Bootstrap inference in econometrics

James G. MacKinnon *Department of Economics, Queen's University*

Abstract. The astonishing increase in computer performance over the past two decades has made it possible for economists to base many statistical inferences on simulated, or bootstrap, distributions rather than on distributions obtained from asymptotic theory. In this paper, I review some of the basic ideas of bootstrap inference. I discuss Monte Carlo tests, several types of bootstrap test, and bootstrap confidence intervals. Although bootstrapping often works well, it does not do so in every case.

Inférence par la méthode d'auto-amorçage (bootstrap) en économétrie. L'incroyable accroissement dans la puissance des ordinateurs au cours des deux dernières décennies a permis aux économistes de fonder plusieurs inférences sur des distributions simulées, ou obtenues par auto-amorçage, plutôt que sur des distributions obtenues par la théorie aymptotique. Dans ce texte, l'auteur passe en revue quelques-unes des idées de base de l'inférence par la méthode d'auto-amorçage. Le texte discute aussi des tests de Monte Carlo, de divers types de tests et des intervalles de confiance obtenus par la méthode d'auto-amorçage. Même si le processus d'auto-amorçage fonctionne souvent bien, cela n'est pas toujours le cas.

1. Introduction

One of the most remarkable examples of technological progress has been the massive increase in the speed of digital computers during the past forty years. For scientific computing, a typical personal computer of today is several hundred times as fast as a typical PC of just a decade ago, although it costs less than half as much. Even the PC of a decade ago was faster than multi-million dollar mainframe com-

This paper was presented as the Presidential Address at the 2002 Annual Meeting of the Canadian Economics Association. The research was supported, in part, by two grants from the Social Sciences and Humanities Research Council of Canada. I am grateful to John Galbraith, Russell Davidson, and Dwayne Benjamin for comments. Email: jgm@qed.econ.queensu.ca

Canadian Journal of Economics / Revue canadienne d'Economique, Vol. 35, No. 4
November / novembre 2002. Printed in Canada / Imprimé au Canada

0008-4085 / 02 / 615–645 /

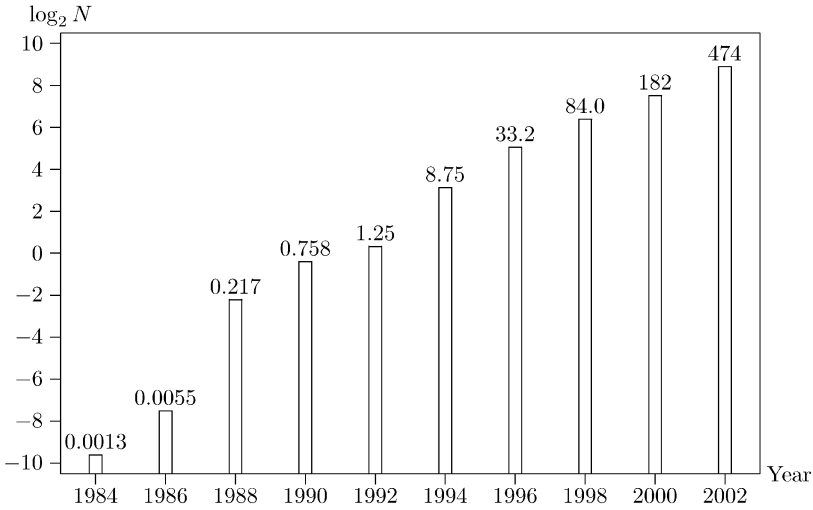

FIGURE 1 Regressions per second on personal computers

puters of just two decades ago. Some of this progress is documented in figure 1, which shows the number of medium-sized ordinary least squares regressions (4000 observations and 20 regressors) that a more or less state-of-the-art, but affordable, personal computer could perform in a single second at various points in time over the past two decades.

Since computer time is now very cheap, it makes sense for applied econometricians to use far more of it than they did just a few years ago. There are at least three ways in which they have been doing so. One approach is to estimate very ambitious structural models, which often involve explicitly modelling choice at the level of individual agents. For many of these models, simulations are needed simply to estimate the model. Because this is very time-consuming, inference is normally based on standard asymptotic results. Important early examples of this approach include Pakes (1986) and Rust (1987). Eckstein and Wolpin (1989) and Stern (1997) provide useful surveys.

A second line of research involves Bayesian estimation using Markov-chain Monte Carlo methods; see, among many others, Albert and Chib (1993), McCulloch and Rossi (1994), Geweke (1999), and Elerian, Chib, and Shephard (2001). With this approach, inference is exact, in the Bayesian sense, but it generally depends upon strong distributional assumptions and the investigator's prior beliefs.

The third line of research, which is the one I will discuss here, is to base statistical inferences on distributions that are calculated by simulation rather than on

1 The numbers in figure 1 are based on my own Fortran programs run on various machines that I have had access to. Times are for an XT clone (1984), a 286/10 (1986), a 386/20 (1988), a 486/25 (1990), a 486DX2/50 (1992), a Pentium 90 (1994), a Pentium Pro 200 (1996), a Pentium II/450 (1998), an Athlon 800 (2000), and a Pentium 4/2200 (2002).

ones that are suggested by asymptotic theory and are strictly valid only when the sample size is infinitely large. In this approach, parameter estimates and test statistics are calculated in fairly conventional ways, but *P* values and confidence intervals are computed using 'bootstrap' distributions obtained by simulation. This bootstrap approach can often, but does not always, lead to much more accurate inferences than traditional approaches are capable of. Like every tool in econometrics, however, it must be used with care.

The reason for using bootstrap inference is that hypothesis tests and confidence intervals based on asymptotic theory can be seriously misleading when the sample size is not large. There are many examples. One is the popular *J* test of non-nested regression models (Davidson and MacKinnon 1981), which always rejects the null hypothesis too often. In extreme cases, even for sample sizes as large as 50, an asymptotic *J* test at the .05 level can reject a true null hypothesis more than 80 per cent of the time; see Davidson and MacKinnon (2002a). Some versions of the information matrix test overreject even more severely. Davidson and MacKinnon (1992) report a simulation in which one such test at the .05 level rejected a true null hypothesis an astounding 99.9 per cent of the time when the sample size was 200.

Of course, asymptotic tests are not always misleading. In many cases, a bootstrap test will yield essentially the same inferences as an asymptotic test based on the same test statistic. Although this does not necessarily imply that the asymptotic test is reliable, the investigator may reasonably feel greater confidence in the results of asymptotic tests that have been confirmed in this way.

Statistical inference in a classical framework involves either testing hypotheses or constructing confidence intervals. In most of this paper I will focus on hypothesis testing, because simulation-based hypothesis testing is generally easier and more reliable than constructing simulation-based confidence intervals. Moreover, hypothesis tests and confidence intervals are very closely related, so that much of what is said about bootstrap tests will also be applicable to bootstrap confidence intervals.

In the next section I discuss Monte Carlo tests, which can be thought of as a special case of bootstrap tests. In section 3 I go on to discuss bootstrap tests more generally. In section 4 I explain why bootstrap tests will often work well and I provide evidence from a simulation experiment for a case in which they work extremely well. In section 5 I consider three common situations in which bootstrap tests do not always work well, and I provide evidence from several simulation experiments that illustrates the problems that can arise. The power of bootstrap tests is dealt with in section 6. Finally, bootstrap confidence intervals are briefly discussed in section 7.

2. Monte Carlo Tests

Statisticians generally make a distinction between two types of simulation-based tests, namely, *Monte Carlo tests* and *bootstrap tests*. In this section, I will begin by discussing a fairly simple example of a Monte Carlo test. In the next section, I will move on to a discussion of bootstrap tests.

One of the best-known test statistics in econometrics is the d statistic proposed by Durbin and Watson (1950, 1951) for testing the null hypothesis that the error terms of a linear regression model are serially uncorrelated. The model under test is

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \tag{1}$$

where there are n observations, and the row vector of regressors $\boldsymbol{X}_t$ is treated as fixed. If $\hat{u}_t$ denotes the t^{th} residual from OLS estimation of (1), then the Durbin-Watson statistic is

$$d = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}. \tag{2}$$

The finite-sample distribution of d depends on $\boldsymbol{X}$, the matrix of regressors with t^{th} row $\boldsymbol{X}_t$. Therefore, it is customary to base inferences on tables that merely provide bounds on the critical values. Since these bounds are often quite far apart, it is frequently impossible to tell whether the null hypothesis of serial independence should be rejected. The exact distribution of d can be calculated, but very few econometric packages do so.

Under the null hypothesis, the vector of OLS residuals is

$$\hat{\boldsymbol{u}} = \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{u}, \quad \text{where} \quad \boldsymbol{M}_{\boldsymbol{X}} \equiv \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Thus, the residuals depend solely on the error terms and the matrix $\boldsymbol{X}$. Since they are evidently proportional to σ, we can see from (2) that the statistic d depends only on the normalized error terms $\epsilon_t \equiv u_t/\sigma$ and the matrix $\boldsymbol{X}$. It does not depend on $\boldsymbol{\beta}$ or σ^2 at all. A statistic like d that does not depend on any unknown parameters is said to be *pivotal*. For any pivotal statistic, we can perform an exact simulation-based test. Such a test is called a Monte Carlo test. When we say that the test is *exact*, we mean that the probability of rejecting the null hypothesis when it is true is precisely equal to α, the nominal level of the test, which would often be .10, .05, or .01.

The steps required to perform a Monte Carlo test based on the Durbin-Watson statistic are as follows:

1. Estimate the linear regression model (1) and compute d.
2. Choose B, the number of simulations. It should be chosen so that $\alpha(B + 1)$ is an integer for all levels α of interest. A common choice is $B = 999$. Since estimating a linear regression model and computing d is very inexpensive, however, $B = 9999$ might be a slightly better choice. It is also possible to start with a small value like 99 and then increase it if the outcome of the test is not clear.
3. Generate B simulated samples, or *bootstrap samples*, indexed by j, by drawing vectors of errors $\boldsymbol{u}_j^*$ from the standard normal distribution. Then regress them on $\boldsymbol{X}$ to generate B simulated residual vectors $\hat{\boldsymbol{u}}_j^*$.

4. For each bootstrap sample, compute d_j^* from the bootstrap residuals $\hat{\boldsymbol{u}}_j^*$ using the formula (2).
5. Use d and the d_j^* to compute a P value. The Durbin-Watson statistic is often treated as a one-tailed test against positive serial correlation, and the null hypothesis is rejected whenever d is sufficiently small. For such a test, the simulated P value is

$$p^*(d) = \frac{1}{B}\sum_{j=1}^{B} I(d_j^* \leq d).$$

   Here $I(\cdot)$ denotes the indicator function, which is equal to 1 if its argument is true and 0 otherwise. Thus, the simulated P value is the fraction of the time that d_j^* is smaller than d. Alternatively, we could test against negative serial correlation by calculating the fraction of the time that d_j^* is larger than d.
6. Reject the null hypothesis that the error terms are serially independent if whichever simulated P value is appropriate is less than α, the level of the test. To test for both positive and negative serial correlation, we could reject whenever the simulated P value for either one-tailed test is less than $\alpha/2$.

