldsymbol{\theta}}_{(i)}^*)\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{(i)}^*), \tag{11}$$

where the functions $\boldsymbol{g}$ and $\boldsymbol{H}$ are the same as the ones used to compute the actual test statistic, and where $\tilde{\boldsymbol{\theta}}_{(i)}^* = [\tilde{\boldsymbol{\theta}}_{1(i)}^* \vdots \mathbf{0}]$. For instance, if an artificial regression corresponding to the unrestricted model is used, then it can be seen from (8) that $LM(\tilde{\boldsymbol{\theta}}_{(i)}^*)$ is the explained sum of squares, or n times the R^2, from this regression; see Davidson and MacKinnon (1990) for details.

At this point, a little care is necessary. Recall that the statistic (7) is defined in terms of functions that are not $O_p(1)$. We therefore rewrite (11) as

$$\big(n^{-1/2}\boldsymbol{g}^\top(\tilde{\boldsymbol{\theta}}_{(i)}^*)\big)\big(-n^{-1}\boldsymbol{H}^{-1}(\tilde{\boldsymbol{\theta}}_{(i)}^*)\big)\big(n^{-1/2}\boldsymbol{g}(\tilde{\boldsymbol{\theta}}_{(i)}^*)\big), \tag{12}$$

where each factor in parentheses is $O_p(1)$. We have

$$
\begin{aligned}
n^{-1}\boldsymbol{H}(\tilde{\boldsymbol{\theta}}^*_{(i)}) &= n^{-1}\boldsymbol{H}(\tilde{\boldsymbol{\theta}}) + O_p(\tilde{\boldsymbol{\theta}}^*_{(i)} - \tilde{\boldsymbol{\theta}}), \text{ and} \\
n^{-1/2}\boldsymbol{g}(\tilde{\boldsymbol{\theta}}^*_{(i)}) &= n^{-1/2}\boldsymbol{g}(\tilde{\boldsymbol{\theta}}) + n^{1/2}O_p(\tilde{\boldsymbol{\theta}}^*_{(i)} - \tilde{\boldsymbol{\theta}}).
\end{aligned} \tag{13}
$$

Note that $n^{-1/2}\boldsymbol{g}(\boldsymbol{\theta}) = O_p(1)$ whenever $n^{1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = O_p(1)$, where $\hat{\boldsymbol{\theta}}$ maximizes the unrestricted loglikelihood, so that $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$. Substituting (13) into (12), we find that

$$
LM(\tilde{\boldsymbol{\theta}}^*_{(i)}) = LM(\tilde{\boldsymbol{\theta}}) + n^{1/2}O_p(\tilde{\boldsymbol{\theta}}^*_{(i)} - \tilde{\boldsymbol{\theta}}). \tag{14}
$$

This is the key result for LM tests.

For Newton's method, when (9) applies, we see from (14) that the difference between $LM(\tilde{\boldsymbol{\theta}}^*_{(i)})$ and $LM(\tilde{\boldsymbol{\theta}})$ is of order $n^{-(2^i-1)/2}$. For the quasi-Newton case, when (10) applies, it is of order $n^{-i/2}$. After just one iteration, when $i = 1$, the difference is of order $n^{-1/2}$ in both cases. Subsequently, the difference diminishes much more rapidly for Newton's method than for quasi-Newton methods. In general, if bootstrap P values are in error at order $n^{-l/2}$, and we are using Newton's method with an exact Hessian, the number of steps needed to achieve at least the same order of accuracy as the bootstrap, m, should be chosen so that $2^m - 1 \geq l$. Thus, for $l = 3$, the smallest suitable m is 2, and for $l = 4$, it is 3. If we are using a quasi-Newton method, we simply need to choose m so that $m \geq l$.

Let us now recapitulate our proposed procedure for bootstrapping an LM test without doing any nonlinear estimations for the bootstrap samples. The procedure is a variant of the general procedure discussed above, and steps 1 and 3 need no further explanation. Step 2 becomes

2a. Draw B bootstrap samples of size n from a bootstrap DGP characterized by $\tilde{\boldsymbol{\theta}}$. Choose the number of Newton steps, m, as a function of the integer l that characterizes the order of the error of the bootstrap P value. If Newton's method is used, $m \geq \log_2(l+1)$; if a quasi-Newton method, $m \geq l$.

2b. For each bootstrap sample, say the j^{th}, perform m Newton steps for the restricted model starting from $\tilde{\boldsymbol{\theta}}^*_{1(0)} = \tilde{\boldsymbol{\theta}}_1$ in order to obtain the iterated estimates $\tilde{\boldsymbol{\theta}}^*_{1(m)}$.

2c. Compute the bootstrap LM statistic $\tau^*_{j(m)} = LM(\tilde{\boldsymbol{\theta}}^*_{(m)})$ using (11).

Bootstrapping LR tests

Likelihood Ratio tests are particularly expensive to bootstrap, because two nonlinear optimizations must normally be performed for each bootstrap sample. However, both of these can be replaced by a small number of Newton steps, starting from the restricted estimates $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_1 \vdots \boldsymbol{0}]$. Under the null, this is done exactly as in step (2b) above for an LM test. Under the alternative, the only difference is that the gradient and Hessian correspond to the unrestricted model, and thus involve derivatives with respect to all components of $\boldsymbol{\theta}$. The starting point for the unrestricted model may well be the same as for the restricted model, but it is probably preferable to start from the endpoint of the restricted iterations, $\tilde{\boldsymbol{\theta}}^*_{(m)} = [\tilde{\boldsymbol{\theta}}^*_{1(m)} \vdots \boldsymbol{0}]$. This endpoint contains possibly relevant information about

the current bootstrap sample, and the difference between it and the unrestricted bootstrap fully nonlinear estimate $\hat{\boldsymbol{\theta}}^*$ is $O_p(n^{-1/2})$, as required.

In the classical ML context, the criterion function to be maximized may be taken as twice the loglikelihood function. The LR statistic is then the difference between the unconstrained maximum with respect to $\boldsymbol{\theta}$ and the constrained maximum with respect to $\boldsymbol{\theta}_1$ only, with $\boldsymbol{\theta}_2 = \mathbf{0}$. In more general contexts, test statistics are often available which can be expressed as the difference between the unconstrained and constrained maxima or minima of a criterion function. A well-known example is Hansen's J statistic for overidentifying restrictions; see Hansen (1982). To avoid confusion with the $O_p(1)$ criterion function Q^n of Section 2, we will let $\ell(\boldsymbol{\theta})$ denote the value of the loglikelihood function; $\boldsymbol{g}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta})$ will denote respectively the gradient and the Hessian of $\ell(\boldsymbol{\theta})$.

For each bootstrap sample, we can compute a bootstrap LR statistic. The true value of this statistic is $2\big(\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\tilde{\boldsymbol{\theta}}^*)\big)$. Consider replacing $\hat{\boldsymbol{\theta}}^*$ by an approximation $\acute{\boldsymbol{\theta}}$ such that $\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^* = O_p(n^{-1/2})$. Since $\boldsymbol{g}(\hat{\boldsymbol{\theta}}^*) = \mathbf{0}$ by the first-order conditions for maximizing $\ell(\boldsymbol{\theta})$, a Taylor expansion gives

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\acute{\boldsymbol{\theta}}) = -\frac{1}{2}\big(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^*\big)^\top \boldsymbol{H}(\bar{\boldsymbol{\theta}})\big(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^*\big),$$

where $\bar{\boldsymbol{\theta}}$ is a convex combination of $\acute{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^*$. Since $\boldsymbol{H}$ is $O_p(n)$, it follows that

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\acute{\boldsymbol{\theta}}) = nO_p\big((\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^*)^2\big). \tag{15}$$

The above result is true for both the restricted and unrestricted loglikelihoods, and is therefore true as well for the LR statistic.

