Queen's Economics Department Working Paper No. 944

# Bootstrap Testing in Nonlinear Models

Russell Davidson
GREQAM, Queen's University

James G. MacKinnon
Queen's University

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

01-1997

# 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

When a model is nonlinear, bootstrap testing can be expensive because of the need to perform at least one nonlinear estimation for every bootstrap sample. We show that it may be possible to reduce computational costs by performing only a fixed, small number of artificial regressions, or Newton steps, for each bootstrap sample. The number of iterations needed is smaller for likelihood ratio tests than for other types of classical tests. The suggested procedures are applied to tests of slope coefficients in the tobit model, where asymptotic procedures often work surprisingly poorly. In contrast, bootstrap tests work remarkably well, and very few iterations are needed to compute them.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada.

January, 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 size of 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 (1996a). 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. Since nonlinear estimation can often be costly, it is of interest to see whether approximate methods can succeed in achieving the accuracy of bootstrap inference without the need for repeated nonlinear estimation. One natural approach is to use artificial linear regressions, which serve as local linearizations of many nonlinear models. Various cases are treated in Davidson and MacKinnon (1984a, 1984b) and Orme (1995). In Davidson and MacKinnon (1990), we laid out the general theory of artificial regressions. Only models estimated by maximum likelihood are treated in that paper, but similar methods can be applied to other classes of models, including models estimated by GMM. When we refer to artificial regressions, we have a wide variety of procedures in mind. In particular, since an artificial regression is really just a procedure for taking approximate Newton steps, genuine Newton steps can also be used.

In this paper, we show that the cost of bootstrapping can often be reduced by using artificial regressions to compute approximations to test statistics, the computation of which would usually require nonlinear estimation. Since the bootstrap is normally accurate only up to some order described by a negative power of the sample size, it is enough to use the artificial regression 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 by $mB$ OLS estimations of artificial regressions, where $m$ is a small integer. In most cases, one nonlinear estimation plus $mB$ linear ones will be less costly than $B+1$ nonlinear estimations.

In the next section, we review the theory of artificial regressions in the context of ML models, and we show that a finite, usually small, number of iterations of such a regression can yield approximations accurate to the same order as the bootstrap. Then, in Section 3, we describe in detail how approximate bootstrap tests may be implemented by use of artificial regressions, for all of the classical testing procedures: Lagrange Multiplier, Likelihood Ratio, $C(\alpha)$, and Wald. Finally, in Section 4, we illustrate the use of these tests in the context of the tobit model. We show that asymptotic tests often work poorly for this model, but bootstrap tests work remarkably well, and very few iterations are needed to compute them.

## 2. Estimation by Artificial Regression

The most convenient way to linearize econometric models is often to use some sort of artificial regression. Following Davidson and MacKinnon (1990), we begin by reviewing artificial regressions for models estimated by maximum likelihood. Consider a fully specified parametrized model characterized by its loglikelihood function, which for a sample of size $n$ can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ is a $k-$vector of model parameters. Any artificial regression associated with this model always involves two things: a regressand, say $\boldsymbol{r}(\boldsymbol{\theta})$, which will be a column vector, and a matrix of regressors, say $\boldsymbol{R}(\boldsymbol{\theta})$, which will be 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 $n$. The artificial regression may be written as

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \text{residuals}. \tag{2}$$

"Residuals" is used here as a neutral term to avoid any implication that (2) is a statistical model. Regression (2) can be evaluated at any point $\boldsymbol{\theta} \in \Theta$, for a parameter space $\Theta \in \mathbb{R}^k$.

The theory of artificial regressions for maximum likelihood models is developed under the assumption that $\boldsymbol{r}(\boldsymbol{\theta})$ and $\boldsymbol{R}(\boldsymbol{\theta})$ have certain defining properties. These can be expressed as follows, where all probability limits are calculated for some data-generating process (DGP) characterized by the loglikelihood (1) for some set of parameters $\boldsymbol{\theta} \in \Theta$:

(i) under the DGP characterized by $\boldsymbol{\theta}$, $\operatorname{plim}\big(n^{-1}\boldsymbol{r}^\top(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\big)$ exists and is a finite, smooth, real-valued function of $\boldsymbol{\theta}$, the value of which is denoted by $\rho(\boldsymbol{\theta})$;

(ii) $\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta}) = \rho(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}(\boldsymbol{\theta})$ is the $k-$vector of scores associated with the model (1); and

(iii) if $\ddot{\boldsymbol{\theta}} - \boldsymbol{\theta} \to \mathbf{0}$, then $n^{-1}\boldsymbol{R}^\top(\ddot{\boldsymbol{\theta}})\boldsymbol{R}(\ddot{\boldsymbol{\theta}}) \to \rho(\boldsymbol{\theta})\mathcal{I}(\boldsymbol{\theta})$, where $\mathcal{I}(\boldsymbol{\theta})$ denotes the information matrix associated with (1).

Artificial regressions that satisfy properties (i) through (iii) are shown in Davidson and MacKinnon (1990) to possess a number of useful properties.

We now consider the use of artificial regressions in an iterative procedure for obtaining the MLE, $\hat{\boldsymbol{\theta}}$. By defining property (ii) above, the first-order conditions for the MLE are:

$$\boldsymbol{R}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{r}(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{3}$$

We shall assume that the estimator $\hat{\boldsymbol{\theta}}$ is root-$n$ consistent, so that, when the DGP is characterized by a true parameter vector $\boldsymbol{\theta}_0$,

$$n^{1/2}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) = O(1).$$

Suppose we have some other parameter vector, say $\boldsymbol{\theta}_{(0)}$, which is also such that

$$n^{1/2}\big(\boldsymbol{\theta}_{(0)} - \boldsymbol{\theta}_0\big) = O(1). \tag{4}$$

It follows that

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

The vector $\boldsymbol{\theta}_{(0)}$ may be random, that is, another root-$n$ consistent estimator, or it may be deterministic, for instance, a vector of parameters specified by some null hypothesis. It may even be a mixture of the two, with some components random and others nonrandom. In the bootstrap context, $\boldsymbol{\theta}_{(0)}$ will generally be the parameter vector that is used to generate the bootstrap data; see Section 3.

The one-step estimator defined by the artificial regression (2) with $\boldsymbol{\theta}_{(0)}$ as starting point is

$$\boldsymbol{\theta}_{(1)} \equiv \boldsymbol{\theta}_{(0)} + \boldsymbol{b}_{(0)}, \text{ with } \boldsymbol{b}_{(0)} = \big(\boldsymbol{R}_{(0)}^\top \boldsymbol{R}_{(0)}\big)^{-1} \boldsymbol{R}_{(0)}^\top \boldsymbol{r}_{(0)}, \tag{5}$$

where $\boldsymbol{R}_{(0)}$ and $\boldsymbol{r}_{(0)}$ denote $\boldsymbol{R}(\boldsymbol{\theta}_{(0)})$ and $\boldsymbol{r}(\boldsymbol{\theta}_{(0)})$, respectively. The structure of an artificial regression ensures that $n^{-1}\boldsymbol{R}^\top(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\theta}) = O(1)$ for all $\boldsymbol{\theta} \in \Theta$, and so, under weak smoothness conditions, a Taylor expansion yields

$$n^{-1}\boldsymbol{R}_{(0)}^\top \boldsymbol{R}_{(0)} = n^{-1}\hat{\boldsymbol{R}}^\top \hat{\boldsymbol{R}} + O(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}), \tag{6}$$

where $\hat{\boldsymbol{R}} \equiv \boldsymbol{R}(\hat{\boldsymbol{\theta}})$. In order to deal with the other factor in $\boldsymbol{b}_{(0)}$, observe that, by defining property (ii) of an artificial regression and Taylor expansion,

$$\begin{aligned} n^{-1/2}\boldsymbol{R}_{(0)}^\top \boldsymbol{r}_{(0)} &= n^{-1/2}\rho_{(0)}\boldsymbol{g}_{(0)} \\ &= n^{-1/2}\big(\hat{\rho} + O(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}})\big)\Big(\boldsymbol{g}(\hat{\boldsymbol{\theta}}) + \boldsymbol{H}(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}})\Big). \end{aligned} \tag{7}$$

Here $\rho_{(0)} \equiv \rho(\boldsymbol{\theta}_{(0)})$, $\boldsymbol{g}_{(0)} \equiv \boldsymbol{g}(\boldsymbol{\theta}_{(0)})$, $\boldsymbol{H}(\boldsymbol{\theta})$ denotes the Hessian of the loglikelihood function (1), and $\bar{\boldsymbol{\theta}}$ is such that $\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\| \leq \|\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}\|$. Of course, by (3), $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \boldsymbol{0}$. Moreover, by the weak smoothness already assumed,

$$n^{-1}\boldsymbol{H}(\bar{\boldsymbol{\theta}}) = n^{-1}\boldsymbol{H}(\hat{\boldsymbol{\theta}}) + O(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}),$$

and, by defining property (iii) and the information matrix equality,

$$n^{-1}\hat{\rho}\boldsymbol{H}(\hat{\boldsymbol{\theta}}) = -n^{-1}\hat{\boldsymbol{R}}^\top \hat{\boldsymbol{R}} + O(n^{-1/2}).$$