The simulated P value $p^*(d)$ makes sense intuitively. If a substantial proportion of the d_j^* is more extreme than d, then the probability of obtaining a test statistic as or more extreme than d must be high, the simulated P value will be large, and we will not reject the null hypothesis that the error terms are serially uncorrelated. Conversely, if very few of the d_j^* are more extreme than d, then the probability of obtaining a test statistic as or more extreme than d must be low, the simulated P value will be small, and we will reject the null hypothesis.

Because d is pivotal, this procedure yields an exact test. Suppose, for concreteness, that $B = 99$ and $\alpha = .05$. Then there are 100 possible values that $p^*(d)$ can take on: $0, 1/99, 2/99, \ldots, 98/99, 1$. If the null hypothesis is true, d and the d_j^* come from exactly the same distribution. Therefore, every possible value of $p^*(d)$ has exactly the same probability, namely, $1/100$. There are five values that will cause us to reject the null: 0, 1/99, 2/99, 3/99, and 4/99. Under the null hypothesis, the probability that one of these five values will occur by chance is precisely .05. Note that this argument would not work if $\alpha(B + 1)$ was not an integer.

Simulation necessarily introduces randomness into our test procedure, and it seems clear that this must have a cost. In this case, the cost is a loss of power. A test based on 99 simulations will be less powerful than a test based on $B = 999$, which in turn will be less powerful than one based on $B = 9999$, and so on. As we will see in section 6, however, the power loss is generally very small indeed when $B \geq 999$.

As I remarked above, it is possible to start with a small value of B, say $B = 99$, and then perform more replications only if the initial results are ambiguous. If, for example, 38 out of the first 99 bootstrap samples yield test statistics more extreme than the actual one, then we can be confident that the null hypothesis is not rejected at any standard significance level, and there is no need to generate any more bootstrap samples. If the initial results are not so clear, we can generate more bootstrap samples until we obtain a sufficiently accurate estimate of the P value. Davidson and MacKinnon (2000) propose a formal procedure for doing this.

Procedures very similar to the one just described for the Durbin-Watson test can also be used to perform a variety of other Monte Carlo tests of linear regression models with fixed regressors and error terms that follow a distribution known up to scale, which does not have to be the normal distribution. These include tests for higher-order serial correlation based on the Gauss-Newton regression, tests for heteroscedasticity, tests for skewness and kurtosis, and tests of certain types of linear restrictions in multivariate linear regression models. As long as the test statistic depends simply on error terms with a known distribution and fixed regressors, it will be pivotal, and we can perform an exact simulation-based test. Monte Carlo tests were first proposed by Dwass (1957). For a much fuller discussion of these tests, see Dufour and Khalaf (2001).

3. Bootstrap tests

When a test statistic is not pivotal, we cannot use a Monte Carlo test. However, we can use bootstrap tests that work in very much the same way. For example, the Durbin-Watson d statistic would no longer be pivotal if the distribution of the error terms were unknown. But we can still generate bootstrap samples (in a somewhat different way), compute the d_j^*, and calculate bootstrap P values exactly as before. These bootstrap P values will not be entirely accurate, but they will often be much more accurate than P values calculated from an asymptotic distribution.

Without the normality assumption, it does not make sense to generate bootstrap errors from the normal distribution. Instead, we want to generate them nonparametrically. One of the simplest and most popular approaches is to obtain the $\boldsymbol{u}_j^*$ by *resampling* the residuals $\hat{u}_t$. To generate a single bootstrap sample, we pick n integers between 1 and n at random with equal probability. If the t^{th} integer is equal to k, we set $u_t^* = \hat{u}_k$. In this way, we effectively generate the bootstrap error terms from the empirical distribution function of the residuals. Resampling is the key idea of the bootstrap as it was originally proposed by Efron (1982).

As every student of econometrics knows, OLS residuals have smaller variance than the error terms on which they are based. Thus it would seem to be desirable to resample not the vector $\hat{\boldsymbol{u}}$ of raw residuals but rather the vector of *rescaled residuals*

$$\tilde{\boldsymbol{u}} \equiv \left(\frac{n}{n-k}\right)^{1/2} \hat{\boldsymbol{u}}, \tag{3}$$

the elements of which have variance σ^2. In the case of the Durbin-Watson statistic and other test statistics that do not depend on σ^2, it makes absolutely no difference whether or not the residuals have been rescaled before we resample them. For many other test statistics, however, it is important to resample $\tilde{\boldsymbol{u}}$ instead of $\hat{\boldsymbol{u}}$.

Other, more complicated, methods of rescaling the residuals can also be used. For example, we could resample the vector with typical element

$$\ddot{u}_t = \left(\frac{n}{n-1}\right)^{1/2} \left(\frac{\hat{u}_t}{(1-h_t)^{1/2}} - \frac{1}{n}\sum_{s=1}^{n} \frac{\hat{u}_s}{(1-h_s)^{1/2}}\right), \tag{4}$$

where h_t denotes the t^{th} diagonal element of the *hat matrix*, that is, the matrix $\boldsymbol{P_X} \equiv \boldsymbol{I} - \boldsymbol{M_X}$. In expression (4), we divide by $(1 - h_t)^{1/2}$ in order to ensure that, if the error terms were homoscedastic, all the transformed residuals would have the same variance. We then subtract the mean of the residuals after the initial transformation so that the transformed residuals will have mean zero, and we then multiply by the square root of $n/(n-1)$ to undo the shrinkage caused by subtracting the mean.

In many cases, the unrestricted model is also a regression model, and we could resample the residuals for that model instead of the residuals for the restricted model. Indeed, several authors, including van Giersbergen and Kiviet (2002), have advocated doing precisely this, on the grounds that it will improve power. The argument is that, when the null is false, the unrestricted residuals will provide a better approximation to the distribution of the error terms. As we will see in section 6, this conjecture appears to be false.

Bootstrap tests based on non-pivotal test statistics are calculated in essentially the same way as Monte Carlo tests. We first calculate a test statistic, say τ, in the usual way. Possibly as a byproduct of doing so, we estimate the model under the null hypothesis and obtain estimates of all the quantities needed to generate bootstrap samples that satisfy the null hypothesis. The data-generating process used to generate these samples is called the *bootstrap DGP*. The bootstrap DGP may be purely parametric, but it often involves some sort of resampling so as to avoid making distributional assumptions. It must always satisfy the null hypothesis. We then generate B bootstrap samples, compute a bootstrap test statistic τ_j^* using each of them, and calculate the bootstrap P value $p^*(\tau)$ as the proportion of the τ_j^* that is more extreme than τ. If $p^*(\tau)$ is less than α, we reject the null hypothesis.

Instead of calculating a P value, some authors prefer to calculate a bootstrap critical value based on the τ_j^* and reject the null hypothesis whenever τ exceeds it. If the test is one for which we reject when τ is large, then the bootstrap critical value c_α^* is the $1 - \alpha$ quantile of the τ_j^*. When $\alpha(B+1)$ is an integer, this is simply number $(1-\alpha)(B+1)$ in the list of the τ_j^*, sorted from smallest to largest. For example, if $B = 999$ and $\alpha = .05$, then c_α^* is number 950 in the sorted list.

Rejecting the null hypothesis whenever τ exceeds c_α^* will yield exactly the same results as rejecting it whenever the bootstrap P value is less than α. Since it does not yield a P value, however, it normally provides less information. But calculating critical values may be desirable when B is small (because the test is expensive to compute) and τ is more extreme than all of the τ_j^*. In such a case, observing that τ greatly exceeds the estimated critical value may give us more confidence that the null is false than merely seeing that $p^*(\tau)$ is equal to 0.

4. When do bootstrap tests perform well?

The reason for using bootstrap tests instead of asymptotic tests is that we hope to make fewer mistakes by doing so. Many asymptotic tests overreject in finite samples, often very severely. Other asymptotic tests underreject, sometimes quite seriously. By using a bootstrap test instead of an asymptotic one, we can usually, but not always, make more accurate inferences. Ideally, a test will have a small *error in*

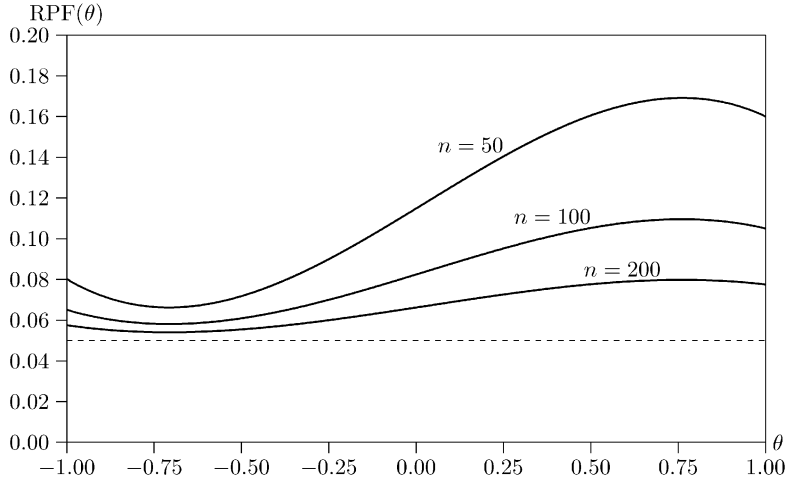

FIGURE 2 Rejection probability functions for three sample sizes

rejection probability, or *ERP*. This is the difference between the actual rejection frequency under the null hypothesis and the level of the test. The ERP of a bootstrap test is usually, but not always, smaller than the ERP of the asymptotic test on which it is based. In many cases, it is a great deal smaller.

The theoretical literature on the finite-sample performance of bootstrap tests, such as Beran (1988) and Hall and Titterington (1989), is primarily concerned with the rate at which the ERP of bootstrap tests declines as the sample size increases. It has been shown that, in a variety of circumstances, the ERP of bootstrap tests declines more rapidly than the ERP of the corresponding asymptotic test. The essential requirement is that the underlying test statistic be *asymptotically pivotal*. This is a much weaker condition than that the statistic be pivotal. It means that, as the sample size tends to infinity, any dependence of the distribution on unknown parameters or other unknown features of the data-generating process must vanish. Of course, any test statistic that has a known asymptotic distribution which is free of nuisance parameters, as do the vast majority of test statistics in econometrics, must be asymptotically pivotal.

It is not hard to see intuitively why the ERP of a bootstrap test based on an asymptotically pivotal test statistic will decline more rapidly than the ERP of an asymptotic test based on the same test statistic. Suppose, for simplicity, that the finite-sample distribution of a test statistic τ depends on just one nuisance parameter, say θ. Then we can graph the rejection probability of the asymptotic test as a function of θ, as in figure 2. If the *rejection probability function*, or *RPF*, is flat, then τ is pivotal, and the bootstrap test will work perfectly. If it is not flat, then the bootstrap test will not work perfectly, because the distribution of the τ_j^*, which is based on an estimate $\hat{\theta}$, will differ from the distribution of τ, which is based on the unknown true value θ_0.