Now recall the results (1) and (2), which give the rate of convergence for Newton's method and quasi-Newton methods, respectively. From (1) and (15), we see that, if Newton's method is used,

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\hat{\boldsymbol{\theta}}^*_{(i)}) = n^{-2^i+1}. \tag{16}$$

From (2) and (15), we see that, if a quasi-Newton method is used,

$$\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\hat{\boldsymbol{\theta}}^*_{(i)}) = n^{-i}. \tag{17}$$

These results imply that, for both Newton and quasi-Newton methods when $l = 3$ and $l = 4$, the minimum number of steps m for computing $\hat{\boldsymbol{\theta}}^*_{(m)}$ and $\tilde{\boldsymbol{\theta}}^*_{(m)}$ needed to ensure that the error in the LR statistic is at most of order $n^{-l/2}$ is just 2.

The results (16) and (17) suggest that, for a given number of steps, the approximation will be better for LR statistics than for LM statistics. The reason for this, of course, is that the loglikelihood functions, restricted and unrestricted, are locally flat at their respective maxima, and hence they are less sensitive to slight errors in the point at which they are evaluated than is the LM statistic, which depends on the slope of the unrestricted loglikelihood function at the restricted estimates.

The result that $m = 2$ is always sufficient to ensure that the approximation error for the LR statistic is of no higher order than the size distortion of the bootstrap test is a striking one. Simulation evidence to be presented in Sections 4 and 5 suggests that it does hold in at least some interesting cases. In contrast, for quasi-Newton methods, LM tests will often require $m = 4$. Therefore, when the approximate bootstrap is used, it may be no more expensive to bootstrap LR tests than to bootstrap LM tests.

Bootstrapping $C(\alpha)$ tests

It can sometimes be convenient to use $C(\alpha)$ tests when maximum likelihood estimation is difficult. All that is needed for a $C(\alpha)$ test is a set of root-n consistent estimates of the parameters of the null hypothesis, which we may denote as $\acute{\boldsymbol{\theta}} = [\acute{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$. These tests were introduced by Neyman (1959); further discussion can be found in Smith (1987) and Dagenais and Dufour (1991). They were extended to the GMM context in Davidson and MacKinnon (1993), where it was shown that they may be computed by artificial regression. Quite generally, a $C(\alpha)$ statistic can be expressed, in the notation of (7), as

$$C(\alpha) = -\boldsymbol{g}^\top(\acute{\boldsymbol{\theta}})\boldsymbol{H}^{-1}(\acute{\boldsymbol{\theta}})\boldsymbol{g}(\acute{\boldsymbol{\theta}}) + \boldsymbol{g}_1^\top(\acute{\boldsymbol{\theta}})\big(\boldsymbol{H}_{11}(\acute{\boldsymbol{\theta}})\big)^{-1}\boldsymbol{g}_1(\acute{\boldsymbol{\theta}}),$$

where $\boldsymbol{g}_1(\boldsymbol{\theta})$ is the gradient of the criterion function with respect to $\boldsymbol{\theta}_1$ only, and $\boldsymbol{H}_{11}(\boldsymbol{\theta})$ is the corresponding block of the Hessian. As with LM tests, there are numerous variants of $C(\alpha)$ tests available, depending on how $\boldsymbol{H}$ is estimated.

Bootstrapping $C(\alpha)$ tests clearly requires no Newton steps, since the criterion function is not optimized. However, it is desirable to take at least one Newton or quasi-Newton step for the restricted model, since $\acute{\boldsymbol{\theta}}_1$ is not in general an asymptotically efficient estimator of $\boldsymbol{\theta}_1$. The one-step estimator $\acute{\boldsymbol{\theta}}_{1\,(1)}$ is asymptotically efficient, and it is therefore preferable to base the bootstrap DGP on it rather than on $\acute{\boldsymbol{\theta}}_1$.

As long as the procedure for obtaining the initial root-n consistent estimate $\acute{\boldsymbol{\theta}}_1$ is not too computationally demanding, bootstrapping a $C(\alpha)$ test will generally be less time-consuming than bootstrapping an LM or LR test.

Bootstrapping Wald and Wald-like tests

Wald tests tend to have poor finite-sample properties, in part because they are not invariant under nonlinear reparametrizations of the restrictions under test; see, among others, Gregory and Veall (1985, 1987) and Phillips and Park (1988). This suggests that it may be particularly important to bootstrap them.

Although the Wald test statistic itself is based entirely on the unrestricted estimates $\hat{\boldsymbol{\theta}}$, estimates that satisfy the null hypothesis must be obtained in order to generate the bootstrap samples. For this purpose, it is probably best to use the restricted estimates $\tilde{\boldsymbol{\theta}}$. However, since the Wald test is often used precisely because $\tilde{\boldsymbol{\theta}}$ is hard to compute, it is desirable to consider other possibilities. Estimation of the unrestricted model gives parameter estimates $\hat{\boldsymbol{\theta}} \equiv [\hat{\boldsymbol{\theta}}_1 \vdots \hat{\boldsymbol{\theta}}_2]$, and $\hat{\boldsymbol{\theta}}_1$ is a vector of root-n consistent, but inefficient, estimates of the parameters of the restricted model. One possibility is thus to proceed as for a $C(\alpha)$ test, using $\hat{\boldsymbol{\theta}}_1$ in place of $\acute{\boldsymbol{\theta}}_1$

in the procedure discussed above. Thus the bootstrap samples would be generated using one-step efficient estimates with $\hat{\boldsymbol{\theta}}_1$ as starting point.

If we compute a $C(\alpha)$ test based on $\hat{\boldsymbol{\theta}}_1$ as the initial root-n consistent estimates of the restricted model, we obtain a test that may be thought of as a "Wald-like" test. However, it may be difficult to obtain such root-n consistent estimates when it is not easy to partition the parameter vector so as to make all the restrictions zero restrictions. In such cases, it may be easier to bootstrap one of the variants of the Wald test itself. Whether one uses a $C(\alpha)$ Wald-like test or a true Wald test, bootstrapping requires us to obtain the unrestricted estimates for each bootstrap sample, or approximations to them based on Newton's method.

The number of steps needed for a given degree of approximation to a Wald statistic can be determined by considering the case in which the restrictions take the form $\boldsymbol{\theta}_2 = \mathbf{0}$. In that case, the Wald statistic can be written as

$$W(\hat{\boldsymbol{\theta}}) = (n^{1/2}\hat{\boldsymbol{\theta}}_2)^\top \hat{\boldsymbol{V}}^{-1}(n^{1/2}\hat{\boldsymbol{\theta}}_2), \tag{18}$$

where $\hat{\boldsymbol{V}}$ is a consistent estimate, based on $\hat{\boldsymbol{\theta}}$, of the asymptotic covariance matrix of $n^{1/2}\hat{\boldsymbol{\theta}}_2$. $\hat{\boldsymbol{V}}$ is thus $O_p(1)$. It may be obtained from any suitable estimate of the information matrix, including the one provided by an artificial regression corresponding to the unrestricted model. If (18) is approximated by $W(\acute{\boldsymbol{\theta}})$, then, by the same arguments as those leading to (14), the approximation error is clearly of order $n^{1/2}O_p(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$. Thus the number of Newton or quasi-Newton steps needed to obtain a given degree of accuracy for (18) is the same as for an LM test.

In the next two sections, we provide some simulation evidence on how well the procedures proposed in this paper actually perform in finite samples. Section 4 deals with the tobit model, and Section 5 deals with tests of common factor restrictions.

4. Bootstrap Testing in the Tobit Model

The simulations discussed in this section concern tests of slope coefficients in a tobit model (Tobin, 1958; Amemiya, 1973). The tobit model is an interesting one to study for several reasons. It is a nonlinear model, which means that bootstrapping is not cheap, but it has a well-behaved loglikelihood function, so that bootstrapping is not prohibitively expensive either. The model can be estimated by Newton's method and by at least two different quasi-Newton methods, and all of the classical tests are readily available.