We therefore find that

$$\begin{aligned} n^{-1/2}\boldsymbol{R}_{(0)}^\top \boldsymbol{r}_{(0)} &= -\big(\hat{\rho} + O(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}})\big)\Big(n^{-1}\hat{\rho}^{-1}\hat{\boldsymbol{R}}^\top \hat{\boldsymbol{R}} \; + O(n^{-1/2})\Big) n^{1/2}\big(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}\big) \\ &= n^{-1}\hat{\boldsymbol{R}}^\top \hat{\boldsymbol{R}}\, n^{1/2}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{(0)}\big) + O(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}). \end{aligned} \tag{8}$$

Thus, by (6) and (8),

$$\begin{aligned} n^{1/2}\boldsymbol{b}_{(0)} &= \big(n^{-1}\boldsymbol{R}_{(0)}^\top \boldsymbol{R}_{(0)}\big)^{-1} n^{-1/2}\boldsymbol{R}_{(0)}^\top \boldsymbol{r}_{(0)} \\ &= \Big(n^{-1}\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{R}} + O(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}})\Big)^{-1}\Big(n^{-1}\hat{\boldsymbol{R}}^\top\hat{\boldsymbol{R}}\, n^{1/2}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{(0)}\big) + O(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}})\Big) \\ &= n^{1/2}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{(0)}\big) + O\big(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}\big). \end{aligned} \tag{9}$$

Therefore, by (9) and the definition (5),

$$\boldsymbol{\theta}_{(1)} - \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{(0)} + \boldsymbol{b}_{(0)} - \hat{\boldsymbol{\theta}} = n^{-1/2}O\big(\boldsymbol{\theta}_{(0)} - \hat{\boldsymbol{\theta}}\big). \tag{10}$$

By assumption (4), the right-most expression here is $O(n^{-1})$.

Now consider what happens when the artificial regression is iterated. The result (10) is expressed in a form which makes it apparent that iteration leads to progressive refinements of the estimator. We have seen that, under assumption (4), the first-round estimator $\boldsymbol{\theta}_{(1)}$ has the property that $\boldsymbol{\theta}_{(1)} - \hat{\boldsymbol{\theta}} = O(n^{-1})$. If now $\boldsymbol{\theta}_{(1)}$ is used as the starting point of a further iteration, then, in obvious notation, the second-round estimator is

$$\boldsymbol{\theta}_{(2)} \equiv \boldsymbol{\theta}_{(1)} + \boldsymbol{b}_{(1)}, \text{ with } \boldsymbol{b}_{(1)} = \big(\boldsymbol{R}_{(1)}^\top \boldsymbol{R}_{(1)}\big)^{-1}\boldsymbol{R}_{(1)}^\top \boldsymbol{r}_{(1)}.$$

Replacing $\boldsymbol{\theta}_{(0)}$ and $\boldsymbol{\theta}_{(1)}$ in equation (10) by $\boldsymbol{\theta}_{(1)}$ and $\boldsymbol{\theta}_{(2)}$, respectively, then shows that $\boldsymbol{\theta}_{(2)} - \hat{\boldsymbol{\theta}} = O(n^{-3/2})$. Clearly, after $i$ iterations, we will have

$$\boldsymbol{\theta}_{(i)} - \hat{\boldsymbol{\theta}} = O(n^{-(i+1)/2}), \tag{11}$$

so that we gain a refinement of order $n^{-1/2}$ with each iteration. If an Edgeworth expansion exists for $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, then this expansion and that for $n^{1/2}(\boldsymbol{\theta}_{(i)} - \boldsymbol{\theta}_0)$ must coincide through order $n^{-i/2}$.

In some cases, it may be easier, or more effective, to use Newton's method than to run an artificial regression. If $\boldsymbol{g}_{(i-1)}$ and $\boldsymbol{H}_{(i-1)}$ denote the gradient and the Hessian matrix evaluated at $\boldsymbol{\theta}_{(i-1)}$, we simply compute

$$b_{(i)} = -\boldsymbol{H}_{(i-1)}^{-1}\boldsymbol{g}_{(i-1)}$$

directly instead of using an artificial regression to compute it. The result (11) obviously continues to apply when $b_{(i)}$ is defined in this way.

The result (11) is the key result of this paper. However, it must be interpreted with caution, because, like all results based on asymptotic expansions, it holds only for $n$ sufficiently large. When $n$ is not large enough, $\boldsymbol{\theta}_{(i)}$ may be much further away from $\hat{\boldsymbol{\theta}}$ than (11) suggests; indeed, it may even be further away from $\hat{\boldsymbol{\theta}}$ than $\boldsymbol{\theta}_{(0)}$. Nevertheless, this result offers the promise that $\boldsymbol{\theta}_{(i)}$ may converge very rapidly to $\hat{\boldsymbol{\theta}}$ in many circumstances.

## 3. Bootstrapping by Artificial Regression

Consider again the model characterized by the loglikelihood function (1). Let the parameter vector $\boldsymbol{\theta}$ 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, we may use a likelihood ratio (LR) test, a Lagrange multiplier (LM) test, a $C(\alpha)$ test, or a Wald test, all of which are asymptotically equivalent under the null hypothesis and under DGPs local to the null. There are many more than four possibilities, since all but the LR test have numerous variants, distinguished by the use of different estimators of the information matrix and, perhaps, by different parametrizations of the model. All of these classical tests may be bootstrapped in order to improve their finite-sample properties.

There are several procedures that may be used for bootstrapping test statistics; see Horowitz (1994) for an alternative one. We recommend using the bootstrap to compute a $P$ 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 ML estimates $\tilde{\boldsymbol{\theta}} \equiv [\tilde{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$ that satisfy the null hypothesis.
2. Using a DGP with parameter vector $\tilde{\boldsymbol{\theta}}$, generate $B$ bootstrap samples, each of size $n$. Use each bootstrap sample to compute a bootstrap test statistic, say $\tau_j^*$, for $j = 1, \ldots, B$.
3. Calculate the estimated bootstrap $P$ value $\hat{p}^*$ as the proportion of bootstrap samples for which $\tau_j^*$ exceeds $\hat{\tau}$. If a formal test at level $\alpha$ is desired, reject the null hypothesis whenever $\hat{p}^* < \alpha$.

This bootstrap procedure is remarkably simple, and it often works remarkably well; see Davidson and MacKinnon (1996a, 1996b) and Section 4 below for evidence on this point. Notice that step 2 uses a parametric bootstrap, in which the bootstrap DGP is the parametric model characterized by the vector $\tilde{\boldsymbol{\theta}}$. A parametric bootstrap is usually appropriate if we are using ML estimation, but it may often be inappropriate if we are using NLS or GMM estimation. Step 2 can often be modified so that it involves less stringent distributional assumptions, but the details will vary from case to case, and the bootstrap DGP will generally depend on $\tilde{\boldsymbol{\theta}}$ to some extent.

Bootstrap tests perform best when $B$ is infinite. In practice, when $B$ is finite, the bootstrap will not perform quite as well, for two reasons. First, the estimated bootstrap $P$ value $\hat{p}^*$ will not equal the true bootstrap $P$ value $p^*$; it will, in fact, generally provide a biased estimate. Second, there will be some loss of power. Both these issues are discussed in Davidson and MacKinnon (1996b). 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 test size $\alpha$ that may be of interest, and so that it is not too small. $B = 399$ is the smallest value that we would recommend in most cases, and larger numbers such as $B = 999$ or $B = 1999$ may be preferable if computing cost is not a serious issue.

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 $\tau_j^*$ computed by means of artificial regressions. We discuss LM, LR, $C(\alpha)$, and Wald tests in that order.

**Bootstrapping LM tests**

It is logical to start with the LM test, in part because it is relatively simple to deal with. Moreover, since LM tests are often computed by means of artificial regressions, it seems particularly natural to use an artificial regression to implement the bootstrap for such tests. In order to compute an LM test in this way, we must use an artificial regression that corresponds to the full, unrestricted model. It is often useful to partition the matrix $\boldsymbol{R}(\boldsymbol{\theta})$ that appears in (2) to correspond to the partition of $\boldsymbol{\theta}$ into a $k_1-$vector $\boldsymbol{\theta}_1$ and a $k_2-$vector $\boldsymbol{\theta}_2$. When this is done, (2) can be rewritten as

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}_1(\boldsymbol{\theta})\boldsymbol{b}_1 + \boldsymbol{R}_2(\boldsymbol{\theta})\boldsymbol{b}_2 + \text{residuals}. \tag{12}$$

The variables in (12) are then evaluated at the restricted estimates $\tilde{\boldsymbol{\theta}}$ defined above. It is well known that $n$ times the uncentered $R^2$ from regression (12) evaluated at $\tilde{\boldsymbol{\theta}}$ is an LM test statistic which is asymptotically distributed as $\chi^2(k_2)$ under the null hypothesis; see Davidson and MacKinnon (1990) for details. In quite a few cases, the total sum of squares will have a plim of unity. In such cases, the explained sum of squares (ESS) is also a valid test statistic, and it is probably a better one to use.