As Davidson and MacKinnon (1999a) showed, the ERP of a bootstrap test depends on the slope of the RPF, but only if $\hat{\theta}$ is biased, and on its curvature, whether or not $\hat{\theta}$ is biased. Whenever τ is asymptotically pivotal, the RPF must converge to a horizontal line as the sample size tends to infinity. This is illustrated in figure 2, which shows RPFs for the same test for three different sample sizes. The fact that the slope and curvature of the RPF become smaller as the sample size increases would, by itself, cause the ERP of the bootstrap test to decrease at the same rate as the ERP of the asymptotic test. But increasing the sample size also causes both the bias and the variance of $\hat{\theta}$ to decrease. This further reduces the ERP of the bootstrap test, but it has no effect on the ERP of the asymptotic test. Therefore, as the sample size increases, the ERP of a bootstrap test should improve more rapidly than that of an asymptotic test based on the same test statistic.

This result does not imply that a bootstrap test will always outperform the corresponding asymptotic test. There may well be values of θ for which the latter happens to perform extremely well and the bootstrap test performs less well. If the ERP of an asymptotic test is large, however, then it is increasingly likely, as the sample size increases, that the ERP of a bootstrap test based on it will be smaller. In practice, it often seems to be very much smaller. Thus, by using bootstrap tests, we may be able to avoid the gross errors of inference that frequently occur when we act as if test statistics actually follow their asymptotic distributions.

Let us now consider a specific example that illustrates the relationship between asymptotic and bootstrap tests. When a regression model includes lagged dependent variables, the Durbin-Watson statistic is not valid. In this situation, one popular way of testing for first-order serial correlation, which was suggested by Durbin (1970) and Godfrey (1978), is to run the original regression again with the lagged OLS residuals added as an additional regressor. The t statistic on the lagged residuals, which we will refer to as the Durbin-Godfrey statistic, can be used to perform an asymptotically valid test for first-order serial correlation; see Davidson and MacKinnon (1993, chap. 10).

The finite-sample distribution of the Durbin-Godfrey statistic depends on the sample size, the matrix of regressors, and the values of all the parameters. The parameter on the lagged dependent variable is particularly important. For purposes of illustration, I generated data from the model

$$y_t = \beta_1 + \sum_{j=2}^{4} \beta_j X_{tj} + \delta y_{t-1} + u_t, \qquad u_t \sim N(0, \sigma^2), \tag{5}$$

where $n = 20$, the X_{tj} display positive serial correlation ($\rho = 0.75$), all the β_j are equal to 1, $\sigma = 0.1$, and δ is allowed to vary between -0.99 and 0.99. Figure 3, which is based on 500,000 replications for each of 199 values of δ, shows the RPF for a Durbin-Godfrey test based on the Student's t distribution at the .05 level.

We can see from figure 3 that, in this particular case, the ordinary Durbin-Godfrey test may either overreject or underreject. In the worst case, when $\delta = 0.96$, it rejects 9.07% of the time. Note that these results are very specific to the model (5) and the parameter values I used. Both the shape and the level of the RPF can be quite different from what they are in the figure.

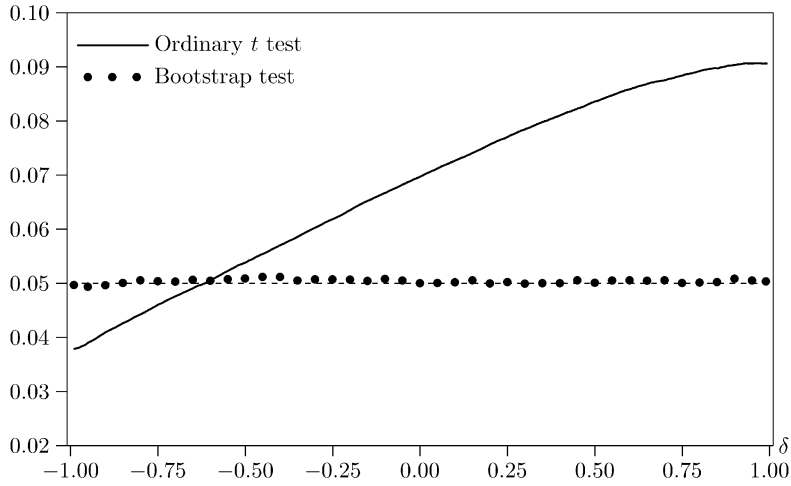

FIGURE 3 Rejection frequencies for Durbin−Godfrey test, $n = 20$

Figure 3 also shows the rejection frequencies for a bootstrap version of the Durbin-Godfrey test for 41 values of δ $(-0.99, -0.95, \ldots, 0.95, 0.99)$. These are based on only 100,000 replications, with $B = 399$, and are shown as bullets. I would not recommend using such a small value of B in practice, but sampling errors tend to cancel out in a Monte Carlo experiment like this one. Data were generated recursively from (5), using the OLS estimates under the null (but with $\hat{\delta}$ constrained not to exceed 0.999), and the bootstrap error terms were generated by resampling residuals rescaled using (3). The actual value y_0 was used for the initial value of the lagged dependent variable.

At first glance, it appears from figure 3 that the bootstrap test works perfectly, as all the observed rejection frequencies are extremely close to 0.05. Closer examination reveals, however, that even though the bootstrap test works extraordinarily well, its performance varies with δ, and it does not actually work perfectly. For example, for every value of δ between -0.80 and -0.05, the bootstrap test always overrejects, although in the worst case it rejects just 5.11 per cent of the time.

In general, bootstrap tests seem to perform extremely well in the context of single-equation models with exogenous or predetermined regressors and errors that are independent and identically distributed. For example, Davidson and MacKinnon (1999b) show that both bootstrap tests of common factor restrictions and bootstrap tests for omitted variables in the tobit model perform very well indeed with samples of modest size. What is more interesting, Davidson and MacKinnon (2002a) consider the J test of nonnested linear regression models, which often overrejects very severely as an asymptotic test. They show, both theoretically and via simulation, that bootstrapping the J test largely eliminates the overrejection in most cases. In certain extreme cases, in which the ordinary bootstrap J test still overrejects noticeably, a more sophisticated bootstrap test proposed by Davidson and MacKinnon (2002b) greatly reduces the remaining overrejection.

Based on this and other evidence, both published and unpublished, I would be very surprised to encounter a bootstrap test that did not work well in the context of a single-equation regression model or a single-equation limited-dependent variable model such as the logit, probit, or tobit models, provided the regressors are exogenous or predetermined and the underlying error terms are homoscedastic and serially uncorrelated.

5. When do bootstrap tests perform badly?

As the qualifications at the end of the preceding paragraph suggest, there are at least three situations in which bootstrap tests cannot be relied upon to perform particularly well. I briefly discuss each of these in this section.

5.1. Models with serial correlation

Economists commonly encounter models with serial correlation of unknown form. This situation frequently arises in the context of GMM estimation, and it almost always arises when we wish to test the null hypothesis that a time series has a unit root. I will focus on the latter case here.

A procedure that is widely used to test the unit root hypothesis is the augmented Dickey-Fuller (or ADF) test, one version of which is based on the regression

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \sum_{j=1}^{p} \delta_j \Delta y_{t-j} + u_t, \tag{6}$$

where y_t is an observation on the time series to be tested, and $\Delta y_t \equiv y_t - y_{t-1}$. One popular test statistic is τ_c, the ordinary t statistic for $\beta_1 = 0$ in regression (6). However, it does not follow the Student's t distribution, even asymptotically. Its asymptotic distribution is known, but it depends on functionals of Wiener processes and must be computed by simulation. Very accurate asymptotic critical values and P values may be obtained by using the program of MacKinnon (1996), which uses response surface estimates based on a large number of simulation experiments.

The lagged values of the dependent variable are included in regression (6) in order to remove any serial correlation that would otherwise be present. For any given p, the asymptotic distribution of τ_c will depend on the pattern of serial correlation in the error terms. If p is allowed to grow with the sample size at a suitable rate, however, this dependence will vanish, and the asymptotic distribution will be the same as if there were no serial correlation; see Galbraith and Zinde-Walsh (1999). In practice, the value of p is usually chosen by some sort of formal or informal testing procedure, which can be misleading; see Ng and Perron (1995, 2001).

Although the asymptotic distribution of τ_c does not depend on the time-series properties of the error terms, the finite-sample distribution certainly does. There have been many Monte Carlo studies on this topic, a classic one being Schwert (1989). It therefore seems natural to use a bootstrap test instead of an asymptotic one. As we will see, there can certainly be something to be gained by doing so, but

the improvement, when there is any, tends to be much less dramatic than it was in the example of the previous section.

If the bootstrap is to work well, we need to generate bootstrap error terms that display the same sort of serial correlation as the real ones, without knowing how the real error terms were generated. This is evidently quite a challenging task. There are two popular, and very different, approaches.

The first approach, which is semiparametric, is called the *sieve bootstrap*. We first impose the unit root null and estimate an autoregressive model of order p, where p is chosen in a way that allows it to increase with the sample size. We then generate simulated innovations by resampling the rescaled residuals from the autoregressive model. The serially correlated bootstrap error terms are then constructed from the model and the innovations. For details, see Bühlmann (1997, 1998), Choi and Hall (2000), Park (2002), and Chang and Park (2002). Although it has merit, this approach is not entirely satisfactory. For samples of moderate size, the AR(p) approximation may not be a good one. Even if it is, the parameter estimates are certain to be biased. Thus, it is likely that the bootstrap samples will differ from the real one in important respects.

The second approach, which is fully non-parametric, is to resample groups of residuals. Conceptually, one of the simplest such methods is the *block bootstrap*, which has been proposed in various forms by Carlstein (1986), Künsch (1989), Politis and Romano (1994), and a number of other authors. One particular block bootstrap procedure works as follows:

- Pick a block length $b < n$.
- Form n blocks of b residuals, each starting with a different one of the n residuals, and (in this version) wrapping around to the beginning if necessary.
- Generate the bootstrap errors by resampling the blocks. If n/b is an integer, there will be n/b blocks. Otherwise, the last block will have to be shorter than b.

Numerous other block bootstrap procedures exist, some of which have better theoretical properties than others; see Lahiri (1999). Although the one just described is probably as good as any, it is far from satisfactory. The blocks of residuals will not have the same properties as the underlying error terms, and the patterns of dependence within each block will be broken between each block and where the blocks wrap around. Nevertheless, it can be shown that, if b is allowed to tend to infinity at the correct rate, which is slower than the rate at which n does so, the bootstrap samples will have the right properties asymptotically.