The model we study is

$$\begin{aligned} y_t' &= \boldsymbol{X}_{1t}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}\boldsymbol{\beta}_2 + u_t, \;\; u_t \sim N(0, \sigma^2), \\ y_t &= \max(0, y_t'), \end{aligned}$$

where y_t is observed but y_t' is not, $\boldsymbol{X}_{1t}$ is a $1 \times k_1$ vector of observations on exogenous regressors, one of which is a constant term, and $\boldsymbol{X}_{2t}$ is a $1 \times k_2$ vector

of observations on other exogenous variables. The loglikelihood function for this model is

$$(19)\quad \sum_{y_t=0} \log\Biggl(\Phi\Bigl(-\tfrac{1}{\sigma}\bigl(\boldsymbol{X}_{1t}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}\boldsymbol{\beta}_2\bigr)\Bigr)\Biggr) + \sum_{y_t>0} \log\Biggl(\tfrac{1}{\sigma}\phi\Bigl(\tfrac{1}{\sigma}\bigl(y_t - \boldsymbol{X}_{1t}\boldsymbol{\beta}_1 - \boldsymbol{X}_{2t}\boldsymbol{\beta}_2\bigr)\Bigr)\Biggr),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative standard normal distribution functions, respectively. Since σ has to be estimated, the total number of parameters is $k_1 + k_2 + 1$. The null hypothesis is that $\boldsymbol{\beta}_2 = \mathbf{0}$.

We consider five different test statistics. The first is the LR statistic, which is twice the difference between (19) evaluated at $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\sigma})$ and at $(\tilde{\boldsymbol{\beta}}_1, \mathbf{0}, \tilde{\sigma})$. The second is the efficient score (ES) form of the LM statistic, which uses the true information matrix evaluated at the restricted estimates. Orme (1995) has recently proposed an ingenious, but rather complicated, double-length artificial regression for tobit models; when it is evaluated at the restricted estimates, its explained sum of squares (ESS) is equal to the ES form of the LM statistic. The third test statistic is the outer product of the gradient (OPG) form of the LM test, which may also be computed via an artificial regression. The regressand is a vector of 1s, the t^{th} element of each of the regressors is the derivative with respect to one of the parameters of the term in (19) that corresponds to observation t, and the test statistic is the ESS; see Davidson and MacKinnon (1993, Chapter 13). The fourth test statistic is a Wald test, using minus the numerical Hessian as the covariance matrix. Because it is based on the standard parametrization used above, we call it W_β. The fifth test statistic, which we call W_γ, is based on an alternative parametrization in which the parameters are $\boldsymbol{\gamma} \equiv \boldsymbol{\beta}/\sigma$ and $\delta \equiv 1/\sigma$. This alternative parametrization was investigated by Olsen (1978), and our program for ML estimation of tobit models uses it. Computer programs (in Fortran 77) for tobit estimation and for all of the test statistics are available via the Internet from the following Web page: `http://www.econ.queensu.ca/pub/faculty/mackinnon`.

In most of our experiments, each of the exogenous variables was independently distributed as $N(0, 1)$. We did try other distributions, and the results were not sensitive to the choice. Since, under the null hypothesis, it is only through the subspace spanned by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ jointly that the exogenous variables in $\boldsymbol{X}_2$ affect the test statistics, there is no loss of generality in assuming that the two sets of exogenous variables are uncorrelated with each other. We consider six different pairs of values of k_1 and k_2: $(2, 2)$, $(2, 5)$, $(2, 8)$, $(5, 5)$, $(5, 8)$, and $(8, 2)$. For any value of k_1, each DGP is characterized by a constant term β_c, a slope coefficient β_s, which is the same for elements 2 through k_1 of $\boldsymbol{\beta}_1$, and a variance σ^2. Depending on the sample size, k_1, k_2, and the parameter values, it is possible for there to be so few nonzero values of y_t that it is impossible to obtain ML estimates, at least under the alternative. Samples for which the number of nonzero y_t was less than $k_1 + k_2 + 1$ were therefore dropped. We tried to design the experiments so that this would not happen very often.

The initial experiments were designed to see how well the five tests perform without bootstrapping and how their performance depends on parameter values. It turns out that the value of σ is particularly important. Figure 1 plots the performance of all five tests against σ for the case in which $k_1 = 5$ and $k_2 = 8$,

with $n = 50$, $\beta_c = 0$, and $\beta_s = 1$. We chose $n = 50$ for this experiment because it is the smallest sample size for which we did not have to drop very many samples. The vertical axis shows the fraction of the time that an asymptotic test at the nominal .05 level actually leads to rejection. The tests that perform worst are the OPG form of the LM test and the W_β test, and the test that performs best is the ES form of the LM test.

[Figure 1 about here]

Figure 1 deals with only one case. Figures 2 and 3 plot the performance of the LR test and the ES LM test, respectively, for all six cases. Note the different scales of the vertical axis. In contrast to the LR test and the other tests, the ES LM test performs much less well in the case $k_1 = 8, k_2 = 2$ than in the case $k_1 = 5, k_2 = 8$. Nevertheless, it is always the test that performs best in these experiments. Figures 2 and 3 suggest that Monte Carlo experiments in which either k_1 or k_2 is always small may yield very misleading results about the finite-sample performance of asymptotic tests in the tobit and related models.

[Figure 2 about here]

[Figure 3 about here]

We also experimented with changing the constant term, β_c, and the slope coefficient, β_s. All of the tests tend to overreject more severely as β_c falls and the number of zero observations consequently increases. Increasing β_s is essentially equivalent to reducing σ and shrinking β_c towards zero in such a way that β_c/σ remains constant, and the results were therefore predictable.

Figure 4 shows how the rejection frequencies for all the tests at the nominal .05 level vary with n, for $k_1 = 5$, $k_2 = 8$, $\beta_c = 1$, $\beta_s = 1$, and $\sigma = 1$. We used $\beta_c = 1$ because we would have had to omit a great many samples for the smaller values of n if we had used $\beta_c = 0$. It is clear from the figure that all the tests approach their asymptotic distributions quite rapidly.

[Figure 4 about here]

The remaining experiments are concerned with the performance of bootstrap tests. In all cases, we used 399 bootstrap samples. This is the smallest value that we would recommend in most cases; see Davidson and MacKinnon (1997). The first question we investigate is whether approximate bootstrap procedures based on small values of m will actually yield estimated approximate bootstrap P values $\check{p}^*$ that are very close to the estimated bootstrap P values $\hat{p}^*$, as the theory suggests. Three different iterative procedures were used: Newton's method using the $(\boldsymbol{\gamma}, \delta)$ parametrization, a quasi-Newton method based on the OPG regression, and one based on the artificial regression of Orme (1995). We performed a number of experiments, each with 1000 replications, and calculated several measures of how close $\check{p}^*$ was to $\hat{p}^*$. Results for some of the experiments are reported in Table 1; results for the other experiments were similar. The table shows the average absolute difference between $\check{p}^*$ and $\hat{p}^*$. We also calculated the correlation between $\check{p}^*$ and $\hat{p}^*$ and the maximum absolute difference between them; all three measures always gave very similar results. When the average absolute difference is less than about 0.001, the maximum absolute difference is usually less than 0.01, and the squared correlation is usually greater than 0.9999. Thus we believe that

most investigators would regard an average absolute difference of less than 0.001 as negligibly small.

[Table 1 about here]

Several results are evident from Table 1. First of all, the OPG regression works dreadfully for the LR test. The approximate bootstrap P values are often far from the true ones, and they are sometimes farther away after two steps than after one step. The OPG regression also works poorly for the other tests, of course. The reason is simply that the OPG regression is apparently a very poor way to estimate tobit models. In contrast, the Orme regression works quite well, and, as the theory of Section 2 predicts, Newton's method works even better. As the theory of Section 3 predicts, m steps of Newton's method always perform better for the LR test than for the LM and Wald tests, and $m = 2$ appears to be adequate in all cases for the LR test. Perhaps surprisingly, $m = 2$ is often adequate for the other two tests as well.