Bootstrapping the LM statistic is conceptually straightforward, but it involves estimating the model $B$ additional times under the null hypothesis. We propose to replace the nonlinear ML estimation by a predetermined, finite, usually small, number of iterations of the artificial regression, as discussed in the previous section, starting from the ML estimates given by the real data. The justification of such a scheme is as follows. The bootstrap test will not be exact, on account of the difference between the bootstrap DGP, say $\tilde{\mu}$, and the true unknown DGP, say $\mu_0$, that actually generated the data. As shown in Davidson and MacKinnon (1996a), the size of the bootstrap test at any given level $\alpha$ will be distorted by an amount that depends on the joint distribution of the test statistic computed from the artificial regression and the (random) level-$\alpha$ critical value for that statistic under the bootstrap DGP $\tilde{\mu}$. It is usually possible to determine an integer $m$ such that the true size of the bootstrap test at nominal level $\alpha$ differs from $\alpha$ by an amount that is $O(n^{-m/2})$. This being so, the same order of accuracy will be achieved even if there is an error that is $O(n^{-m/2})$ in the computation of the bootstrap test statistics.

The "true" value of the parameters for the bootstrap DGP is $\tilde{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\theta}}_1 \vdots \boldsymbol{0}]$. If we denote the ML estimates from a bootstrap sample by $\tilde{\boldsymbol{\theta}}_1^*$, then by construction we have that $\tilde{\boldsymbol{\theta}}_1^* - \tilde{\boldsymbol{\theta}}_1 = O(n^{-1/2})$. Thus $\tilde{\boldsymbol{\theta}}_1$ is a suitable starting point for a set of iterations of the artificial regression

$$\boldsymbol{r}(\boldsymbol{\theta}_1) = \boldsymbol{R}_1(\boldsymbol{\theta}_1)\boldsymbol{b}_1 + \text{residuals}. \tag{13}$$

This artificial regression corresponds to the model subject to the restrictions of the null hypothesis. The iterations are

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}^*_{1\,(0)} &= \tilde{\boldsymbol{\theta}}_1 \\
\tilde{\boldsymbol{\theta}}^*_{1\,(1)} &= \tilde{\boldsymbol{\theta}}^*_{1\,(0)} + \boldsymbol{b}_{1\,(0)}, \\
\tilde{\boldsymbol{\theta}}^*_{1\,(2)} &= \tilde{\boldsymbol{\theta}}^*_{1\,(1)} + \boldsymbol{b}_{1\,(1)}, \text{ and so on.}
\end{aligned} \tag{14}
$$

By the result (11), the iterated estimates $\tilde{\boldsymbol{\theta}}^*_{1\,(i)}$, $i = 0, 1, 2, \ldots$, satisfy

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

At this point, a little care is necessary. Consider the explained sum of squares from the artificial regression (12). Suppose that the variables in (12) are evaluated, not at $\tilde{\boldsymbol{\theta}}$, but somewhere nearby, say $\acute{\boldsymbol{\theta}}$. Then the explained sum of squares will change by an amount of order $n^{1/2}(\acute{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})$ rather than of order $(\acute{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})$. To see this, observe that the explained sum of squares can be written as

$$n^{-1/2}\acute{\boldsymbol{r}}^\top\acute{\boldsymbol{R}}\big(n^{-1}\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{R}}\big)^{-1}n^{-1/2}\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{r}}.$$

From (8), it can be seen that the difference between $n^{-1/2}\acute{\boldsymbol{R}}^\top\acute{\boldsymbol{r}}$ and $n^{-1/2}\tilde{\boldsymbol{R}}^\top\tilde{\boldsymbol{r}}$ is of order $n^{1/2}(\acute{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})$, from which the result follows. Thus, the ESS from regression (12) when (12) is evaluated at $\tilde{\boldsymbol{\theta}}^*_{1\,(i)}$ will differ from the ESS when (12) is evaluated at $\tilde{\boldsymbol{\theta}}^*_1$ by an amount that is of order $n^{-i/2}$. The same is true for the LM statistic, whether it is computed as ESS or as $nR^2$. This implies that $m$ iterations on a bootstrap sample, starting from $\tilde{\boldsymbol{\theta}}_1$, will yield approximate bootstrap LM statistics that differ from the actual one by amounts of order $n^{-m/2}$. Choosing $m = 3$ or, in some cases, $m = 4$ will cause the approximation error to be of the same order as the size distortion of the bootstrap test.

Let us now recapitulate our proposed procedure for bootstrapping an LM test without doing any nonlinear estimations for the bootstrap samples. Since it is simply a variant of the general procedure discussed above, we number the steps so that the relationship is clear.

1. Compute a vector of ML estimates $\tilde{\boldsymbol{\theta}} \equiv [\tilde{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$ that satisfy the null hypothesis. Then use regression (12) to compute the LM statistic $\hat{\tau}_{LM}$.

2a. Draw $B$ bootstrap samples of size $n$ from the DGP characterized by the loglikelihood function (1) evaluated at $\tilde{\boldsymbol{\theta}}$.

2b. For each bootstrap sample, say the $j^{\text{th}}$, set up the artificial regression (13), evaluating its variables initially at $\tilde{\boldsymbol{\theta}}^*_{1\,(0)} \equiv \tilde{\boldsymbol{\theta}}_1$. Then use the OLS estimates $\boldsymbol{b}_{1\,(0)}$ to construct $\tilde{\boldsymbol{\theta}}^*_{1\,(1)} = \tilde{\boldsymbol{\theta}}^*_{1\,(0)} + \boldsymbol{b}_{1\,(0)}$. Repeat this step a total of $m$ times in order to obtain the iterated estimates $\tilde{\boldsymbol{\theta}}^*_{1\,(m)}$, as in (14).

2c. Run regression (12) evaluated at $\tilde{\boldsymbol{\theta}}^*_{(m)} \equiv [\tilde{\boldsymbol{\theta}}^*_{1\,(m)} \vdots \mathbf{0}]$. Compute the bootstrap LM statistic, $\tau^*_{j\,(m)}$, as the $nR^2$, or possibly as the ESS, from this regression.

3. Calculate the estimated approximate bootstrap $P$ value $\check{p}^*$ as the proportion of bootstrap samples for which $\tau^*_{j\,(m)}$ exceeds $\hat{\tau}_{LM}$.

Only at steps 2b and 2c does this procedure differ from the general one described previously. Instead of using $\tau^*_j$ computed from regression (12) evaluated at $\tilde{\boldsymbol{\theta}}^*$ to estimate the bootstrap $P$ value, we use $\tau^*_{j\,(m)}$ computed from the same regression evaluated at $\tilde{\boldsymbol{\theta}}^*_{(m)}$.

**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 avoided. Estimation under the null is replaced by $m$ iterations of regression (13), exactly as in step 2b above, yielding the iterated estimates $\tilde{\boldsymbol{\theta}}^*_{(m)}$. Estimation under the alternative is replaced by $m$ iterations of regression (12), starting from $\tilde{\boldsymbol{\theta}}^*_{(m)}$, yielding the iterated estimates $\hat{\boldsymbol{\theta}}^*_{(m)}$. It would also be valid to commence these iterations at $\tilde{\boldsymbol{\theta}}$, but that would be throwing away information about the current bootstrap sample.

The true bootstrap LR statistic is $2\big(\ell(\hat{\boldsymbol{\theta}}^*) - \ell(\tilde{\boldsymbol{\theta}}^*)\big)$. If we replace $\hat{\boldsymbol{\theta}}^*$ and $\tilde{\boldsymbol{\theta}}^*$ by $\hat{\boldsymbol{\theta}}^*_{(m)}$ and $\tilde{\boldsymbol{\theta}}^*_{(m)}$, respectively, we are introducing errors of order $n^{-(m+1)/2}$, by the result (11). However, the approximate LR statistic differs from the true one by a quantity that is only of order $n^{-m}$. Therefore, $m = 2$ should be sufficient to ensure that the approximation error for the LR statistic is of the same or lower order as the size distortion of the bootstrap test itself.

This result can be seen, for both the unrestricted and restricted loglikelihoods, by considering the difference between a maximized loglikelihood $\ell(\hat{\boldsymbol{\theta}})$ and $\ell(\acute{\boldsymbol{\theta}})$, where $\acute{\boldsymbol{\theta}}$ is close to $\hat{\boldsymbol{\theta}}$. Since $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ by the first-order conditions for maximizing $\ell(\boldsymbol{\theta})$, a Taylor expansion gives

$$\ell(\hat{\boldsymbol{\theta}}) - \ell(\acute{\boldsymbol{\theta}}) = -\tfrac{1}{2}\big(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\big)^\top \boldsymbol{H}(\bar{\boldsymbol{\theta}})\big(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\big), \tag{15}$$

where, as in (7), $\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\| \leq \|\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\|$, and $\boldsymbol{H}$ is the Hessian of the loglikelihood. Since $\boldsymbol{H}$ is $O(n)$, it follows that, if $\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}$ is $O(n^{-(m+1)/2})$, then (15) is of order $n^{-m}$, as claimed. Since this is true for both the restricted and the unrestricted loglikelihood functions, it must be true for the LR statistic.