To illustrate these two approaches, I have undertaken a few Monte Carlo experiments that examine the performance of the ADF test based on regression (6), with $p = 4$, in a very special case. The data were generated by

$$\Delta y_t = u_t, \qquad u_t = \rho u_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0,1), \tag{7}$$

in which the error terms follow an AR(1) process with parameter ρ. The parameters β_0 and β_1 that appear in (6) do not appear here, because both are equal to 0 under

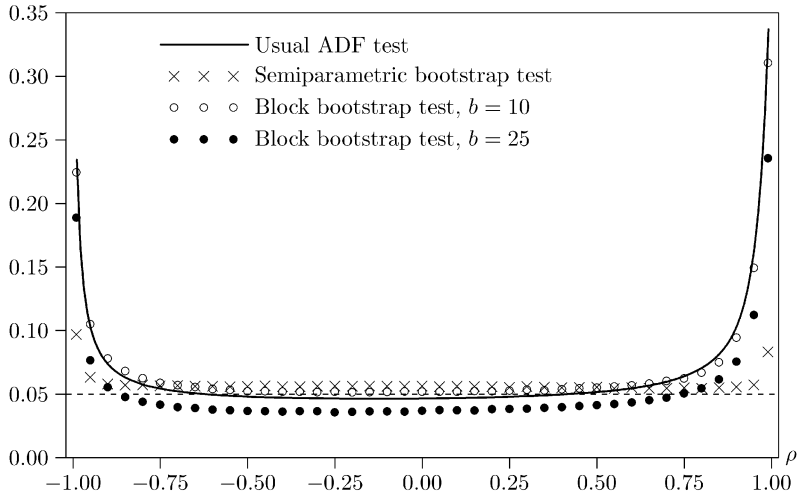

FIGURE 4 Rejection frequencies for ADF tests, $n = 100$

the null hypothesis of a unit root. In all the experiments, there were $n = 100$ observations.

Figure 4 shows rejection frequencies at the .05 level as a function of ρ for four different tests. The solid line, which is based on one million replications for each of 199 equally spaced values of ρ between -0.99 and 0.99, shows the rejection frequencies for the usual ADF test based on the critical value -2.8906, which is what the program of MacKinnon (1996) gives for a sample of size 100. The various symbols show rejection frequencies for three different bootstrap versions of the same test, based on 100,000 replications with $B = 399$ for 41 different values of ρ ($-0.99, -0.95, -0.90, \ldots, 0.90, 0.95, 0.99$). The crosses correspond to a semiparametric bootstrap test, in which the investigator estimates the model (7) by maximum likelihood (to ensure that $|\hat{\rho}| < 1$) and then resamples the residuals. This should work better than the sieve bootstrap does, because we are estimating the correct model. The circles and bullets correspond to non-parametric bootstrap tests, in which the block bootstrap with, respectively, $b = 10$ and $b = 25$ is used to generate the bootstrap errors from the observed values of Δy_t.

The results in figure 4 are not particularly encouraging. The semiparametric bootstrap test overrejects slightly for most values of ρ and quite substantially for $|\rho| = 0.99$. For extreme values of ρ, however, it overrejects much less severely than the asymptotic test does. The two block bootstrap tests give mixed results. The one with $b = 10$ performs very well for moderate values of ρ, but it performs just as badly as the usual ADF test when $|\rho|$ is large. The one with $b = 25$ underrejects quite noticeably for values of ρ between -0.85 and 0.75. It does outperform the usual ADF test for large absolute values of ρ, but it overrejects much more severely than the semiparametric bootstrap test.

The case I have examined here is particularly favourable to the bootstrap. An AR(1) process is very simple, and, in practice, the investigator will almost never be sure that the error terms follow such a process. Therefore, it seems very likely that the sieve bootstrap will perform less well than the semiparametric one did here, and there is certainly no reason to believe that the nonparametric bootstrap will perform any better. The poor performance of the block bootstrap in this case is consistent with theoretical results which suggest that the block bootstrap is likely to provide only modest improvements over asymptotic tests; see Härdle, Horowitz, and Kreiss (2001) for a review of the highly technical literature on this topic. Other references on bootstrapping time series include Li and Maddala (1996), Berkowitz and Kilian (2000), and van Giersbergen and Kiviet (2002). In the current state of the art, it appears that bootstrap tests should be used with caution for models in which the error terms display substantial serial correlation.

5.2. Models with heteroskedasticity

It is also challenging to make the bootstrap work well in models with heteroscedastic error terms when the form of the heteroscedasticity is unknown. In this situation, we must generate bootstrap samples in such a way that we retain the relationship between the variance of each error term and the corresponding regressors. Therefore, we cannot simply resample the residuals. Instead, two methods of generating bootstrap samples are widely used.

The simplest way to deal with heteroscedasticity, which was originally proposed by Freedman (1981), is called *bootstrapping pairs* or the *pairs bootstrap*. Consider the linear regression model

$$y_t = \boldsymbol{X}_t \boldsymbol{\beta} + u_t, \quad u_t = \sigma_t \epsilon_t, \quad \mathrm{E}(\epsilon_t^2) = 1, \tag{8}$$

where σ_t^2, the variance of the error terms, depends on the regressors in an unknown fashion. The idea of bootstrapping pairs is to resample the regressand and regressors together. Thus, the t^{th} row of each bootstrap regression is

$$y_t^* = \boldsymbol{X}_t^* \boldsymbol{\beta} + u_t^*, \tag{9}$$

where the row vector $[y_t^* \;\; \boldsymbol{X}_t^*]$ is equal to each of the row vectors $[y_s \;\; \boldsymbol{X}_s]$, for $s = 1, \ldots, n$, with probability $1/n$. In this way, we do not specify a parametric bootstrap DGP at all. Instead, we ensure that the bootstrap data are generated from the empirical distribution function of the real data. Since the regressor matrix will be different for each of the bootstrap samples, the pairs bootstrap does not make sense if the regressors are thought of as fixed in repeated samples.

When using the pairs bootstrap, we cannot impose a parametric null hypothesis on $\boldsymbol{\beta}$. In order to compute a bootstrap P value, we need to change the null hypothesis to one that is compatible with the data. If the hypothesis of interest is that β_1 equals some specified value, then we need to compare the actual statistic for testing this hypothesis with the distribution of the bootstrap statistics for the hypothesis that $\beta_1 = \hat{\beta}_1$. Bootstrap P values are then computed in the usual way.

An alternative way to deal with heteroscedasticity is to use what is called the *wild bootstrap*, which was proposed by Liu (1988) and further developed by Mammen (1993). Once again, consider the model (9). For testing restrictions on this model, the wild bootstrap DGP would be

$$y_t = \boldsymbol{X}_t \tilde{\boldsymbol{\beta}} + f(\tilde{u}_t) v_t, \tag{10}$$

where $\tilde{\boldsymbol{\beta}}$ denotes the OLS estimates subject to the restriction that is being tested, $f(\tilde{u}_t)$ is a transformation of the t^{th} residual $\tilde{u}_t$ associated with $\tilde{\boldsymbol{\beta}}$, and v_t is a random variable with mean 0 and variance 1.

A simple choice for the function $f(\cdot)$ is

$$f(\tilde{u}_t) = \frac{\tilde{u}_t}{(1 - h_t)^{1/2}},$$

which ensures that the $f(\tilde{u}_t)$ would have constant variance if the error terms were homoscedastic. We do not have to subtract the mean from $f(\tilde{u}_t)$, because the fact that v_t has mean 0 ensures that $f(\tilde{u}_t)v_t$ does so as well.

There are, in principle, many ways to specify the random variable v_t. By far the most popular is the two-point distribution,

$$F_1\colon \quad v_t = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}). \end{cases}$$

This distribution was suggested by Mammen (1993). A much simpler two-point distribution, called the *Rademacher distribution*, is

$$F_2\colon \quad v_t = \begin{cases} -1 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2. \end{cases}$$

Davidson and Flachaire (2001) have recently shown, on the basis of both theoretical analysis and simulation experiments, that wild bootstrap tests based on the Rademacher distribution F_2 will usually perform better, in finite samples, than those based on F_1.

In some respects, the error terms for the wild bootstrap DGP (10) do not resemble those of the true DGP (8) at all. When a two-point distribution is used, as it almost always is, the bootstrap error term can take on only two possible values for each observation. With F_2, these are just plus and minus $f(\tilde{u}_t)$. Nevertheless, the wild bootstrap does mimic the essential features of the true DGP well enough for it to be useful in many cases.

In order to investigate the performance of the pairs and wild bootstraps, I conducted a number of simulation experiments for the model

$$y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + u_t, \qquad u_t = \sigma_t \epsilon_t, \qquad \epsilon_t \sim N(0,1), \tag{11}$$

where both regressors were drawn randomly from the standard lognormal distribution, $\beta_0 = \beta_1 = 1$, $\beta_2 = 0$, and

$$\sigma_t = z(\gamma)(\beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2})^{\gamma}, \tag{12}$$

$z(\gamma)$ being a scaling factor chosen to ensure that the average variance of u_t is equal to 1. Thus changing γ changes the pattern of heteroscedasticity but does not, on average, change the variance of the error terms. This model was deliberately chosen to make heteroscedasticity-robust inference difficult. Because the regressors are lognormal, samples will often contain a few observations on the X_{tj} that are quite extreme, and the most extreme observation in each sample will tend to become more so as the sample size increases.

The most common way to test the hypothesis that $\beta_2 = 0$ in (11) is to estimate the model by ordinary least squares and calculate a heteroscedasticity-robust t statistic. This can be done in various ways. Following MacKinnon and White (1985), I divided the OLS estimate $\hat{\beta}_2$ by the square root of the appropriate diagonal element of the heteroscedasticity-robust covariance matrix

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{13}$$

where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with typical diagonal element $\hat{u}_t^2/(1 - h_t)$. Here, $\hat{u}_t$ is the t^{th} OLS residual, and h_t is the t^{th} diagonal element of the hat matrix $\boldsymbol{P}_{\boldsymbol{X}}$ for the unrestricted model. The heteroscedasticity-robust LM test proposed by Davidson and MacKinnon (1985) would almost certainly work better than the heteroscedasticity-robust t statistic that I have chosen to study. The latter is more commonly employed, however, and my objective here is not to find the best possible heteroscedasticity-robust test but to investigate the effect of bootstrapping.

The results of two sets of experiments are shown in figures 5 and 6. The solid lines show rejection frequencies for the asymptotic test. They are based on 500,000 replications for each of 41 values of γ between 0 and 2 at intervals of 0.05. The points show results for three different bootstrap tests. They are based on 100,000 replications with $B = 399$ for each of 17 values of γ at intervals of 0.125. The bootstrap tests use the wild bootstrap based on F_2 (bullets), the wild bootstrap based on F_1 (circles), and the pairs bootstrap (plus signs).

When $n = 50$, the asymptotic test overrejects for small values of γ and underrejects for large ones. The pairs bootstrap test does likewise. It always rejects less frequently than the asymptotic test, and it underrejects severely when γ is large. In contrast, both wild bootstrap tests always underreject. The underrejection is fairly modest for small values of γ, but it becomes much more severe as γ increases, especially for the test based on F_1.