Unfortunately, although it works very well, the approximate bootstrap procedure is not quite as useful in this case as these results may suggest. The reason is that Newton's method converges very rapidly, just as (1) says it should. For the middle case in Table 1, the average numbers of steps needed to obtain the restricted and unrestricted estimates when $n = 50$ are only 3.87 and 4.06, respectively. These numbers decline to 3.67 and 3.81 for $n = 100$ and to 3.42 and 3.40 for $n = 200$. The restricted estimates are used as starting values for the unrestricted estimation, and this undoubtedly makes the number of steps for the latter lower than it would otherwise be. Thus, although using the approximate bootstrap with $m = 2$ does save a considerable amount of CPU time, the savings are not really dramatic.

[Figure 5 about here]

The technique of this paper would not be of interest if bootstrap tests did not work well. For the tobit model, they seem to work very well indeed. Figure 5 shows P value discrepancy plots for all five bootstrap tests for six of the roughly twenty experiments that we ran. The level of the test is plotted on the horizontal axis, and the difference between the actual rejection frequency and the level is plotted on the vertical axis. For a test that performed perfectly, the plot would be a horizontal line at zero, plus a little experimental noise; see Davidson and MacKinnon (1998). Because all the tests performed so well, it was necessary to use a great many replications in order to distinguish between genuine discrepancies and experimental noise. All of the experiments therefore used 100,000 replications, each with 399 bootstrap samples.

It is clear from Figure 5 that the bootstrap tests work well, but not quite perfectly. For the case of $k_1 = 2$ and $k_2 = 8$, in panels A and B, all the tests work very well, although their performance does deteriorate a bit when σ is increased from 0.1 to 5.0. For the case of $k_1 = 5$ and $k_2 = 8$, in the remaining four panels, the deterioration as σ increases is much more marked. All the tests underreject noticeably when $\sigma = 5.0$, while the OPG LM test overrejects slightly when $\sigma = 0.1$. The worst case is shown in panel E. In contrast, panel F, which has the same parameter values but with $n = 100$, shows that all the tests except the OPG LM test work essentially perfectly, and even the OPG LM test improves dramatically.

We also ran experiments for several other cases. In the $(2, 5)$ case, bootstrap tests always worked better than in the $(2, 8)$ case of panels A and B. In the $(5, 5)$ and $(8, 2)$ cases, they worked less well than in the $(2, 8)$ case, but generally better than in the $(5, 8)$ case of panels C through F. Based on the results in Table 1, we computed true bootstrap P values for experiments with $n = 50$ but used the approximate bootstrap with $m = 2$ for experiments with $n = 100$. There is nothing in panel F to suggest that this harmed the performance of the bootstrap in any way.

Although some size distortions are evident in Figure 5, it is important to remember that all the bootstrap tests always perform dramatically better than the corresponding asymptotic tests. For example, at the .05 level, the worst bootstrap test (OPG LM) rejects 3.91% of the time for the $(5, 8)$ case shown in panel E of Figure 5, while the corresponding asymptotic test rejects 25.75% of the time. Even the best asymptotic test, the ES LM test, rejects more than 12.5% of the time at the .05 level in some of our experiments (see Figure 3). In contrast, the worst performance we observed for the bootstrap version of this test was a rejection rate of 4.77% in the case $k_1 = 8, k_2 = 2, \sigma = 5.0$.

5. Tests of Common Factor Restrictions

In order to provide some evidence on the performance of approximate bootstrap tests in a time-series context, we study their application to tests of common factor restrictions in linear regression models. Consider the model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \sum_{j=1}^{p} \rho_j u_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma^2), \tag{20}$$

in which y_t is an observation on a dependent variable, $\boldsymbol{X}_t$ is a $1 \times k$ vector of observations on exogenous regressors, $\boldsymbol{\beta}$ is a $k-$vector of parameters, and u_t is an error term that follows an AR(p) process with coefficients ρ_j and innovations ε_t. If we drop the first p observations, (20) can be rewritten as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \sum_{j=1}^{p} \rho_j y_{t-j} - \sum_{j=1}^{p} \rho_j \boldsymbol{X}_{t-j}\boldsymbol{\beta} + \varepsilon_t. \tag{21}$$

It is easy to see that (21) is a special case of the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \sum_{j=1}^{p} \rho_j y_{t-j} + \sum_{j=1}^{p} \boldsymbol{X}_{t-j}\boldsymbol{\gamma}_j + \varepsilon_t, \tag{22}$$

where the $\boldsymbol{\gamma}_j$ are $k-$vectors of parameters. We can obtain (21) from (22) by imposing the common factor restrictions

$$\boldsymbol{\gamma}_j = -\rho_j\boldsymbol{\beta}, \quad j = 1, \ldots, p. \tag{23}$$

There appear to be pk restrictions. However, when $\boldsymbol{X}_t$ includes a constant term, trend terms, seasonal dummy variables, a lagged dependent variable, or more

than one lag of the same independent variable, the number of restrictions will be less than pk, because the unrestricted model (22) will not be identified without further restrictions. For an elementary exposition, and references, see Davidson and MacKinnon (1993, Section 10.9).

It is important to test the common factor restrictions (23), because many types of misspecification can give rise to the appearance of serial correlation. A natural way to do so is to estimate (21) by nonlinear least squares and (22) by ordinary least squares, in order to obtain the restricted and unrestricted sums of squared residuals, respectively, and then compute an ordinary F test. However, this F test will not be exact in finite samples, even if the ε_t are normally distributed, because (21) is nonlinear and because both it and (22) are dynamic models. It is therefore natural to bootstrap the F test. We do this by using the parametric bootstrap. The bootstrap samples are generated from (20), under the assumption that the errors are normally distributed, using the estimated parameters from NLS estimation of (21). The condition that the u_t be stationary is imposed when generating the bootstrap data. If the estimated ρ_j do not satisfy the stationarity conditions (a very rare event in our experiments), they are shrunk proportionately until they do satisfy them.

Instead of an F test, it would be possible to use an LR test based on the same two regressions or an LM test based on a Gauss-Newton regression for the unrestricted model evaluated at the restricted estimates. In the bootstrap context, there is no point considering either of these tests, however. It is well known that LR and F tests for slope coefficients in regression models are monotonically related; see Davidson and MacKinnon (1993, Section 13.4). In this case, because the Gauss-Newton regression for the LM test must have the same sum of squared residuals as regression (22), the LM test is also monotonically related to the F test. Therefore, even though these three tests may have very different finite-sample properties, bootstrapping them must yield identical bootstrap P values.

For a common factor test, only the restricted model (22) requires nonlinear estimation. We estimate it using the Gauss-Newton regression, a quasi-Newton method that is easy to implement and seems to work well for this type of model. When the iterative procedure is terminated after only a few steps, the restricted sum of squared residuals will be somewhat too large. As a consequence, the bootstrap test statistic τ_j^* will be too large, so that too many of the τ_j^* will exceed $\hat{\tau}$, and the estimated bootstrap P value will therefore tend to be too large. This means that an approximate bootstrap test will reject the null hypothesis less frequently than a genuine bootstrap test. Thus we can see how well the approximate bootstrap procedure works simply by comparing the rejection frequencies of the approximate and genuine bootstrap tests.