This result suggests that fewer iterations of the artificial regression will be needed in order to obtain any desired degree of approximation for the LR statistic than for the LM statistic. 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.

To summarize, here is the procedure for bootstrapping an LR test without doing any nonlinear estimations for the bootstrap samples.

1. Compute two vectors of ML estimates, $\tilde{\boldsymbol{\theta}}$ that satisfies the null hypothesis and $\hat{\boldsymbol{\theta}}$ that does not. Then compute the LR statistic $\hat{\tau}_{LR}$ as $2\big(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\big)$.

2a. Draw $B$ bootstrap samples of size $n$ from the DGP characterized by $\ell(\tilde{\boldsymbol{\theta}})$.

2b. For each bootstrap sample, run the artificial regression (13), evaluating its variables initially at $\tilde{\boldsymbol{\theta}}_1$, and then using the OLS estimates to construct new estimates according to (14). Repeat a total of $m$ times. This step yields $\tilde{\boldsymbol{\theta}}^*_{(m)}$.

2c. For each bootstrap sample, run the artificial regression (12), evaluating its variables initially at $\tilde{\boldsymbol{\theta}}^*_{(m)}$, and then using the OLS estimates to construct new estimates as in step 2b. Repeat a total of $m$ times. This step yields $\hat{\boldsymbol{\theta}}^*_{(m)}$.

2d. For each bootstrap sample, compute the bootstrap LR statistic, $\tau^*_{j\,(m)}$, as $2\big(\ell(\hat{\boldsymbol{\theta}}^*_{(m)}) - \ell(\tilde{\boldsymbol{\theta}}^*_{(m)})\big)$.

3. Calculate the estimated approximate bootstrap $P$ value $\check{p}^*$ as the proportion of bootstrap samples for which $\tau^*_{j\,(m)}$ exceeds $\hat{\tau}_{LR}$.

In the above summary, we have not specified the method of maximum likelihood estimation to be used with the real data. Clearly, this estimation can also be carried out by using artificial regressions. However, it is not advisable to specify in advance the number of iterations used to obtain the ML estimates, because there is no way in general to be sure of starting the iterative process sufficiently close to the unknown true parameters. It is therefore preferable to choose in advance some convergence criterion, and then iterate for as long as necessary.

The result that $m = 2$ is 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, and simulation evidence to be presented in Section 4 suggests that it does hold in at least one interesting case. This result implies that an approximate LR statistic can be computed using just four artificial regressions, instead of two nonlinear optimizations. In contrast, an approximate LM statistic requires either four or five artificial regressions, depending on whether $m = 3$ or $m = 4$ is sufficient. Thus, if 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 $\acute{\boldsymbol{\theta}}_1$. The artificial regression (13) is then evaluated and run at $\acute{\boldsymbol{\theta}}_1$ and the artificial regression (12) is evaluated and run at $\acute{\boldsymbol{\theta}} \equiv [\acute{\boldsymbol{\theta}}_1 \vdots \mathbf{0}]$. The test statistic is either the difference between the $nR^2$ from (12) and the $nR^2$ from (13) or, in many but not all cases, the difference between the ESS from (12) and the ESS from (13).

Observe that running (13) allows us to obtain one-step, asymptotically efficient, estimates of $\boldsymbol{\theta}_1$, by (5) applied to the null model. Because of their asymptotic efficiency, it is preferable to use these one-step estimates, rather than $\acute{\boldsymbol{\theta}}_1$, in order to

set up the bootstrap DGP. Bootstrapping $C(\alpha)$ tests requires no iteration of the artificial regressions. Otherwise, the bootstrap procedure is very similar to the one for LM tests:

1a. Obtain the root-$n$ consistent estimates $\acute{\boldsymbol{\theta}}_1$ of the parameters of the null model. Then compute the test statistic $\hat{\tau}_C$ by running the artificial regressions (13) and (12), evaluating the former at $\acute{\boldsymbol{\theta}}_1$ and the latter at $\acute{\boldsymbol{\theta}}$. As discussed above, there are two possible test statistics.

1b. Obtain one-step efficient estimates, $\grave{\boldsymbol{\theta}}_1$, of the null model parameters by adding the artificial parameter estimates from (13) to $\acute{\boldsymbol{\theta}}_1$.

2a. Draw $B$ bootstrap samples of size $n$ from the DGP characterized by $\ell(\grave{\boldsymbol{\theta}}_1 \vdots \mathbf{0})$.

2b. For each bootstrap sample, obtain root-$n$ consistent estimates $\acute{\boldsymbol{\theta}}_1^*$ by the same procedure as that used in the first step to obtain $\acute{\boldsymbol{\theta}}_1$.

2c. Set up (13) and (12), evaluating them at $\acute{\boldsymbol{\theta}}_1^*$ and $\acute{\boldsymbol{\theta}}^* \equiv [\acute{\boldsymbol{\theta}}_1^* \vdots \mathbf{0}]$, respectively. Compute the bootstrap test statistic, $\tau_j^*$ in the same way that $\hat{\tau}_C$ was computed.

3. Calculate the estimated bootstrap $P$ value $\hat{p}^*$ as the proportion of bootstrap samples for which $\tau_j^*$ exceeds $\hat{\tau}_C$.

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. For each bootstrap sample, it is necessary to run two artificial regressions, instead of $m+1$ (with $m = 3$ or $m = 4$, for the LM test) or $2m$ (with $m = 2$, for the LR test) that are needed to implement the approximate bootstrap procedures suggested earlier.

**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 ML 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 ML estimates for each bootstrap sample, and this can be done by iterating an artificial regression.

The number of iterations of an artificial regression 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

$$\hat{\tau}_W = (n^{1/2}\hat{\boldsymbol{\theta}}_2)^\top \hat{\boldsymbol{V}}^{-1} (n^{1/2}\hat{\boldsymbol{\theta}}_2), \tag{16}$$

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 $O(1)$. 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 (16) is approximated by $\acute{\tau}_W$, obtained by evaluating (16) at some $\acute{\boldsymbol{\theta}}$ close to $\hat{\boldsymbol{\theta}}$, the approximation error is clearly of order $n^{1/2} O(\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$. In particular, if $\acute{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}$ is $O(n^{-(m+1)/2})$, then $\acute{\tau}_W - \hat{\tau}_W$ is $O(n^{-m/2})$.

The order of approximation of a Wald test after $m$ iterations of an artificial regression is thus the same as for an LM test. Clearly, it is the same for a Wald-like $C(\alpha)$ test as well, since $C(\alpha)$ tests are based on quantities that take the form of either $nR^2$ or an explained sum of squares, just like LM tests. Theory suggests that 3 or 4 iterations of the artificial regression will be needed for LM and Wald tests, compared with just 2 for LR tests.

To summarize, here is the approximate bootstrap procedure that we are proposing for Wald and Wald-like tests:

1a. Compute a vector $\hat{\boldsymbol{\theta}} \equiv [\hat{\boldsymbol{\theta}}_1 \vdots \hat{\boldsymbol{\theta}}_2]$ of ML estimates of the unrestricted model, and use it to compute the test statistic $\hat{\tau}_W$. For a true Wald test, this is given by (16) with some suitable choice of $\hat{\boldsymbol{V}}$. For a Wald-like test, $\hat{\tau}_W$ is computed as a $C(\alpha)$ test, by use of the artificial regressions (12) and (13).

1b. Obtain root-$n$ consistent estimates, $\grave{\boldsymbol{\theta}}_1$, of the parameters of the restricted model. If possible, use the restricted ML estimates $\tilde{\boldsymbol{\theta}}_1$. If not, use one or more iterations of (13) starting from $\hat{\boldsymbol{\theta}}_1$.

2a. Draw $B$ bootstrap samples of size $n$ from the DGP characterized by $\ell(\grave{\boldsymbol{\theta}}_1 \vdots \mathbf{0})$.

2b. For each bootstrap sample, run the artificial regression (12), evaluating its variables initially at $\hat{\boldsymbol{\theta}}$, and then using the OLS estimates to construct new estimates. Repeat this step a total of $m$ times in order to obtain the iterated estimates $\hat{\boldsymbol{\theta}}^*_{(m)}$.

2c. Compute the bootstrap Wald or Wald-like statistic, $\tau^*_{j\,(m)}$, in precisely the same way as $\hat{\tau}_W$ was computed.

3. Calculate the estimated approximate bootstrap $P$ value $\check{p}^*$ as the proportion of bootstrap samples for which $\tau^*_{j\,(m)}$ exceeds $\hat{\tau}_W$.

## 4. Bootstrap Testing in the Tobit Model

In this section, we provide some simulation evidence on how well the procedures proposed in this paper perform in finite samples. All of the simulations deal with 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. All of the classical tests are readily available for this model, but little appears to be known about their finite-sample properties. Thus a secondary objective of this section is to shed light on these properties, for both asymptotic and bootstrap tests.