When $n = 400$, the asymptotic test continues to overreject for small values of γ and underreject for large ones. As before, the pairs bootstrap test always rejects less frequently than the asymptotic test. Both wild bootstrap tests perform extremely well for small values of γ, with the F_1 version overrejecting slightly and the F_2 version underrejecting very slightly. For larger values of γ, both underreject. The

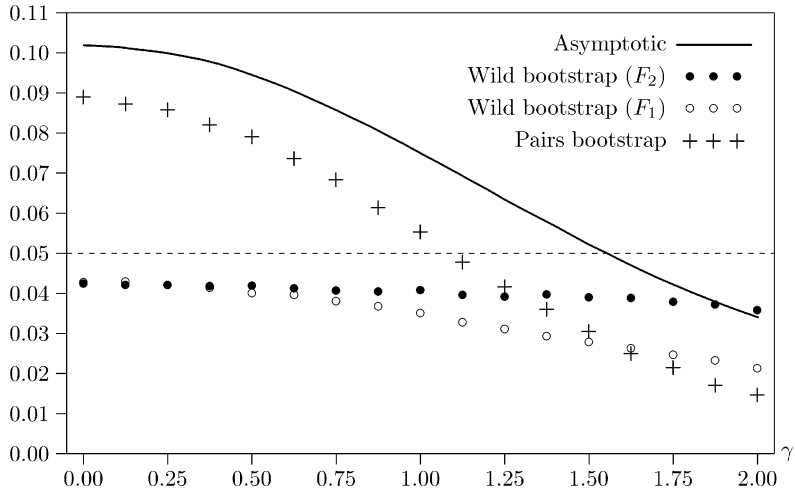

FIGURE 5 Rejection frequencies for heteroscedasticity-consistent tests, $n = 50$

test based on F_1 performs only moderately better than the asymptotic test for large values of γ, while the test based on F_2 performs very much better.

These results confirm the findings of Davidson and Flachaire (2001) and suggest that, if one is going to use the wild bootstrap, one should use the F_2 version of it. Figures 5 and 6 also suggest that bootstrapping pairs can be extremely unreliable, and that tests based on either version of the wild bootstrap may not be particularly reliable in samples of modest size. The sensitivity of the results to γ implies that the relative performance of the various tests may be highly model dependent. Nevertheless, for data sets of reasonable size, the F_2 version of the wild bootstrap does appear to be a promising technique.

5.3. Simultaneous equations models

Bootstrapping even one equation of a simultaneous equations model is a good deal more complicated than bootstrapping an equation in which all the explanatory variables are exogenous or predetermined. The problem is that the bootstrap DGP must provide a way to generate all of the endogenous variables, not just one of them. The class of model for which two-stage least squares is appropriate can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Y}\boldsymbol{\gamma} + \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{u} \\ \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\Pi} + \boldsymbol{V}, \end{aligned} \tag{14}$$

where $\boldsymbol{y}$ is a vector of observations on an endogenous variable of particular interest, $\boldsymbol{Y}$ is a matrix of observations on other exogenous variables, $\boldsymbol{X}$ is a matrix of observations on exogenous or predetermined variables, and $\boldsymbol{X}_1$ consists of some of the columns of $\boldsymbol{X}$. The first equation of (14) can be estimated consistently by two-stage

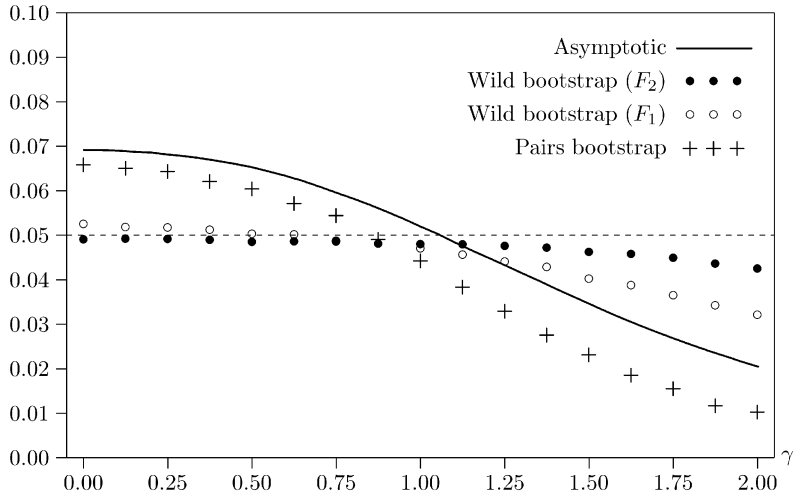

FIGURE 6 Rejection frequencies for heteroscedasticity-consistent tests, $n = 400$

least squares, but 2SLS estimates are usually biased in finite samples. They can be seriously misleading, even when the sample size is large, if some of the reduced form equations for $\boldsymbol{Y}$ have little explanatory power; see Staiger and Stock (1997), among many others.

In order to bootstrap the 2SLS estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we need to generate bootstrap samples containing both $\boldsymbol{y}^*$ and $\boldsymbol{Y}^*$. For a semiparametric bootstrap, we need estimates of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\Pi}$. These would normally be 2SLS estimates of the parameters of the first (structural) equation and OLS estimates of the parameters of the remaining (reduced form) equations. We can then obtain the bootstrap error terms by resampling rows of the residual matrix $[\hat{\boldsymbol{u}} \;\; \hat{\boldsymbol{V}}]$, perhaps after rescaling. Alternatively, we could assume normality and use a fully parametric bootstrap.

A simpler approach, which also allows for heteroscedasticity, is to use the pairs bootstrap; this was proposed by Freedman and Peters (1984). As we have seen, however, this approach is less than ideal for testing hypotheses, and it can be expected to work even less well than the semiparametric approach when the error terms are homoscedastic.

The finite-sample distributions of the 2SLS estimates are quite sensitive to some of the parameters that appear in the bootstrap DGP. Thus, although bootstrapping may work well in some cases, it would be unrealistic to expect it to work well all the time. As an illustration, I generated data from a special case of (14). The model was

$$\begin{aligned} y_t &= \beta + \gamma Y_t + u_t \\ Y_t &= \boldsymbol{X}_t \boldsymbol{\pi} + v_t, \end{aligned} \tag{15}$$

where all coefficients were equal to 1, $\boldsymbol{X}_t$ consisted of a constant and three independent standard normal random variables, and the error terms were jointly nor-

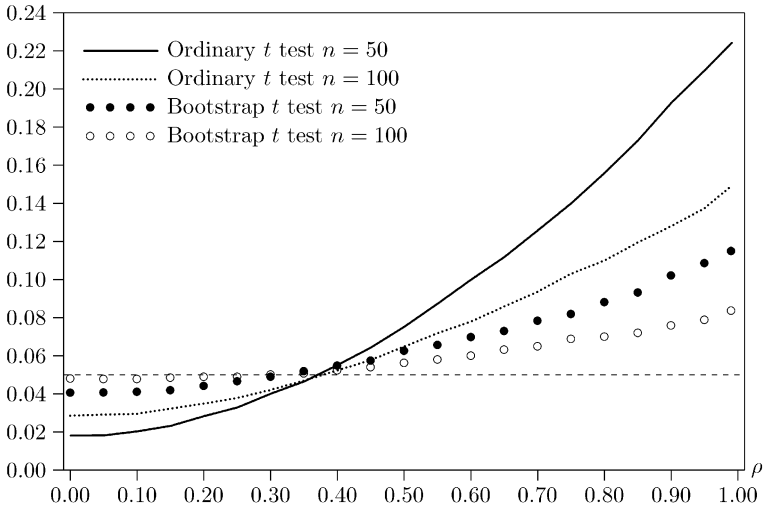

FIGURE 7 Rejection frequencies for t tests in 2SLS regression

mally distributed with $\mathrm{Var}(u_t) = 1$, $\mathrm{Var}(v_t) = 16$, and correlation ρ. The first equation of (15) was estimated by 2SLS, and the ordinary t statistic was used to test the true null hypothesis that $\gamma = 1$. Bootstrap samples were generated by a semiparametric bootstrap procedure that used estimates of β and $\boldsymbol{\pi}$ under the null and obtained the error terms by resampling pairs of rescaled residuals.

Figure 7 shows the results of a few experiments designed to illustrate how rejection frequencies vary with ρ. These are based on 100,000 replications, with $B = 399$, and sample sizes of 50 or 100. We see that both tests underreject for small values of ρ and overreject for larger ones. The bootstrap test performs far from perfectly, although it almost always performs better than the ordinary t test, and it seems to perform relatively better for the larger sample size. These results are, of course, very sensitive to the other parameters of the model. In particular, both tests would work much better if $\mathrm{Var}(v_t)$ were smaller and the second equation therefore fit better.

There is no doubt that the bootstrap can be useful for multivariate models. For example, Rilstone and Veall (1996) provide encouraging evidence on the performance of certain bootstrap procedures in the context of seemingly unrelated regressions, and Inoue and Kilian (2002) do so in the context of vector autoregressions. But in neither case, and even less so for simultaneous equations models, should we expect the sort of astonishingly good performance that was observed in figure 3.

6. The power of bootstrap tests

The probability that a test will reject the null hypothesis when some alternative is true is called its *power*. Economists have traditionally paid surprisingly little atten-

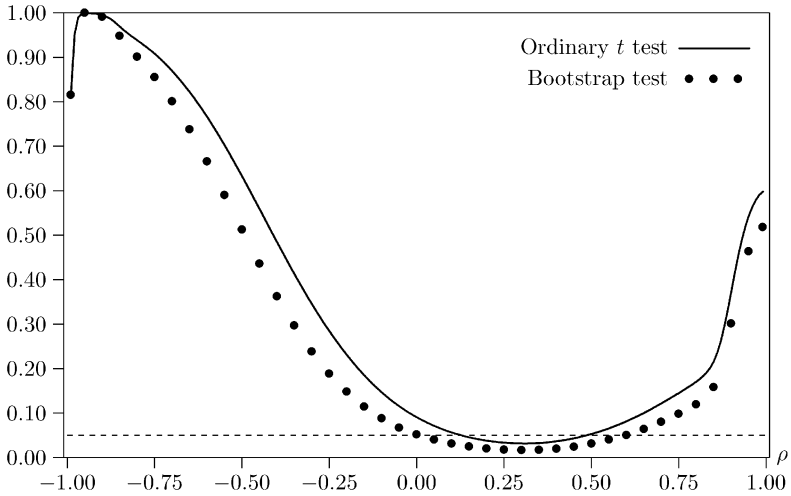

FIGURE 8 Power of Durbin-Godfrey test, $n = 20$, $\delta = 0.9$

tion to power, even though there is not much point in performing a hypothesis test if it does not have reasonably high power when the null hypothesis is violated to an economically meaningful extent.