For our Monte Carlo experiments, $\boldsymbol{X}_t$ consisted of a constant term and 3 other regressors, which were generated from independent AR(1) processes with parameter 0.5. Preliminary work showed that the F test always worked quite well, but by no means perfectly, that the LR test rejected more often than the F test, frequently much more often, and that the LM test rejected less often than the F test, frequently much less often. The F test overrejected in some cases and underrejected in others. For the bootstrapping experiments, we selected two

cases, one for which the F test overrejected and one for which it underrejected. In case 1, $p = 1$, $\rho_1 = 0.9$, and there were 3 restrictions. In case 2, $p = 2$, $\rho_1 = -0.3$, $\rho_2 = 0.1$, and there were 6 restrictions. We did 11 experiments for each case, with sample sizes 20, 25, 30, ..., 70. Each of the experiments involved 399 bootstrap samples and 100,000 replications. These experiments required a great deal of CPU time, but the large number of replications was essential to obtain reasonably accurate results.

[Figure 6 about here]

[Figure 7 about here]

Figures 6 and 7 show rejection frequencies as a function of the sample size for cases 1 and 2, respectively. In Figure 6, we see that the F test always overrejects. In contrast, the bootstrap test overrejects slightly for the smallest sample sizes, but it seems to perform very well (allowing for experimental error) for $n \geq 40$. The approximate bootstrap test with $m = 2$ always underrejects noticeably, although its performance is quite acceptable for $n \geq 50$. The approximate bootstrap test with $m = 3$ underrejects much less severely, and it is essentially indistinguishable from the genuine bootstrap test for $n \geq 35$.

In Figure 7, we see that the F test always underrejects, as does the bootstrap test for small values of n. The approximate bootstrap tests never underreject as severely as the F test, however, and the one that uses $m = 3$ performs very well for $n \geq 40$. Although the approximate bootstrap test with $m = 2$ does underreject noticeably for all sample sizes, its performance is quite acceptable for $n \geq 50$.

[Figure 8 about here]

Because the estimation method used here is a quasi-Newton method, the gains from using the approximate bootstrap are greater than the ones observed for the tobit model. Figure 8 shows the numbers of steps needed to obtain restricted parameter estimates that are accurate to 4 or 5 digits, as a function of the sample size, for cases 1 and 2. As can be seen from the figure, the computational savings from using a small, fixed number of steps are substantial, especially in case 2.

6. Final remarks

We have shown that the cost of bootstrap testing for nonlinear models can be reduced by using a small number of steps of an iterative procedure based either on Newton's method or on a quasi-Newton method that uses an artificial regression, when computing the bootstrap test statistics. The theory implies that just 2 steps of the iterative procedure should be sufficient for likelihood ratio tests, while 2, 3, or 4 steps will be needed for Lagrange multiplier and Wald tests. For a given number of steps greater than 1, the approximate bootstrap P values that result should be more accurate when Newton's method is used than when a quasi-Newton method is used.

Of course, since these results are based on asymptotic theory, they may or may not provide a good guide in any actual case. It is always possible that the sample size may be too small for a given iterative procedure to work sufficiently well for the theory to apply. However, it is much more likely that such a procedure

will work well in the bootstrap context than in the context of real data, because the bootstrap data are generated from the model that is being estimated, and because the starting values, which are the parameters used to generate the data, are guaranteed to be distant from the bootstrap estimates by an amount that is $O_p(n^{-1/2})$ in the sample size n. Thus we believe that the techniques proposed in this paper will often be useful in practice. They appear to work well for the two examples that we studied, namely, tests of slope coefficients in tobit models and tests of common factor restrictions.

References

Amemiya, T., "Regression Analysis when the Dependent Variable is Truncated Normal," *Econometrica* 41 (1973), 997–1016.

Beran, R., "Prepivoting Test Statistics: a Bootstrap View of Asymptotic Refinements," *Journal of the American Statistical Association* 83 (1988), 687–697.

Dagenais, M. G. and J.-M. Dufour, "Invariance, Nonlinear Models and Asymptotic Tests," *Econometrica* 59 (1991), 1601–1615.

Davidson, R. and J. G. MacKinnon, "Specification Tests Based on Artificial Regressions," *Journal of the American Statistical Association* 85 (1990), 220–227.

Davidson, R. and J. G. MacKinnon, *Estimation and Inference in Econometrics* (New York: Oxford University Press, 1993).

Davidson, R. and J. G. MacKinnon, "The Size Distortion of Bootstrap Tests," GREQAM Document de Travail No. 96A15 and Queen's Institute for Economic Research Discussion Paper No. 936, 1996.

Davidson, R. and J. G. MacKinnon, "Bootstrap Tests: How Many Bootstraps?" Queen's Institute for Economic Research Discussion Paper No. 951, 1997.

Davidson, R. and J. G. MacKinnon, "Graphical Methods for Investigating the Size and Power of Hypothesis Tests," *The Manchester School* 66 (1998), forthcoming.

Gregory, A. W. and M. R. Veall, "On Formulating Wald Tests for Nonlinear Restrictions," *Econometrica* 53 (1985), 1465–1468.

Gregory, A. W. and M. R. Veall, "Formulating Wald Tests of the Restrictions Implied by the Rational Expectations Hypothesis," *Journal of Applied Econometrics* 2 (1987), 61–68.

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

Hansen, L. P., "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50 (1982), 1029–1054.

Horowitz, J. L., "Bootstrap-Based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61 (1994), 395–411.

Newey, W. K. and D. McFadden, "Large Sample Estimation and Hypothesis Testing," in R. F. Engle and D. McFadden, eds., *Handbook of Econometrics*, Vol. IV (Amsterdam: North-Holland, 1994), 2111–2245.

Neyman, J., "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in U. Grenander, ed., *Probability and Statistics* (New York: John Wiley, 1959), 213–234.

Olsen, R. J., "Note on the Uniqueness of the Maximum Likelihood Estimator of the Tobit Model," *Econometrica* 46 (1978), 1211–1215.

Orme, C., "On the Use of Artificial Regressions in Certain Microeconometric Models," *Econometric Theory* 11 (1995), 290–305.

Phillips, P. C. B. and J. Y. Park, "On the formulation of Wald tests of nonlinear restrictions," *Econometrica* 56 (1988), 1065–1083.

Quandt, R. E., "Computational Problems and Methods," in Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*, Vol. I, (Amsterdam: North-Holland, 1983), 699–764.

Robinson, P. M., "The Stochastic Difference Between Econometric Statistics," *Econometrica* 56 (1988), 531–548.

Smith, R. J., "Alternative Asymptotically Optimal Tests and Their Application to Dynamic Specification," *Review of Economic Studies*, 54 (1987), 665–680.

Tobin, J., "Estimation of Relationships for Limited Dependent Variables," *Econometrica* 26 (1958), 24–36.

TABLE 1

AVERAGE ABSOLUTE DIFFERENCES BETWEEN $\check{p}^*$ AND $\hat{p}^*$

n	m	LR (OPG)	LR (Orme)	LR (N)	LM (N)	W_γ (N)
$k_1 = 5, k_2 = 8, \sigma = 0.1$						
50	1	0.1678	0.0617	0.0023	0.0149	0.0252
50	2	0.2735	0.0011	0.0000	0.0004	0.0004
100	1	0.1527	0.0215	0.0008	0.0055	0.0060
100	2	0.1811	0.0001	0.0000	0.0000	0.0000
200	1	0.0866	0.0098	0.0003	0.0023	0.0022
200	2	0.0546	0.0000	0.0000	0.0000	0.0000
$k_1 = 5, k_2 = 8, \sigma = 1.0$						
50	1	0.1699	0.0512	0.0044	0.0159	0.0487
50	2	0.2744	0.0014	0.0000	0.0005	0.0014
100	1	0.1317	0.0182	0.0013	0.0058	0.0163
100	2	0.1201	0.0002	0.0000	0.0001	0.0001
200	1	0.0580	0.0082	0.0005	0.0027	0.0068
200	2	0.0368	0.0001	0.0000	0.0000	0.0000
$k_1 = 5, k_2 = 8, \sigma = 5.0$						
50	1	0.1606	0.0403	0.0047	0.0125	0.0604
50	2	0.2365	0.0015	0.0001	0.0005	0.0023
100	1	0.0846	0.0146	0.0013	0.0045	0.0219
100	2	0.0575	0.0002	0.0000	0.0001	0.0002
200	1	0.0272	0.0064	0.0006	0.0019	0.0093
200	2	0.0164	0.0001	0.0000	0.0000	0.0000

Note: In columns 3 through 7, the heading indicates which test is being bootstrapped and, in parentheses, the method used for iteration: "OPG" means the OPG regression, "Orme" means the artificial regression proposed by Orme (1995), and "N" means Newton's Method.