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

$$\sum_{y_t=0} \log\Big(\Phi\Big(-\tfrac{1}{\sigma}\big(\boldsymbol{X}_{1t}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}\boldsymbol{\beta}_2\big)\Big)\Big) + \sum_{y_t>0} \log\Big(\tfrac{1}{\sigma}\phi\Big(\tfrac{1}{\sigma}\big(y_t - \boldsymbol{X}_{1t}\boldsymbol{\beta}_1 - \boldsymbol{X}_{2t}\boldsymbol{\beta}_2\big)\Big)\Big), \tag{17}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative standard normal distribution functions, respectively. Since $\sigma$ 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 (17) 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 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 ESS is equal to the ES form of the LM statistic. The third test statistic is the OPG form of the LM test, which may also be computed via an artificial regression. The regressand is a vector of 1s, the $t^{\text{th}}$ element of each of the regressors is the derivative with respect to one of the parameters of the term in (17) 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 $W_\beta$. The fifth test statistic, which we call $W_\gamma$, 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://qed.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 $\beta_c$, a slope coefficient $\beta_s$, which is the same for elements 2 through $k_1$ of $\boldsymbol{\beta}_1$, and a variance $\sigma^2$. 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 $y_t$ were 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 first set of experiments was 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 $\sigma$ is particularly important. Figures 1 through 5 therefore plot rejection frequencies (observed $P$ values) for the five tests as functions of $\sigma$, all for $n = 50$, based on 50,000 replications for each of a large number of values of $\sigma$. We chose $n = 50$ because it is the smallest sample size for which we did not have to drop very many samples for the $k_1 = 5, k_2 = 8$ case. In the experiments reported in these figures, $\beta_c = 0$ and $\beta_s = 1$; this implies that, on average, $y_t = 0$ half the time. The vertical axis in these figures shows the fraction of the time that an asymptotic test at the nominal .05 level actually leads to rejection. Notice that the scale of the vertical axis is not the same in all the figures. The tests that perform worst are the OPG form of the LM test and the $W_\beta$ test, and the test that performs best, usually by a wide margin, is the ES form of the LM test. In almost all cases, test performance deteriorates as $\sigma$ becomes smaller.

Figures 1 through 5 make it clear that all the tests perform less well as either $k_1$ or $k_2$ is increased. For all but one of the tests, increases in either $k_1$ or $k_2$ lead to increased overrejection. For the ES LM test, however, increases in $k_1$ lead to increased overrejection, while increases in $k_2$ have the opposite effect, offsetting the overrejection in some cases and leading to underrejection in others. The $(8,2)$ case was included precisely because it is the one for which the ES LM test performs worst. These results 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 6 shows that test performance also depends on $\beta_c$, which determines the fraction of the sample for which $y_t$ is nonzero. This figure graphs the $P$ value functions for all the tests in what is usually the worst case ($k_1 = 5, k_2 = 8$) against $\beta_c$. For this figure, $n = 50$, $\beta_s = 1$, and $\sigma = 1$. Note that, for the smallest values of

$\beta_c$, it was sometimes necessary to omit quite a lot of samples (up to about 40%). It is clear from the figure that all the tests, except perhaps the ES LM test, overreject more severely as the number of zero observations increases. The LR test and the OPG LM test deteriorate particularly rapidly.

We also experimented with changing the slope coefficient $\beta_s$. The results of these experiments were predictable. When $\beta_c$ is zero, increasing $\beta_s$ is equivalent to reducing $\sigma$, and the effect on finite-sample test size is therefore precisely what one would predict from Figures 1 through 5. When $\beta_c$ is not zero, changing $\beta_s$ is equivalent to changing both $\beta_c$ and $\sigma$, and, again, the results are precisely what one would predict.

Finally, we conducted a series of experiments in which we changed the sample size, $n$. As expected, increasing $n$ led to improved performance by all the tests. Figure 7 shows how the true size of all the tests at the nominal .05 level varies 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. For all but one of the tests, the rate at which they do so appears to be roughly proportional to $n^{-1}$. For the OPG LM test, the rate appears to be slower than $n^{-1}$ but faster than $n^{-1/2}$.

The experiments discussed up to this point were designed to investigate the performance of the asymptotic tests. If these tests always worked perfectly, there would be no point bootstrapping them. On the other hand, if their performance did not depend on the parameters of the model, bootstrap testing would work perfectly; see Davidson and MacKinnon (1996a). Thus Figures 1 through 7 make it clear that it may well be attractive to bootstrap all these tests, and that it is of interest to see how the bootstrap tests perform. The remaining experiments therefore concern the performance of bootstrap tests. In all cases, bootstrap $P$ values were calculated using 399 bootstrap samples.

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, one based on the OPG regression, one based on the artificial regression of Orme (1995), and one based on Newton's method, using the $(\boldsymbol{\gamma}, \delta)$ parametrization. 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.

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 rounds than after one round. The OPG regression also works dreadfully for the other tests, of course. In contrast, the Orme regression works quite well, and Newton's method works extremely well. This suggests that how well approximate bootstrap methods will work is likely to vary greatly from case to case, and will depend on precisely what iterative procedure is used.

As the theory predicts, $m$ rounds 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. What is perhaps surprising is that $m = 2$ is often adequate for the other two tests as well. For $n = 200$, a single round of Newton's method always provides an adequate approximation for the LR test, although it does not do so for the other tests. Unfortunately, the approximate bootstrap procedure is not quite as useful in this case as these results suggest. It appears that maximizing the loglikelihood generally requires no more than three Newton steps, so using the approximate bootstrap with $m = 2$ does not actually save a great deal of CPU time.

The remaining experiments were designed to see how well bootstrap tests work. Figure 8 shows $P$ value discrepancy plots for all five bootstrap tests for six of the roughly twenty experiments that we ran. Nominal test size is plotted on the horizontal axis, and the difference between actual and nominal size 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 (1996c). 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 8 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 $\sigma$ 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 $\sigma$ 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. Note that, 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.

These results are entirely consistent with the theory of the size distortion of bootstrap tests developed in Davidson and MacKinnon (1996a). Bootstrap tests should not work perfectly, because the $P$ value functions in Figures 1 through 6 are not flat, but they should work well, because the functions are neither very steep nor very nonlinear. Because the parameters will be estimated less precisely when $\sigma$ is larger, it is reasonable for performance to become worse as $\sigma$ increases. In general, the size distortion of a bootstrap test should be $O(n^{-1})$ lower than the order of the size distortion of the corresponding asymptotic test. This is entirely consistent with what we see in panels E and F, when we recall that the asymptotic version of the OPG LM test seems to converge more slowly than the other asymptotic tests.

Although some size distortions are evident in Figure 8, 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 8, 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 2). In contrast, the worst performance we observed for the bootstrap version of this test was a rejection rate of 4.77% in the case $k_1 = 8, k_2 = 2, \sigma = 5.0$.

Since all the bootstrap tests perform well in terms of size, the choice of which one(s) to use may depend in large part on their power. Of course, asymptotic theory suggests that all the tests should have similar power, on a size-adjusted basis, and the results of Davidson and MacKinnon (1996b) suggest that bootstrapping them should not affect size-adjusted power very much. We performed a moderately large number of experiments, all of which yielded qualitatively similar results that confirmed these theoretical results, with one exception. Because all the results were similar, we report only one case, for $n = 50$, $k_1 = 5$, $k_2 = 8$, and $\sigma = 5$, with all the elements of $\boldsymbol{\beta}_2$ equal to 1. Each column of $\boldsymbol{X}_2$ was correlated with one of the columns of $\boldsymbol{X}_1$, the correlation coefficient being 0.5; other things held constant, higher correlations reduce power. Results based on 50,000 replications, each using 399 bootstrap samples, are shown in Figure 9.

The left panel of Figure 9 plots the power of the five bootstrap tests against their nominal size, for size up to 0.25. The OPG LM test appears to be noticeably less powerful than the other four tests, which are quite closely grouped, with $W_\beta$ being somewhat less powerful than the other three. However, plotting power against nominal size is not really appropriate when all the tests do not have exactly the correct size under the null. The right panel therefore plots power against actual size, where the latter is evaluated by a Monte Carlo experiment in which the data are generated by an approximation to the pseudo-true null; see Davidson and MacKinnon (1996b). Some of the variation in power that was evident in the left panel has now vanished, reflecting the fact that some tests are more prone to underreject, both under the null and under the alternative, than others. The LR, ES LM, and $W_\gamma$ tests are now indistinguishable, and the $W_\beta$ test is only very slightly less powerful than they are. However, the OPG LM test remains noticeably less powerful than the other tests.

This result is typical. In every one of our experiments, the OPG LM test had less power than any of the other tests. Of course, as asymptotic theory predicts, its loss of power became less substantial as the sample size was increased, although the rate of improvement was fairly slow. The poor performance of the OPG test is evidently related to the way it estimates the information matrix and not to the fact that it uses only estimates under the null, since the ES LM test was always either the most powerful test or very close to being the most powerful.

## 5. Conclusions

We have shown that the cost of bootstrap testing for nonlinear models can, in principle, be reduced by using a small number of rounds of an iterative procedure based either on an artificial regression or on Newton's method, instead of nonlinear estimation. This is possible because root-$n$ consistent parameter estimates are always available in the context of bootstrap sampling. The theory implies that just two rounds of the iterative procedure should be sufficient for likelihood ratio tests, while three or four will be needed for Lagrange multiplier and Wald tests. Of course, since these results are based on asymptotic theory, they may or may not provide a good guide in any actual case.