It is natural to worry that bootstrapping a test will reduce its power. This can certainly happen. Indeed, some loss of power is inevitable whenever B, the number of bootstrap samples, is finite. More important, if an asymptotic test overrejects under the null, a bootstrap test based on it will reject less often both under the null and under many alternatives. Conversely, if an asymptotic test underrejects under the null, a bootstrap test based on it will reject more often. There is no reason to believe that bootstrapping a test, using a large value of B, will reduce its power more substantially than will any other method of improving its finite-sample properties under the null.

The relationship between the power of bootstrap and asymptotic tests is studied in Davidson and MacKinnon (2001). It is shown that, if the power of an asymptotic test is adjusted in a plausible way to account for its tendency to overreject or underreject under the null hypothesis, then the resulting 'level-adjusted' power is very similar to the power of a bootstrap test based on the same underlying test statistic. Thus, if bootstrapping does result in a loss of power when B is large, that loss arises simply because bootstrapping corrects the tendency of the asymptotic test to overreject.

As an illustration, consider once again the Durbin-Godfrey test for serial correlation in the linear regression model (5). Figure 8 graphs the power of this test, at the .05 level, as a function of ρ, for the case in which $\sigma = 0.1$, $n = 20$, and $\delta = 0.90$. This was a case in which the test overrejected quite severely under the null; see figure 3. The power of the asymptotic test, shown as the solid line, is based

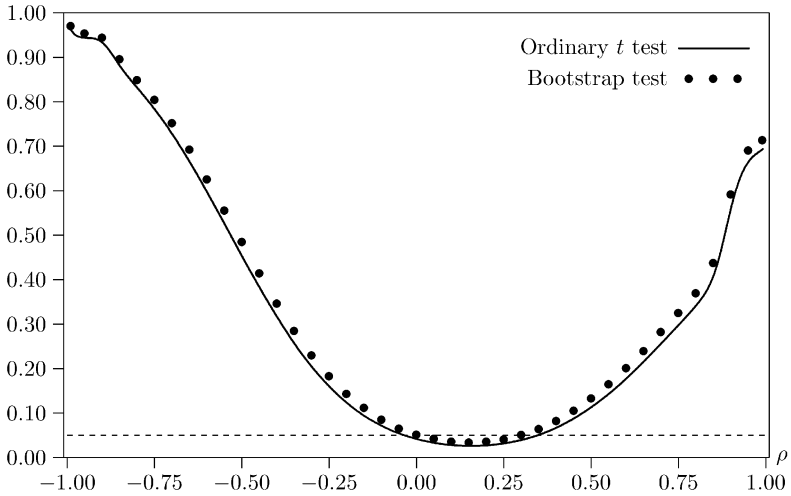

FIGURE 9 Power of Durbin-Godfrey test, $n = 20$, $\delta = -0.9$

on 500,000 replications for each of 199 values of ρ between -0.99 and 0.99. The power of the bootstrap test, shown as bullets for 41 values of ρ ($-0.99, -0.95, \ldots, 0.95, 0.99$), is based on only 100,000 replications, with $B = 399$.

Figure 8 contains a number of striking results. Contrary to what asymptotic theory suggests, for neither test does the power function achieve its minimum at $\rho = 0$ or increase monotonically as $|\rho|$ increases. Instead, power actually declines sharply as ρ approaches -1. Moreover, the asymptotic test has less power for values of ρ between 0 and about 0.62 than it does for $\rho = 0$. In consequence, the bootstrap test actually rejects less than 5 per cent of the time for values between 0 and about 0.61. Thus, in this particular case, the Durbin-Godfrey test is essentially useless for detecting positive serial correlation of the magnitude that we are typically concerned about.

Because the asymptotic test actually rejects 9.01 per cent of the time at the 5 per cent level, the bootstrap test almost always rejects less frequently than the asymptotic one. The magnitude of this 'power loss' depends on ρ. It is largest for values between about -0.25 and -0.65, and it is quite small for very large negative values of ρ.

The results in figure 8 are highly dependent on the sample size, the parameter values, and the way in which the regressors are generated. To illustrate this, figure 9 shows that changing δ from 0.9 to -0.9 has a substantial effect on the shape of the power functions. There is now a much smaller region in which the bootstrap test rejects less than 5 per cent of the time, and there is no drop in power as ρ approaches -1. Moreover, because the asymptotic test rejects just 4.16 per cent of the time at the 5 per cent level, the bootstrap test always rejects more often than the ordinary t test on which it is based.

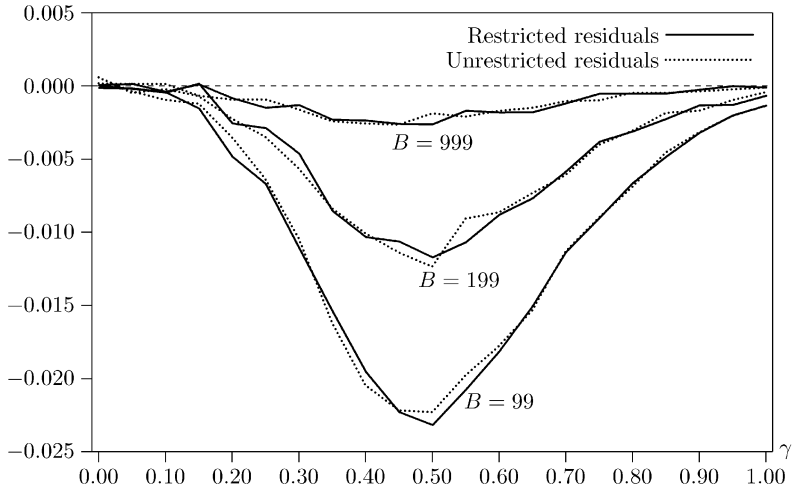

FIGURE 10 Power loss from bootstrapping

As I mentioned in section 2, the power of simulation-based tests generally increases with B. As Davidson and MacKinnon (2000) discuss, however, any loss of power is generally quite modest, except perhaps when B is a very small number. To illustrate this power loss, I conducted yet another simulation experiment. The null was a linear regression model with a constant term and two other regressors, and the alternative was the same model with nine additional regressors. The test statistic was the ordinary F statistic for the coefficients on the additional regressors to be zero. The error terms were normally distributed, and there were 20 observations. Therefore, under the null hypothesis, the F statistic actually followed the F distribution with 9 and 8 degrees of freedom.

I also performed two varieties of bootstrap test, using resampled residuals rescaled according to the formula (4), and then computing bootstrap P values in the usual way. One type of test used the residuals from the restricted model, together with the diagonals of the hat matrix for that model, and the other used the residuals and hat matrix from the unrestricted model. If the argument of van Giersbergen and Kiviet (2002) is correct, the bootstrap test that uses unrestricted residuals should be more powerful than the one that uses restricted residuals. To investigate this argument, I set the coefficients on all the additional regressors to γ and calculated the power of the F and bootstrap tests for various values of γ.

Figure 10 shows the difference between the power of the F test and the power of various bootstrap tests as γ varies between 0 and 1. The null hypothesis is true when $\gamma = 0$. The effects of experimental randomness are visible in the figure, since there were only 100,000 replications. Except for very small values of γ, there is always some loss of power from bootstrapping. In the case of $B = 999$, this power loss is very small, never exceeding 0.00264. Since it is roughly proportional to $1/B$, however, it is about ten times as large for $B = 99$. The power loss is greatest for inter-

mediate values of γ. When γ is small, even the F test has little power, so there is not much power to be lost. When γ is large, the evidence against the null is so strong that the additional randomness introduced by simulation reduces power only slightly.

It is interesting that there is no systematic tendency for the bootstrap test based on the unrestricted residuals to be more powerful than the one based on the restricted residuals. Depending on γ, either of the tests may be very slightly more powerful than the other. I also tried using (3) instead of (4) to rescale the residuals. This affected test power about as much as switching from restricted to unrestricted residuals. These experiments suggest that there is no reason not to use residuals from the restricted model when generating bootstrap samples.

7. Bootstrap confidence intervals

The statistical literature has put far more emphasis on using the bootstrap to construct confidence intervals than on using it to test hypotheses. In my view, this is somewhat unfortunate, for two reasons. The first reason is that there are many more ways to construct bootstrap confidence intervals than there are to perform bootstrap tests. Given a procedure for generating the bootstrap data, it is very straightforward to compute a bootstrap P value or a bootstrap critical value. In contrast, there are generally many alternative ways to compute bootstrap confidence intervals, and they may yield quite different results. Thus, the literature on bootstrap confidence intervals easily can be confusing. The second reason is that, for a given model and dataset, bootstrap tests generally tend to be more reliable than bootstrap confidence intervals.

The most important thing to understand about confidence intervals is that, in principle, they can always be obtained by 'inverting' a suitable test statistic. A confidence interval for a parameter θ is simply the set of values of θ_0 for which the hypothesis that $\theta = \theta_0$ is not rejected. The confidence intervals that we are most familiar with are obtained by inverting t statistics, with critical values based on either the Student's t or the standard normal distribution. One of the most popular bootstrap confidence intervals is obtained in exactly the same way, but with critical values that are quantiles of a distribution of bootstrap t statistics. It is called the *bootstrap t* or *percentile t* confidence interval.

Suppose that $\hat{\theta}$, which has standard error s_θ, is an estimate of the parameter θ in which we are interested. Then, a t statistic for the hypothesis that $\theta = \theta_0$ is

$$t(\theta_0) = \frac{\hat{\theta} - \theta_0}{s_\theta}. \tag{16}$$

Under quite weak conditions, asymptotically, this statistic follows the standard normal distribution. Under very much stronger conditions, in finite samples, it follows the Student's t distribution with a known number of degrees of freedom.

Let $1 - \alpha$ be the level of the confidence interval we are trying to construct. We can find such an interval by inverting the t statistic (16). The two ends of the interval are the values of θ_0 that solve the equations

$$\frac{\hat{\theta} - \theta_0}{s_\theta} = t_{\alpha/2} \quad \text{and} \quad \frac{\hat{\theta} - \theta_0}{s_\theta} = t_{1-\alpha/2}, \tag{17}$$

where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution that $t(\theta_0)$ is assumed to follow. For a .95 confidence interval based on the standard normal distribution, $t_{\alpha/2} = -1.96$ and $t_{1-\alpha/2} = 1.96$. The interval based on (17) is

$$[\hat{\theta} - s_\theta t_{1-\alpha/2},\ \hat{\theta} - s_\theta t_{\alpha/2}]. \tag{18}$$

Notice that the lower limit of this interval depends on the upper-tail critical value $t_{1-\alpha/2}$, and the upper limit depends on the lower-tail critical value $t_{\alpha/2}$. This may seem strange, but after enough reflection it can be seen to make sense.