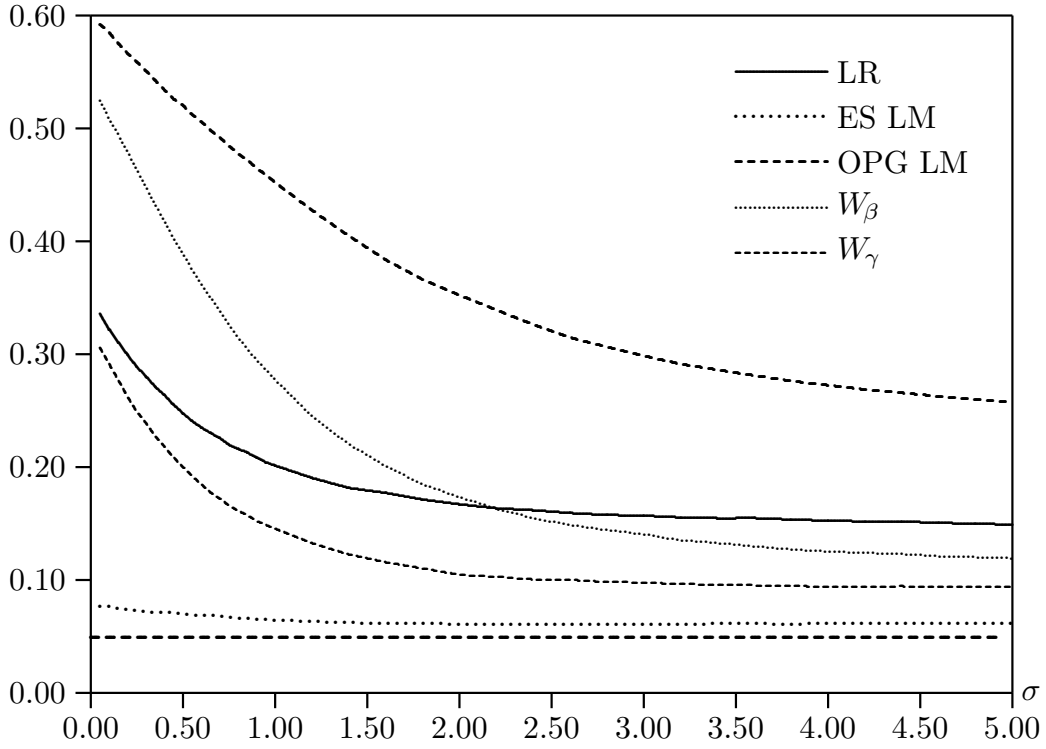

Figure 1. Rejection frequencies for all tests at .05 level, $n = 50$**,** $k_1 = 5$**,** $k_2 = 8$