The approximate bootstrap technique was applied to tests of slope coefficients in the tobit model. Most of the classical tests were found to perform quite poorly when asymptotic $P$ values were used, but they always performed dramatically better when $P$ values were calculated using the bootstrap. All of the tests had essentially the same power after bootstrapping, except for the OPG form of the LM test, which had noticeably less power. The OPG LM test also suffered from the greatest size distortions when bootstrapped. Differences between the size of the other bootstrapped tests were generally small, but the most reliable test appears to be the efficient score form of the LM test. However, this test is the most complicated to program. The LR test also works very well, is easy to program, and should be about equally expensive to compute if approximate bootstrap $P$ values are used.

## References


Amemiya, T. (1973). "Regression analysis when the dependent variable is truncated normal," *Econometrica*, 41, 997–1016.

Davidson, R. and J. G. MacKinnon (1984a). "Model specification tests based on artificial linear regressions," *International Economic Review*, 25, 485–502.

Davidson, R. and J. G. MacKinnon (1984b). "Convenient specification tests for logit and probit models," *Journal of Econometrics*, 25, 241–262.

Davidson, R. and J. G. MacKinnon (1990) "Specification tests based on artificial regressions," *Journal of the American Statistical Association*, 85, 220–227.

Davidson, R. and J. G. MacKinnon (1996a). "The size distortion of bootstrap tests," GREQAM Document de Travail No. 96A15 and Queen's Institute for Economic Research Discussion Paper No. 936.

Davidson, R. and J. G. MacKinnon (1996b). "The power of bootstrap tests," Queen's Institute for Economic Research Discussion Paper No. 937.

Davidson, R. and J. G. MacKinnon (1996c). "Graphical methods for investigating the size and power of hypothesis tests," Queen's Institute for Economic Research Discussion Paper No. 903 (revised).

Gregory, A. W., and M. R. Veall (1985). "On formulating Wald tests for nonlinear restrictions," *Econometrica*, 53, 1465–1468.

Gregory, A. W., and M. R. Veall (1987). "Formulating Wald tests of the restrictions implied by the rational expectations hypothesis," *Journal of Applied Econometrics*, 2, 61–68.

Horowitz, J. L. (1994). "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, 61, 395–411.

Olsen, R. J. (1978). "Note on the uniqueness of the maximum likelihood estimator of the tobit model," *Econometrica*, 46, 1211–1215.

Orme, C. (1995). "On the use of artificial regressions in certain microeconometric models," *Econometric Theory*, 11, 290–305.

Phillips, P. C. B., and J. Y. Park (1988). "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, 56, 1065–83.

Tobin, J. (1958). "Estimation of relationships for limited dependent variables," *Econometrica*, 26, 24–36.

**Table 1. Average absolute differences between $\check{p}^*$ and $\hat{p}^*$**

| $n$ | $m$ | LR (OPG) | LR (Orme) | LR (N) | LM (N) | $W_\gamma$ (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.