Because the standard normal and Student's t distributions are symmetric around the origin, $t_{\alpha/2} = -t_{1-\alpha/2}$. Therefore, the interval (18) can also be written as

$$[\hat{\theta} - s_\theta t_{1-\alpha/2},\ \hat{\theta} + s_\theta t_{1-\alpha/2}]. \tag{19}$$

This form is more familiar than (18), but it is valid only for symmetric distributions. In the familiar case in which $\alpha = .05$ and the standard normal distribution is used, the interval (19) has endpoints at $\hat{\theta}$ plus and minus 1.96 standard errors.

A bootstrap t confidence interval is constructed in very much the same way as the interval (18). The only difference is that the quantiles of the theoretical distribution are replaced by quantiles of a bootstrap distribution. The steps required are as follows:

1. Estimate the model without restrictions to compute $\hat{\theta}$, s_θ, and whatever other quantities are needed to generate the bootstrap samples.
2. Choose B such that $\alpha(B + 1)/2$ is an integer, and generate B bootstrap samples. For each bootstrap sample, estimate the unrestricted model to obtain $\hat{\theta}_j^*$, and calculate the bootstrap test statistic

$$t_j^* \equiv \frac{\hat{\theta}_j^* - \hat{\theta}}{s_j^*}. \tag{20}$$

   This requires calculating a standard error s_j^* for each bootstrap sample. It would be a very bad idea to replace s_j^* by s_θ, because the t_j^* would no longer be asymptotically pivotal. Notice that $\hat{\theta}$ is playing the role of θ_0 in expression (20), because the bootstrap data are not constrained to satisfy a null hypothesis.
3. Find $t_{\alpha/2}^*$ and $t_{1-\alpha/2}^*$, the $\alpha/2$ and $1 - \alpha/2$ quantiles of the t_j^*. These are simply the values numbered $(\alpha/2)(B + 1)$ and $(1 - \alpha/2)(B + 1)$ in the list of the t_j^* sorted from smallest to largest.
4. Calculate the bootstrap t interval as

$$[\hat{\theta} - s_\theta t_{1-\alpha/2}^*,\ \hat{\theta} - s_\theta t_{\alpha/2}^*]. \tag{21}$$

   Notice that, unless the distribution of the t_j^* happens to be symmetric around the origin, this will not be a symmetric interval.

As we will see in a moment, the bootstrap t interval is far from perfect. Nevertheless, it is attractive for at least three reasons:

- The way in which we construct it is very similar to the way in which we construct more familiar confidence intervals like (18).
- It has excellent theoretical properties. Provided the test statistic (16) is asymptotically pivotal, it can be shown that a bootstrap t interval based on it will be asymptotically valid. Moreover, its accuracy will increase more rapidly, as the sample size increases, than that of the standard interval (18) when the latter is valid only asymptotically. See Hall (1992) for a detailed discussion and proof. In the unlikely event that the t statistic (16) is exactly pivotal, a bootstrap t interval based on it will be exact.
- When $\hat{\theta}$ is biased, the bootstrap t interval tends to correct the bias. Suppose, for concreteness, that $\mathrm{E}(\hat{\theta}) < \theta$. Then we would expect $t^*_{\alpha/2}$ to be a larger negative number than $t_{\alpha/2}$, and $t^*_{1-\alpha/2}$ to be a smaller positive number than $t_{1-\alpha/2}$. If so, both limits of the interval (21) will be larger than the corresponding limits of the asymptotic interval (18), which is exactly what we want when $\hat{\theta}$ is biased downwards.

Unfortunately, the actual performance of bootstrap t intervals in finite samples is often not as good as theory suggests. These intervals generally work well if the test statistic on which they are based, expression (16), is approximately pivotal. When this is not the case, however, the distribution of the t^*_j may differ substantially from the distribution of $t(\theta_0)$, and the interval (21) may be quite inaccurate.

There are a great many other ways to construct bootstrap confidence intervals. One very widely applicable approach is to calculate the *bootstrap standard error*

$$s^*_\theta = \left(\frac{1}{B-1} \sum_{j=1}^{B} (\hat{\theta}^*_j - \bar{\theta}^*)^2 \right)^{1/2}, \quad \text{where} \quad \bar{\theta}^* = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}^*_j, \tag{22}$$

which is simply the standard deviation of the bootstrap estimates $\hat{\theta}^*_j$. We can then construct the *bias-corrected bootstrap interval*

$$[2\hat{\theta} - \bar{\theta}^* - s^*_\theta t_{1-\alpha/2}, \qquad 2\hat{\theta} - \bar{\theta}^* + s^*_\theta t_{1-\alpha/2}]. \tag{23}$$

This interval is quite similar to (19), but it is centred on the bias-corrected estimate $2\hat{\theta} - \bar{\theta}^*$, and it uses the bootstrap standard error s^*_θ instead of s_θ. The bias-corrected estimate used in (23) is obtained by subtracting the estimated bias $\bar{\theta}^* - \hat{\theta}$ from $\hat{\theta}$. The bootstrap is commonly used for bias correction in this way; see MacKinnon and Smith (1998), which also discusses more sophisticated types of bias correction. In theory, the interval (23) should generally not work as well as the bootstrap t interval (21), but it may actually work better when s_θ is unreliable, and it can be used in situations where s_θ cannot be computed at all. Of course, when bias is not a problem, we can use an interval similar to (23) that is centred at $\hat{\theta}$ rather than at $2\hat{\theta} - \bar{\theta}^*$.

Even for bootstrap t intervals, quite a few variants are available. For example, there may often be several plausible ways to compute standard errors, each of which

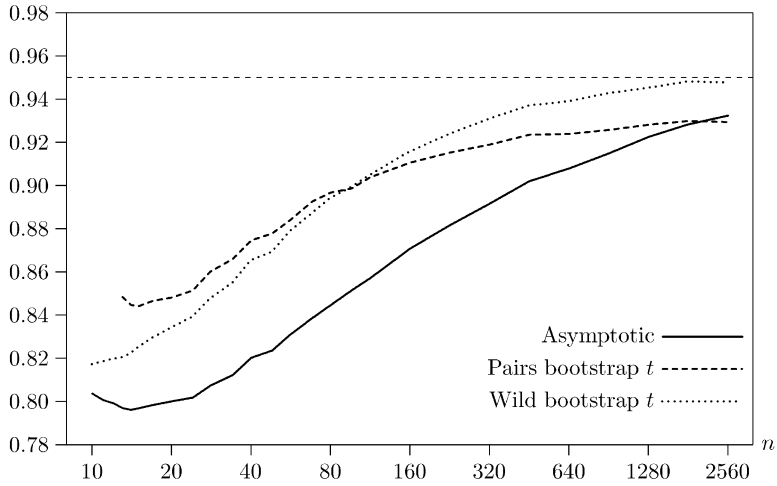

FIGURE 11 Coverage of .95 confidence intervals for heteroscedastic regression

will lead to a different confidence interval. It may also be possible to transform the parameter(s) and then use an interval constructed in terms of the transformed parameters to obtain an interval for θ. Another possibility, if we believe that the distribution of $t(\theta_0)$ is symmetric around the origin, is to estimate the $1 - \alpha$ quantile of the absolute values of the t_j^* and use it to define both ends of the interval.

Because neither bootstrap t intervals nor intervals based on bootstrap standard errors (with or without bias correction) always perform well, many other types of bootstrap confidence intervals have been proposed. For reasons of space, I will not discuss any of these. See Hall (1992), Efron and Tibshirani (1993), DiCiccio and Efron (1996), Davison and Hinkley (1997), Hansen (1999), Davidson (2000), Horowitz (2001), and van Giersbergen and Kiviet (2002), among many others.

To illustrate the performance of bootstrap confidence intervals, I performed two sets of simulation experiments. The first involved a regression model with heteroscedasticity of unknown form. The data were generated by a variant of the model given by (11) and (12), with all the β_j equal to 1 and $\gamma = 1$. Asymptotic confidence intervals were based on the heteroscedasticity-robust covariance matrix estimator (13). I examined two different bootstrap t confidence intervals, both of which also used standard errors based on (13). One used the pairs bootstrap, and the other used the F_2 version of the wild bootstrap.

The proportion of the time that a confidence interval includes or covers the true parameter value is called the *coverage* of the interval. The coverage of three .95 intervals, one asymptotic and two bootstrap t, is shown in figure 11 for a large number of sample sizes, ranging from 10 to 2560. These results are based on 100,000 replications, with 399 bootstraps for each sample size. For the pairs bootstrap, the smallest sample size is 13 rather than 10, because, for smaller sample sizes, the resampled $\boldsymbol{X}^*$ matrix was sometimes singular.

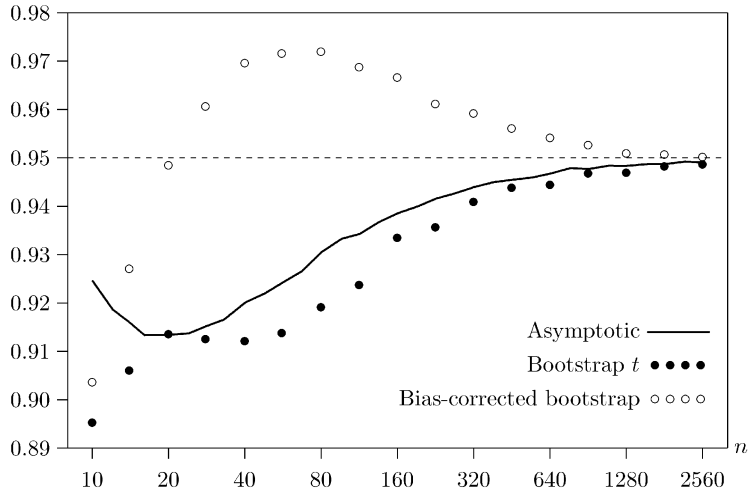

FIGURE 12 Coverage of .95 confidence intervals for AR(1) model

It is evident from figure 11 that all three intervals undercover quite severely for the smaller sample sizes. Both bootstrap *t* intervals always outperform the asymptotic interval, except for the pairs bootstrap for $n = 2560$, but neither performs particularly well in samples of moderate size. It is interesting to note that the pairs bootstrap interval is the most accurate of the three for small sample sizes and the least accurate of them for the largest sample size. For sample sizes greater than about 100, the wild bootstrap interval always performs the best.

The second experiment concerned the autoregressive model

$$y_t = \beta + \rho y_{t-1} + u_t, \qquad u_t \sim \text{IID}(0, \sigma^2),$$

where $\beta = 0$, $\rho = 0.9$, and $\sigma = 0.1$. This model can readily be estimated by ordinary least squares, and the asymptotic confidence interval (19) for ρ at the .95 level can be constructed using the OLS standard error on $\hat{\rho}$. For the bootstrap, I generated data recursively using the OLS parameter estimates, resampling the rescaled residuals. I then constructed bootstrap *t* intervals based on (21) as well as bias-corrected bootstrap intervals based on (23). It may be appropriate to use the latter in this case, because $\hat{\rho}$ will be biased downwards.