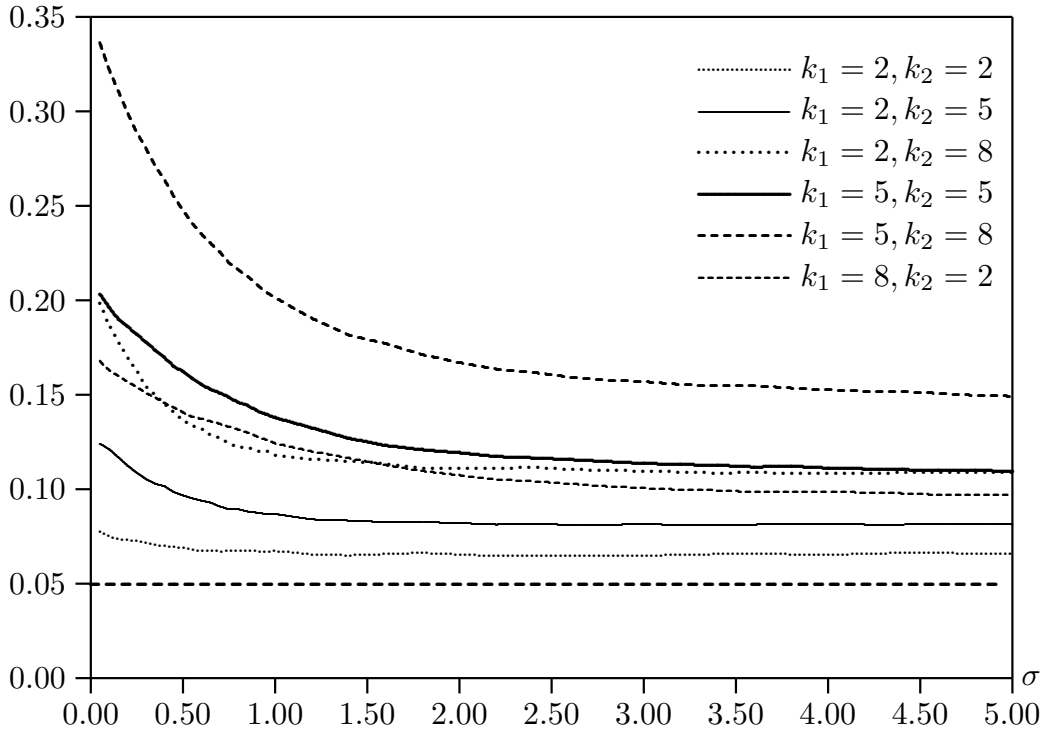

Figure 2. Rejection frequencies for LR tests at .05 level, $n = 50$

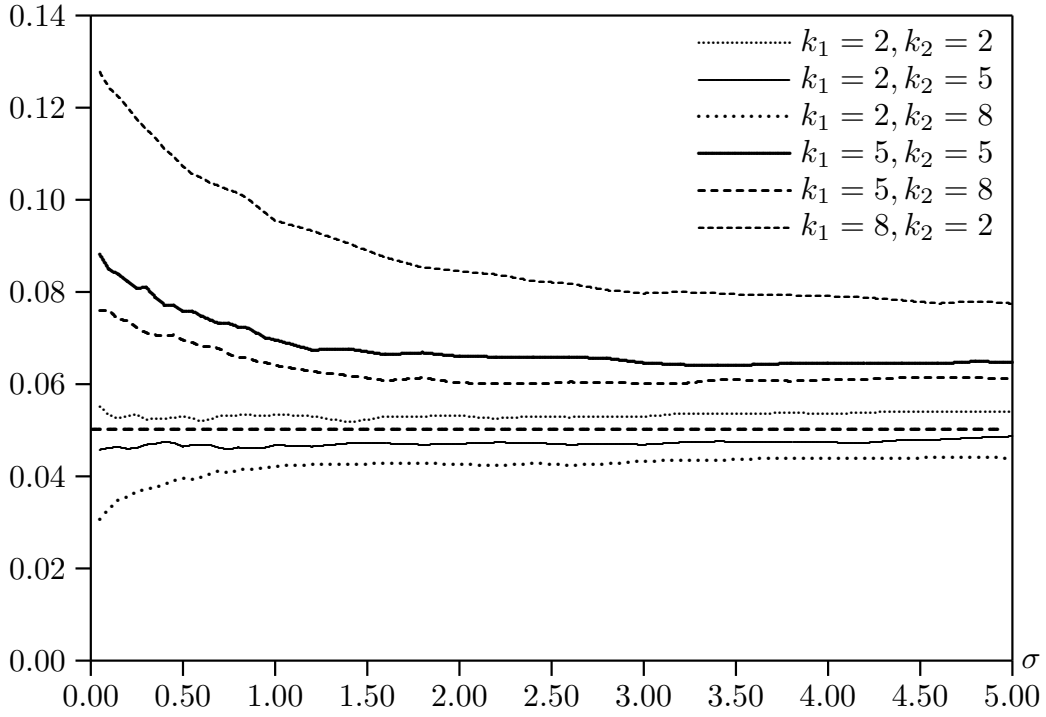

Figure 3. Rejection frequencies for ES LM tests at .05 level, $n = 50$

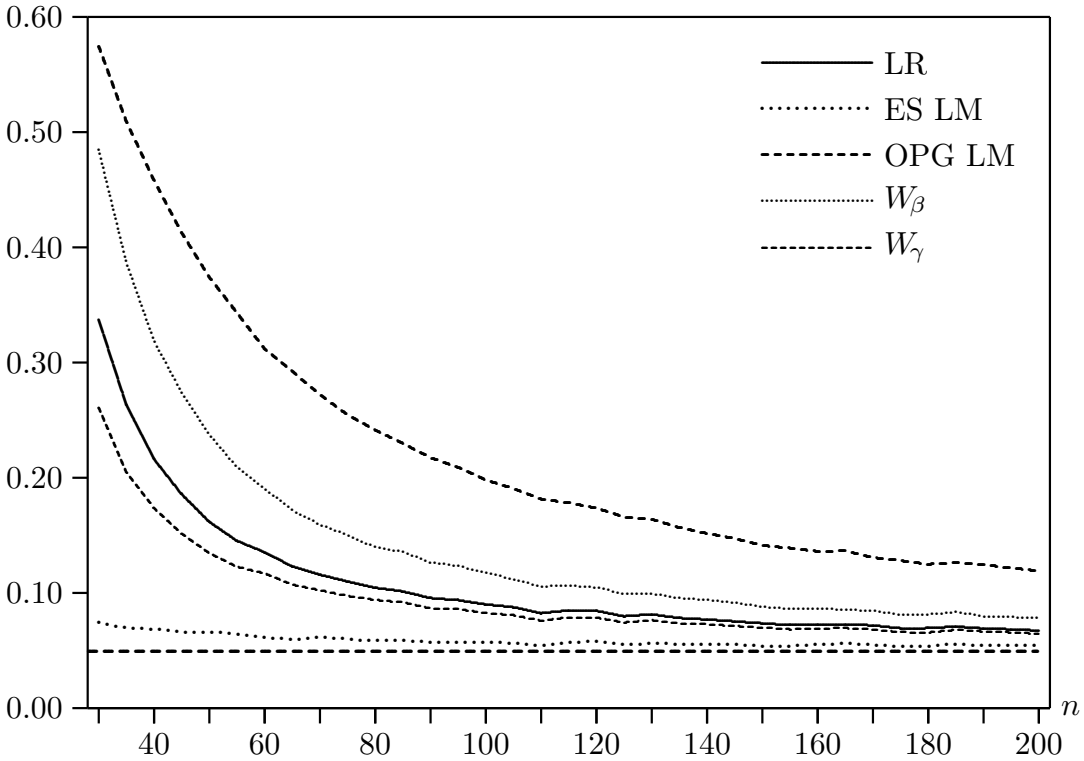

Figure 4. Rejections at .05 level as a function of sample size

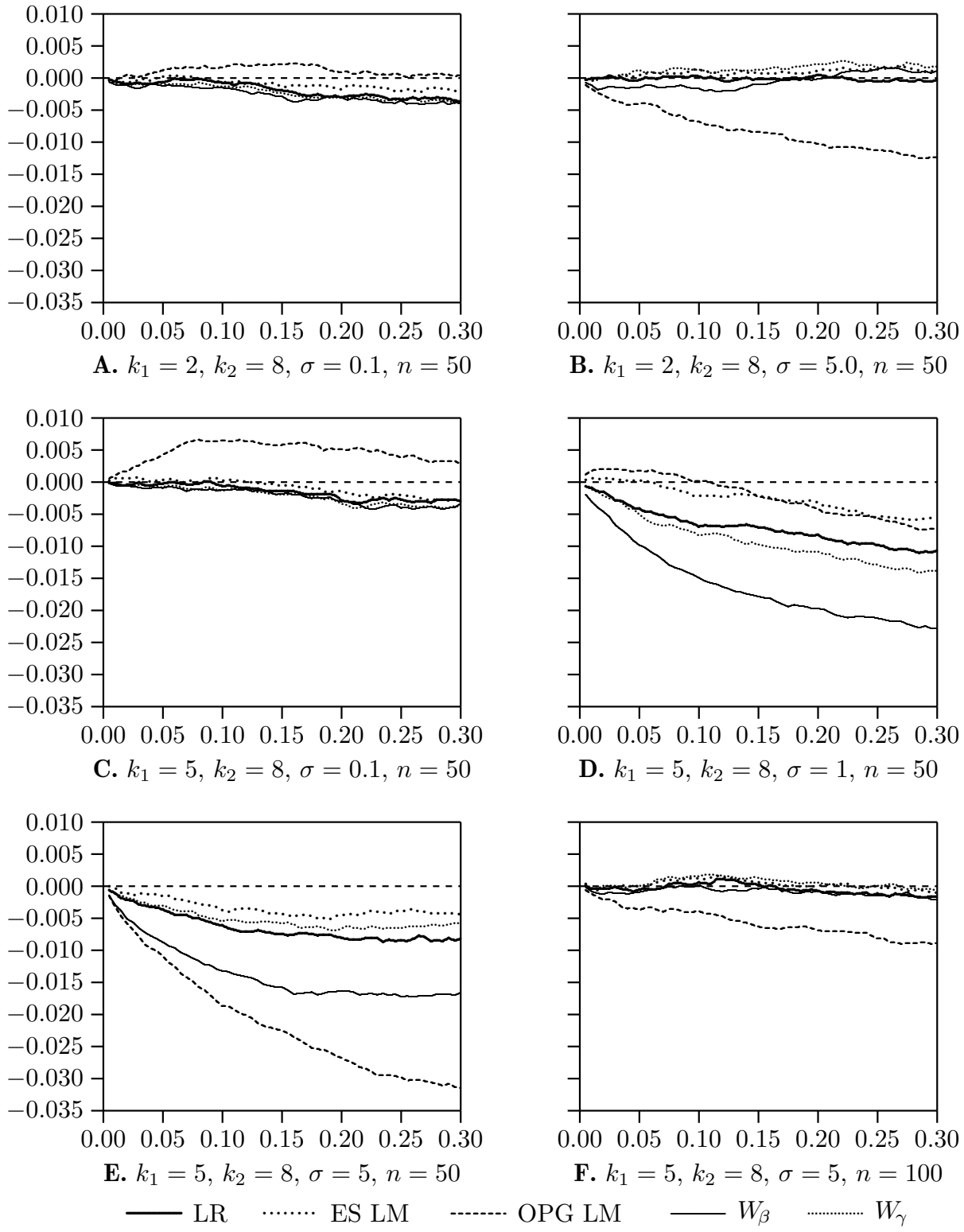