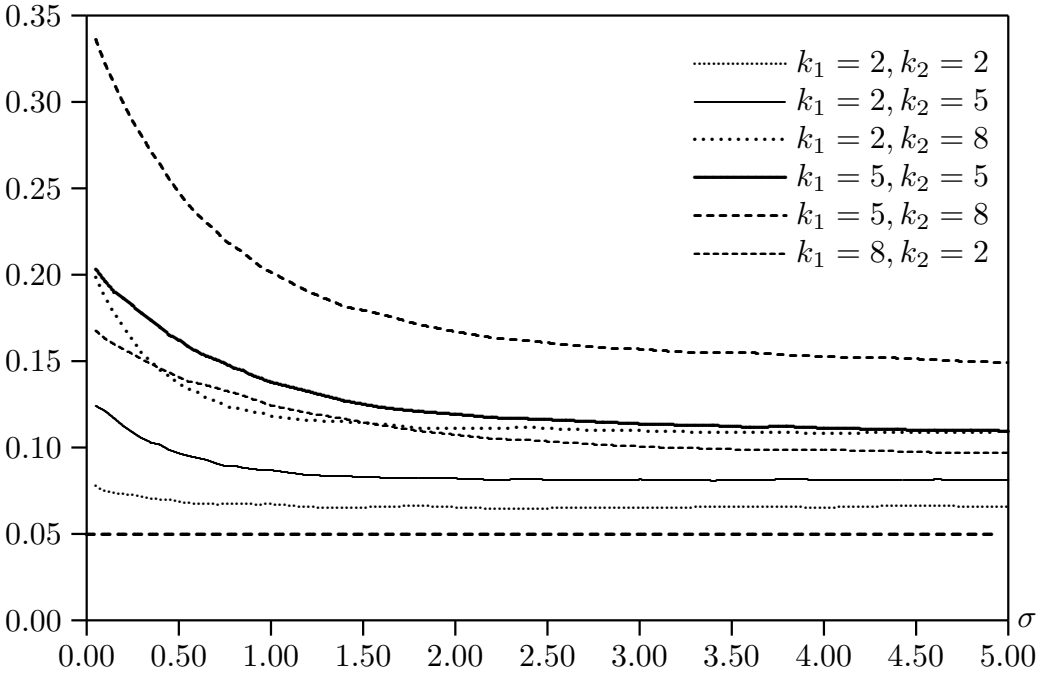



**Figure 1. $P$ value functions for LR tests at .05 level, $n = 50$**

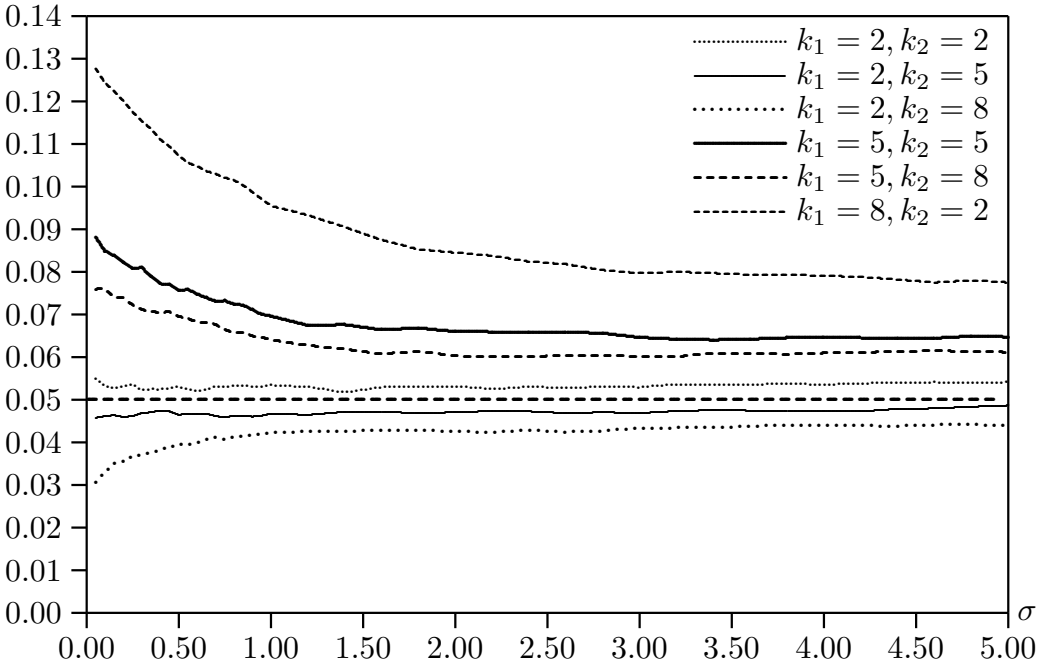



**Figure 2. $P$ value functions for ES LM tests at .05 level, $n = 50$**

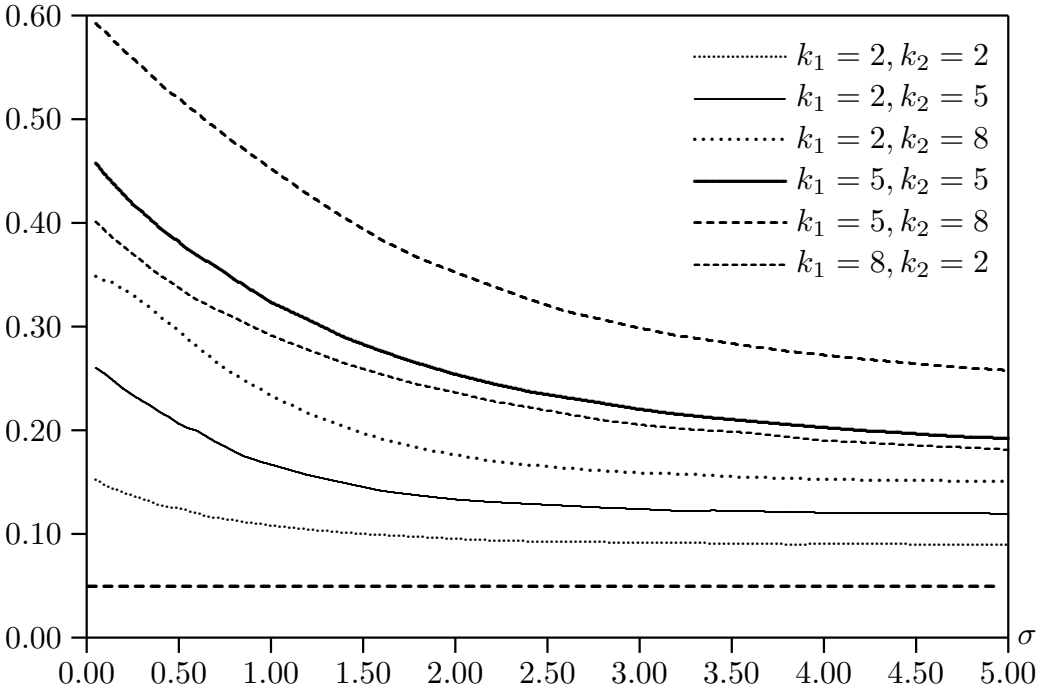



**Figure 3. $P$ value functions for OPG LM tests at .05 level, $n = 50$**

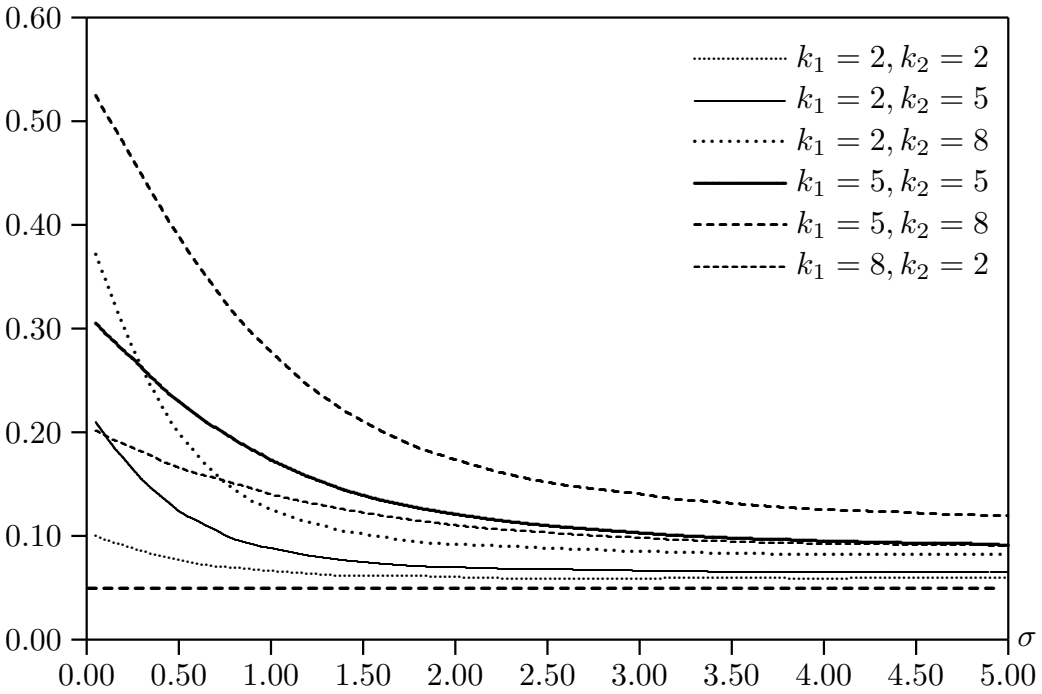



**Figure 4. $P$ value functions for Wald tests ($W_\beta$) at .05 level, $n = 50$**

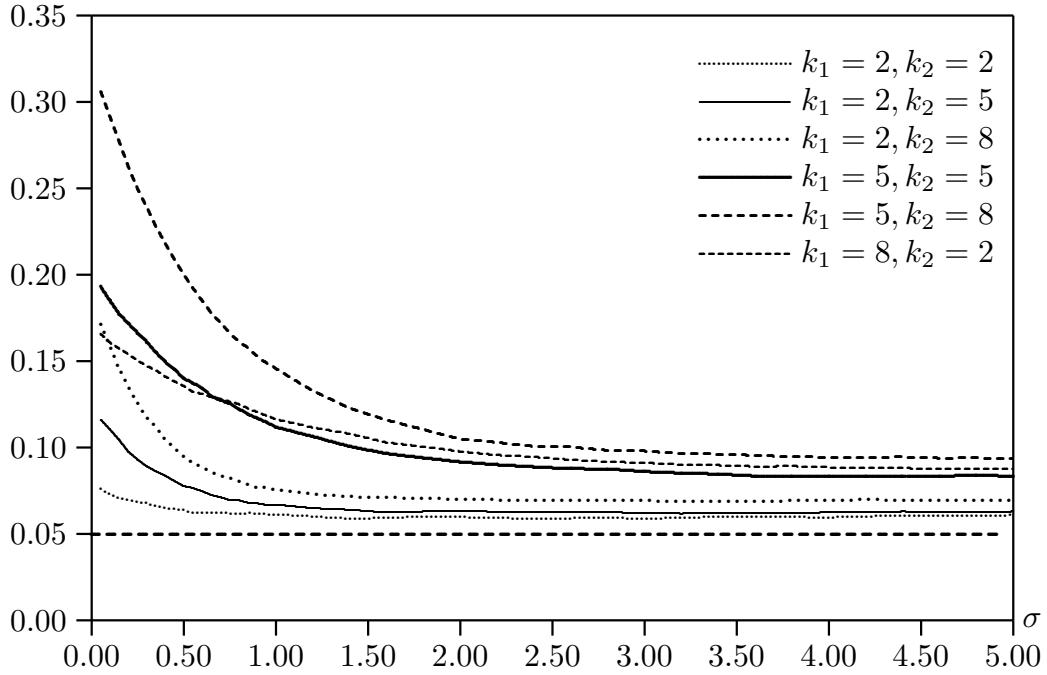