The coverage of asymptotic and bootstrap confidence intervals at the .95 level for a number of sample sizes between 10 and 2560 are shown in figure 12. The simulations used 500,000 replications for the asymptotic intervals and 100,000 replications, with $B = 399$, for the bootstrap ones. Surprisingly, the asymptotic intervals always perform better than the bootstrap *t* intervals. This happens despite the fact that $\hat{\rho}$ is biased towards 0. What seems to be happening is that the asymptotic intervals undercover severely at the upper end but overcover at the lower end, while the bootstrap *t* intervals (which tend to be considerably shorter for the smaller

sample sizes) undercover at both ends. The bias-corrected bootstrap intervals perform quite differently from both the asymptotic or bootstrap *t* intervals. Unlike the latter, they tend to overcover, except for quite small sample sizes. Note that more sophisticated bootstrap confidence intervals, such as those proposed by Hansen (1999) and Davidson (2000), can be expected to perform better in this case than either of the bootstrap intervals considered here.

8. Conclusion

We have seen that, in many cases, using simulated distributions to perform tests and construct confidence intervals is conceptually simple. This approach often yields inferences that are substantially more accurate than those based on asymptotic theory, and it rarely yields inferences that are substantially less accurate. The extraordinarily rapid development of computing technology over the past decade means that the computational costs of using the bootstrap and other simulation-based procedures are often negligible. Thus, it is not surprising that these procedures have become a standard part of the applied econometrician's toolkit.

When the rather stringent conditions needed for a test statistic to be pivotal are satisfied, a Monte Carlo test will always be exact. Under very much weaker conditions, bootstrap tests will be asymptotically valid. Under the relatively weak conditions needed for a test statistic to be asymptotically pivotal, for large enough sample sizes bootstrap tests should perform better than asymptotic tests. However, they cannot be relied upon to perform well in all cases. This is especially true when there is either serial correlation or heteroscedasticity of unknown form. There are many ways to construct bootstrap confidence intervals. These may or may not be more reliable than intervals based on asymptotic theory, and their performance can sometimes be far from satisfactory.

For most econometric models, there is more than one reasonable way to generate bootstrap samples. In any given situation, some of the applicable methods will undoubtedly work better than others. Especially when the residuals show evidence of heteroscedasticity or serial correlation of unknown form, the number of bootstrapping procedures that can be used is generally very great. At the present time, however, we often do not know which of them, if any, can be relied upon to yield reliable inferences in samples of the size typically encountered in applied work.

References

Albert, J., and S. Chib (1993) 'Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts,' *Journal of Business and Economic Statistics* 11, 1–15

Berkowitz, J., and L. Kilian (2000) 'Recent developments in bootstrapping time series,' *Econometric Reviews* 19, 1–48

Beran, R. (1988) 'Prepivoting test statistics: a bootstrap view of asymptotic refinements,' *Journal of the American Statistical Association* 83, 687–97

Bühlmann, P. (1997) 'Sieve bootstrap for time series,' *Bernoulli* 3, 123–48
— (1998) 'Sieve bootstrap for smoothing nonstationary time series,' *Annals of Statistics* 26, 48–83
Carlstein, E. (1986) 'The use of subseries methods for estimating the variance of a general statistic from a stationary time series,' *Annals of Statistics* 14, 1171–9
Chang, Y., and J.Y. Park (2002) 'A sieve bootstrap for the test of a unit root,' *Journal of Time Series Analysis* 23, forthcoming
Choi, E., and P. Hall (2000) 'Bootstrap confidence regions computed from autoregressions of arbitrary order,' *Journal of the Royal Statistical Society*, Series B, 62, 461–77
Davidson, R. (2000) 'Comment on "Recent developments in bootstrapping time series,"' *Econometric Reviews* 19, 49–54
Davidson, R., and E. Flachaire (2001) 'The wild bootstrap, tamed at last,' GREQAM Document de Travail 99A32, revised
Davidson, R., and J.G. MacKinnon (1981) 'Several tests for model specification in the presence of alternative hypotheses,' *Econometrica* 49, 781–93
— (1985) 'Heteroskedasticity-robust tests in regression directions,' *Annales de l'INSÉE* 59/60, 183–218
— (1992) 'A new form of the information matrix test,' *Econometrica* 60, 145–57
— (1993) *Estimation and Inference in Econometrics* (New York: Oxford University Press)
— (1999a) 'The size distortion of bootstrap tests,' *Econometric Theory* 15, 361–76
— (1999b) 'Bootstrap testing in nonlinear models,' *International Economic Review* 40, 487–508
— (2000) 'Bootstrap tests: How many bootstraps?' *Econometric Reviews* 19, 55–68
— (2001) 'The power of bootstrap and asymptotic tests,' unpublished paper
— (2002a) 'Bootstrap *J* tests of nonnested linear regression models,' *Journal of Econometrics* 109, 167–93
— (2002b) 'Fast double bootstrap tests of nonnested linear regression models,' *Econometric Reviews*, forthcoming
Davison, A.C., and D.V. Hinkley (1997) *Bootstrap Methods and Their Application* (Cambridge: Cambridge University Press)
DiCiccio, T.J., and D. Efron (1996) 'Bootstrap confidence intervals' (with discussion), *Statistical Science* 11, 189–228
Dufour, J.-M., and L. Khalaf (2001) 'Monte Carlo test methods in econometrics,' in *A Companion to Econometric Theory*, ed. B. Baltagi (Oxford: Blackwell Publishers), 494–519
Durbin, J. (1970) 'Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables,' *Econometrica* 38, 410–21
Durbin, J., and G.S. Watson (1950) 'Testing for serial correlation in least squares regression I,' *Biometrika* 37, 409–28
— (1951) 'Testing for serial correlation in least squares regression II,' *Biometrika* 38, 159–78
Dwass, M. (1957) 'Modified randomization tests for nonparametric hypotheses,' *Annals of Mathematical Statistics* 28, 181–7
Eckstein, Z., and K. Wolpin (1989) 'The specification and estimation of dynamic, stochastic discrete choice models: a survey,' *Journal of Human Resources* 24, 562–98
Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans* (Philadelphia: Society for Industrial and Applied Mathematics)
Efron, B., and R.J. Tibshirani (1993) *An Introduction to the Bootstrap* (New York: Chapman and Hall)
Elerian, O., S. Chib, and N. Shephard (2001) 'Likelihood inference for discretely observed nonlinear diffusions,' *Econometrica* 69, 959–93
Freedman, D.A. (1981) 'Bootstrapping regression models,' *Annals of Statistics* 9, 1218–28
Freedman, D.A., and S.C. Peters (1984) 'Bootstrapping an econometric model: some empirical results,' *Journal of Business and Economic Statistics* 2, 150–8

Galbraith, J.W., and V. Zinde-Walsh (1999) 'On the distributions of Augmented Dickey-Fuller statistics in processes with moving average components,' *Journal of Econometrics* 93, 25–47

Geweke, J. (1999) 'Using simulation methods for Bayesian econometric models: inference, development and communication' (with discussion and reply), *Econometric Reviews* 18, 1–126

van Giersbergen, N.P.A., and J.F. Kiviet (2002) 'How to implement the bootstrap in static or stable dynamic regression models: test statistic versus confidence interval approach,' *Journal of Econometrics* 108, 133–56

Godfrey, L.G. (1978) 'Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables,' *Econometrica* 46, 1293–301

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

Hall, P., and D.M. Titterington (1989) 'The effect of simulation order on level accuracy and power of Monte-Carlo tests,' *Journal of the Royal Statistical Society*, Series B, 51, 459–67

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

Härdle, W., J.L. Horowitz, and J.-P. Kreiss (2001) 'Bootstrap methods for time series,' paper presented at the 2001 European Meeting of the Econometric Society, Lausanne

Horowitz, J.L. (2001) 'The bootstrap,' in *Handbook of Econometrics*, Vol. 5, ed. J.J. Heckman and E.E. Leamer (Amsterdam: North-Holland)

Inoue, A., and L. Kilian (2002) 'Bootstrapping smooth functions of slope parameters and innovation variances in VAR(∞) models,' *International Economic Review* 43, 309–31

Kreiss, J.-P. (1992) 'Bootstrap procedures for AR(∞) processes,' in *Bootstrapping and Related Techniques*, ed. K.H. Jöckel, G. Rothe, and W. Sender, (Heidelberg: Springer-Verlag), 107–13

Künsch, H.R. (1989) 'The jackknife and the bootstrap for general stationary observations,' *Annals of Statistics* 17, 1217–41

Lahiri, S.N. (1999) 'Theoretical comparisons of block bootstrap methods,' *Annals of Statistics* 27, 386–404

Li, H., and G.S. Maddala (1996) 'Bootstrapping time series models' (with discussion), *Econometric Reviews* 15, 115–95

Liu, R.Y. (1988) 'Bootstrap procedures under some non-I.I.D. models,' *Annals of Statistics* 16, 1696–708

MacKinnon, J.G. (1996) 'Numerical distribution functions for unit root and cointegration tests,' *Journal of Applied Econometrics* 11, 601–18

MacKinnon, J.G., and A.A. Smith, Jr (1998) 'Approximate bias correction in econometrics,' *Journal of Econometrics* 85, 205–30

MacKinnon, J.G., and H. White (1985) 'Some heteroscedasticity consistent covariance matrix estimators with improved finite sample properties,' *Journal of Econometrics* 29, 305–25

Mammen, E. (1993) 'Bootstrap and wild bootstrap for high dimensional linear models,' *Annals of Statistics* 21, 255–85

McCulloch, R., and P. Rossi (1994) 'An exact likelihood analysis of the multinomial probit model,' *Journal of Econometrics* 64, 207–40

Ng, S., and P. Perron (1995) 'Unit root tests in ARMA Models with data dependent methods for the selection of the truncation lag,' *Journal of the American Statistical Association* 90, 268–81

— (2001) 'Lag length selection and the construction of unit root tests with good size and power,' *Econometrica* 69, 1519–54

Pakes, A.S. (1986) 'Patents as options: Some estimates of the value of holding European patent stocks,' *Econometrica* 54, 755–84
Park, J.Y. (2002) 'An invariance principle for sieve bootstrap in time series,' *Econometric Theory* 18, 469–90
Politis, D.N., and J.P. Romano (1994) 'Large sample confidence regions based on subsamples under minimal assumptions,' *Journal of the American Statistical Association* 22, 2031–50
Rilstone, P., and M.R. Veall (1996) 'Using bootstrapped confidence intervals for improved inferences with seemingly unrelated regression equations,' *Econometric Theory* 12, 569–80
Rust, J. (1987) 'Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher,' *Econometrica* 55, 999-1033
Schwert, G.W. (1989) 'Testing for unit roots: a Monte Carlo investigation,' *Journal of Business and Economic Statistics* 7, 147–59
Staiger, D., and J.H. Stock (1997) 'Instrumental variables regressions with weak instruments,' *Econometrica* 65, 557–86.
Stern, S. (1997) 'Simulation-based estimation,' *Journal of Economic Literature* 35, 2006–39