Figure 5. P **value discrepancy plots for bootstrap tests**

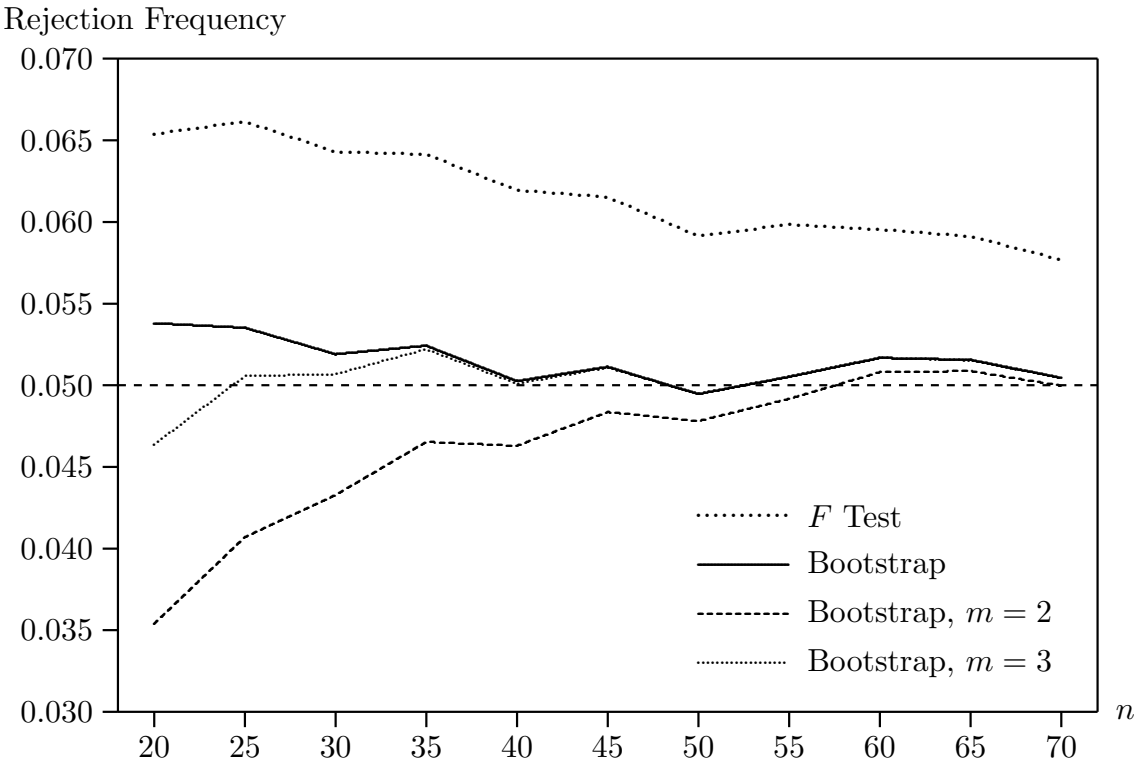

Figure 6. Rejection frequencies for common factor tests at .05 level, case 1

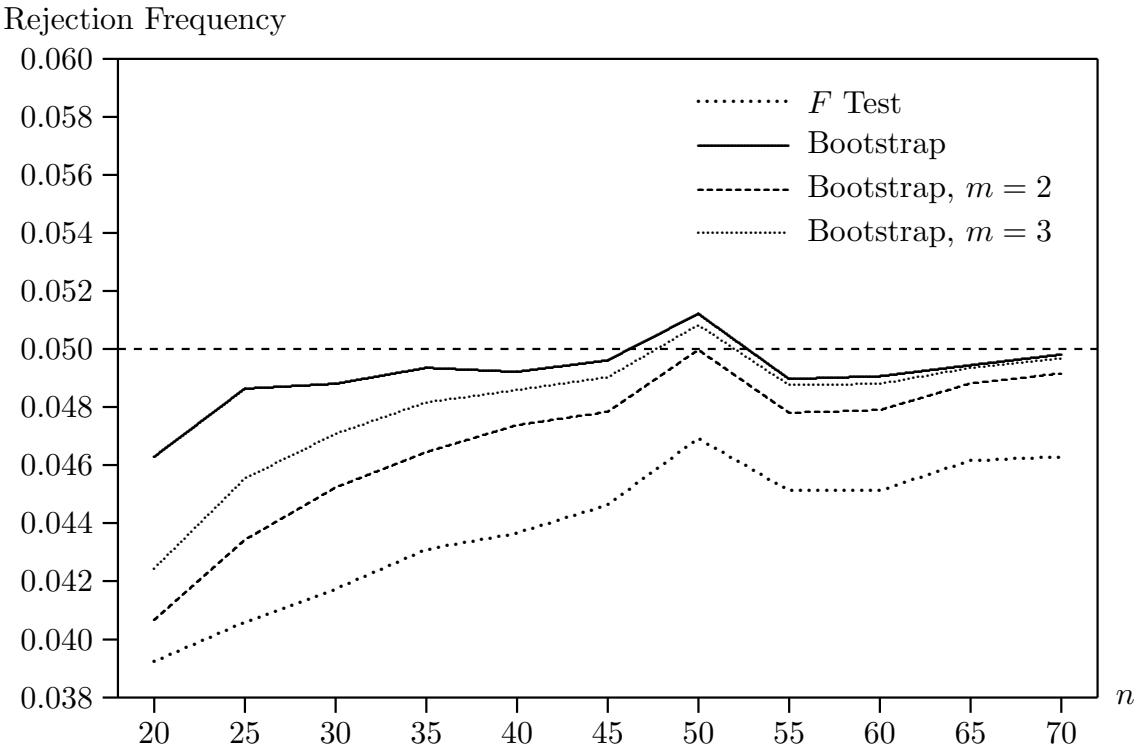

Figure 7. Rejection frequencies for common factor tests at .05 level, case 2

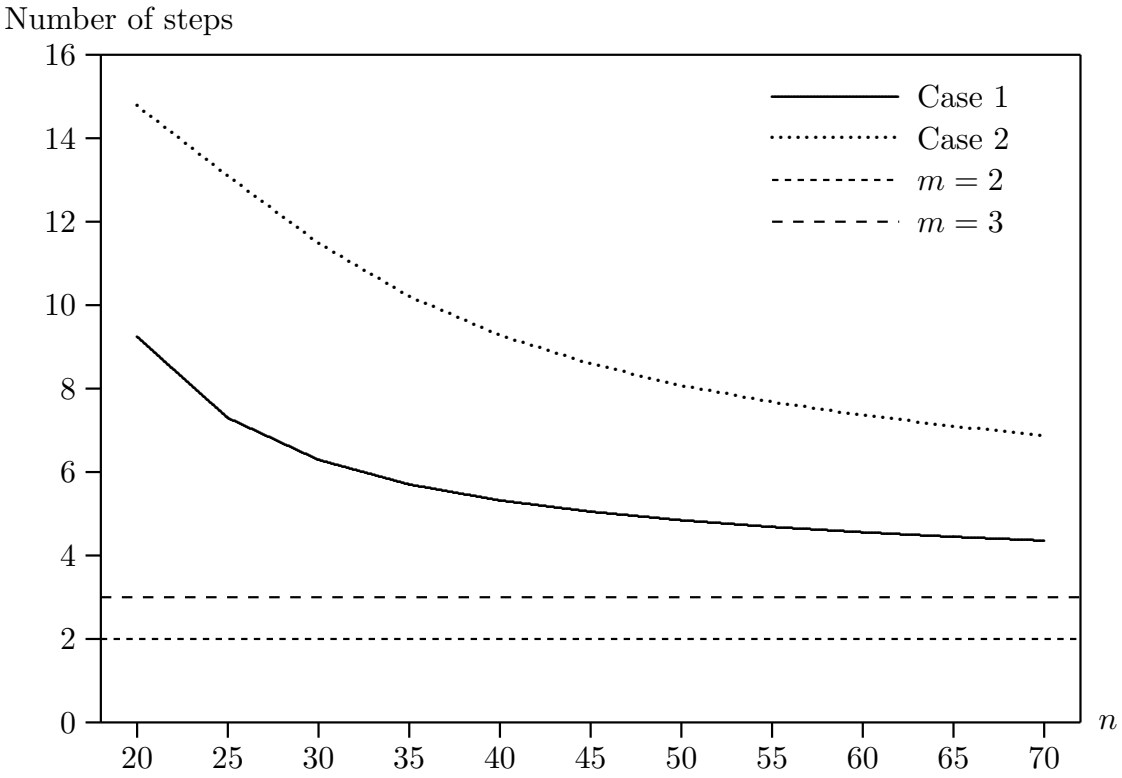

Figure 8. Number of steps for restricted estimation