**Figure 5. $P$ value functions for Wald tests ($W_\gamma$) at .05 level, $n = 50$**

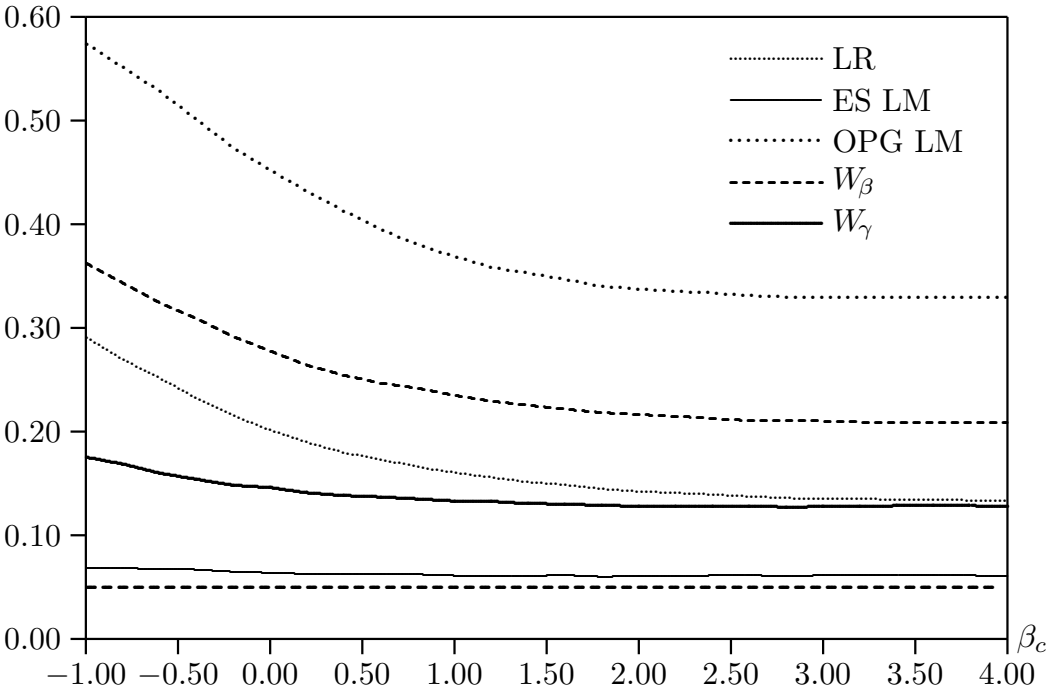



**Figure 6. $P$ value functions for tests at .05 level, $k_1 = 5$, $k_2 = 8$, $n = 50$**

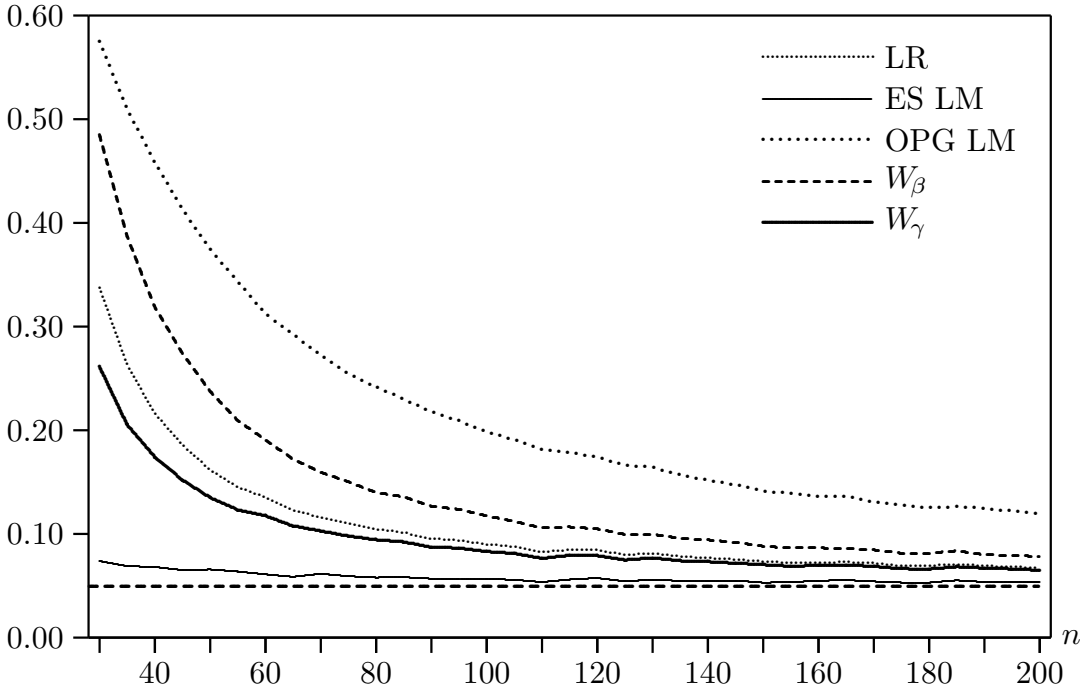



**Figure 7. Rejections at .05 level as a function of sample size**

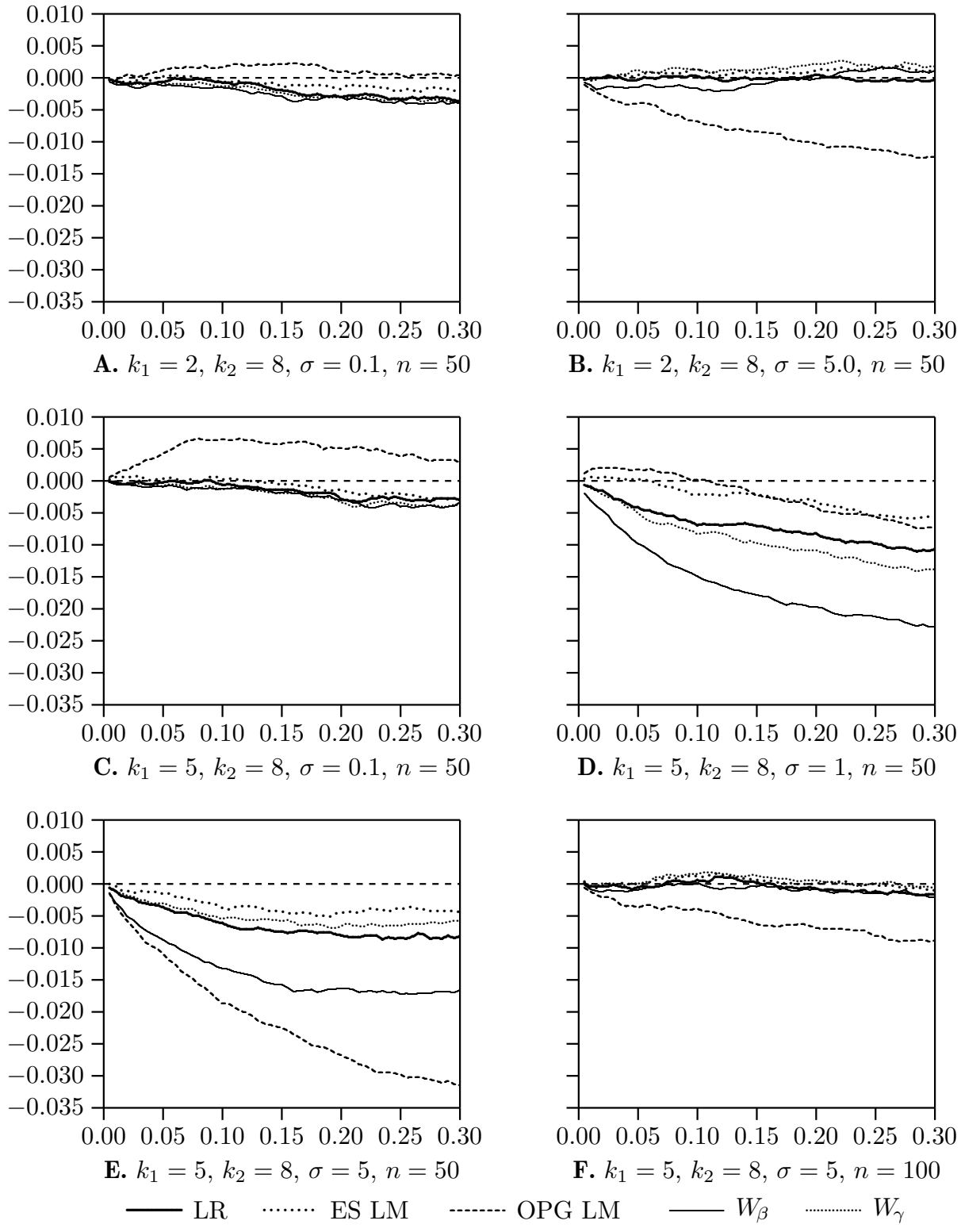
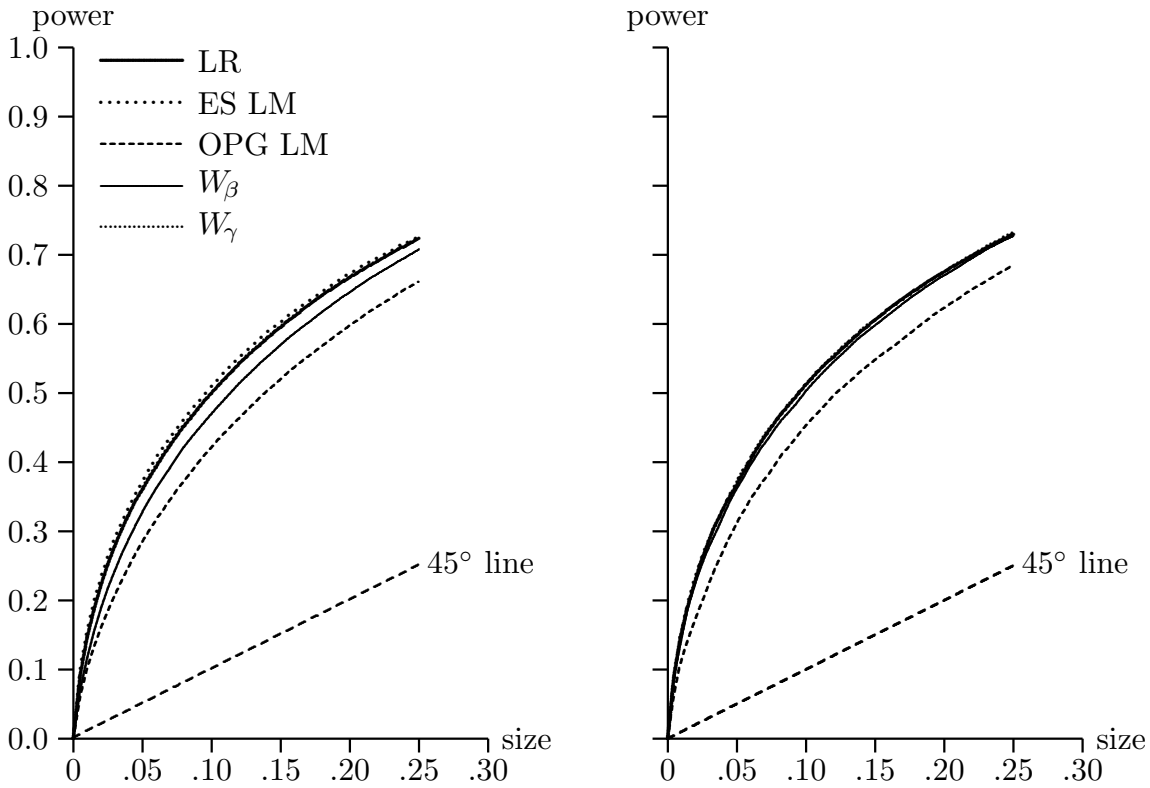



**Figure 8. *P* value discrepancy plots for bootstrap tests**




**A.** Power versus nominal size

**B.** Power versus actual size

**Figure 9. Size-power curves for bootstrap tests, $n = 50$**