Queen’s Economics Department Working Paper No. 1127

Bootstrap Hypothesis Testing

James MacKinnon
Queen’s University

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

6-2007

Bootstrap Hypothesis Testing

James G. MacKinnon

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

jgm@econ.queensu.ca
`http://www.econ.queensu.ca/faculty/mackinnon/`

Abstract

This paper surveys bootstrap and Monte Carlo methods for testing hypotheses in econometrics. Several different ways of computing bootstrap P values are discussed, including the double bootstrap and the fast double bootstrap. It is emphasized that there are many different procedures for generating bootstrap samples for regression models and other types of model. As an illustration, a simulation experiment examines the performance of several methods of bootstrapping the supF test for structural change with an unknown break point.

June 2007; minor revisions, November 2008 and June 2009

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. The computations were performed on a Linux cluster funded by a grant from the Canada Foundation for Innovation, for which I am very grateful to Russell Davidson. I would like to thank John Nankervis and two anonymous referees for helpful comments. An earlier version was presented at the Conference on Computational and Financial Econometrics at the University of Geneva in April 2007.

1. Introduction

The basic idea of any sort of hypothesis test is to compare the observed value of a test statistic, say $\hat{\tau}$, with the distribution that it would follow if the null hypothesis were true. The null is then rejected if $\hat{\tau}$ is sufficiently extreme relative to this distribution. In certain special cases, such as t and F tests on the coefficients of a linear regression model with exogenous regressors and normal errors, this distribution is known, and we can perform exact tests. In most cases of interest to econometricians, however, the distribution of the test statistic we use is not known. We therefore have to compare $\hat{\tau}$ with a distribution that is only approximately correct. In consequence, the test may overreject or underreject.

Traditionally, the approximations that we use in econometrics have been based on asymptotic theory. But advances in computing have made an alternative approach increasingly attractive. This approach is to generate a large number of simulated values of the test statistic and compare $\hat{\tau}$ with the empirical distribution function, or EDF, of the simulated ones. Using the term "bootstrap" in a rather broad sense, I will refer to this approach as **bootstrap testing**, although the term **simulation-based testing** is more general and might perhaps be more accurate. Bootstrap testing can work very well indeed in some cases, but it is, in general, neither as easy nor as reliable as practitioners often seem to believe.

Although there is a very large literature on bootstrapping in statistics, a surprisingly small proportion of it is devoted to bootstrap testing. Instead, the focus is usually on estimating bootstrap standard errors and constructing bootstrap confidence intervals. Two classic books are Efron and Tibshirani (1993) and Davison and Hinkley (1997). There have been many useful survey papers, including DiCiccio and Efron (1996), MacKinnon (2002), Davison, Hinkley, and Young (2003), and Horowitz (2001, 2003).

The next section discusses the basic ideas of bootstrap testing and its relationship with Monte Carlo testing. Section 3 explains what determines how well bootstrap tests perform under the null hypothesis. Section 4 discusses double bootstrap and fast double bootstrap tests. Section 5 discusses various bootstrap data generating processes. Section 6 discusses tests of multiple hypotheses. Section 7 presents some simulation results for a particular case which illustrate how important the choice of bootstrap DGP can be, and Section 8 concludes.

2. Bootstrap and Monte Carlo Tests

Suppose that $\hat{\tau}$ is the observed value of a test statistic τ, and we wish to perform a test at level α that rejects when $\hat{\tau}$ is in the upper tail. Then the ***P* value**, or **marginal significance level**, of $\hat{\tau}$ is

$$p(\hat{\tau}) = 1 - F(\hat{\tau}), \tag{1}$$

where $F(\tau)$ is the cumulative distribution function of τ under the null hypothesis. If we knew $F(\tau)$, we would simply calculate $p(\hat{\tau})$ and reject the null whenever $p(\hat{\tau}) < \alpha$. This is equivalent to rejecting whenever $\hat{\tau}$ exceeds the **critical value** $F_{1-\alpha}(\tau)$, which is the $1 - \alpha$ quantile of $F(\tau)$. When we do not know $F(\tau)$, which is usually the case, it is common to use an asymptotic approximation to it. This may or may not work well.

An increasingly popular alternative is to perform a bootstrap test. We first generate B **bootstrap samples**, or simulated data sets, indexed by j. The procedure for generating the bootstrap samples is called a **bootstrap data generating process**, or **bootstrap DGP**, and there are often a number of choices. Some bootstrap DGPs may be fully parametric, others may be fully nonparametric, and still others may be partly parametric; see Section 5. Each bootstrap sample is then used to compute a **bootstrap test statistic**, say τ_j^*, most commonly by the same procedure used to calculate $\hat{\tau}$ from the real sample. It is strongly recommended that the bootstrap samples should satisfy the null hypothesis, but this is not always possible. When they do not satisfy the null, the τ_j^* cannot be calculated in quite the same way as $\hat{\tau}$ itself.

If we wish to reject when $\hat{\tau}$ is in the upper tail, the bootstrap P value is simply

$$\hat{p}^*(\hat{\tau}) = 1 - \hat{F}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau_j^* > \hat{\tau}), \tag{2}$$

where $\hat{F}^*$ denotes the **empirical distribution function**, or **EDF**, of the τ_j^*, and $\mathrm{I}(\cdot)$ denotes the **indicator function**, which is equal to 1 when its argument is true and 0 otherwise. The inequality would be reversed if we wished to reject when $\hat{\tau}$ is in the lower tail, as with many unit root tests. Thus the bootstrap P value is, in general, simply the proportion of the bootstrap test statistics τ_j^* that are more extreme than the observed test statistic $\hat{\tau}$. In the case of (2), it is in fact the empirical analog of (1). Of course, rejecting the null hypothesis whenever $\hat{p}^*(\hat{\tau}) < \alpha$ is equivalent to rejecting it whenever $\hat{\tau}$ exceeds the $1 - \alpha$ quantile of $\hat{F}^*$.

Perhaps surprisingly, this procedure can actually yield an exact test in certain cases. The key requirement is that the test statistic τ should be **pivotal**, which means that its distribution does not depend on anything that is unknown. This implies that τ and the τ_j^* all follow the same distribution if the null is true. In addition, the number of bootstrap samples B must be such that $\alpha(B+1)$ is an integer, where α is the level of the test. If a bootstrap test satisfies these two conditions, then it is exact. This sort of test, which was originally proposed in Dwass (1957), is generally called a **Monte Carlo test**. For an introduction to Monte Carlo testing, see Dufour and Khalaf (2001).

It is quite easy to see why Monte Carlo tests are exact. Imagine sorting all $B + 1$ test statistics. Then rejecting the null whenever $\hat{p}(\hat{\tau}) < \alpha$ implies rejecting it whenever $\hat{\tau}$ is one of the largest $\alpha(B + 1)$ statistics. But, if $\hat{\tau}$ and the τ_j^* all follow the same distribution, this happens with probability precisely α. For example, if $B = 999$ and $\alpha = .01$, we reject the null whenever $\hat{\tau}$ is one of the 10 largest test statistics.

Since a Monte Carlo test is exact whenever $\alpha(B + 1)$ is an integer, it is tempting to make B very small. In principle, it could be as small as 19 for $\alpha = .05$ and as small as 99 for $\alpha = .01$. There are two problems with this, however. The first problem is that the smaller is B the less powerful is the test. The loss of power is proportional to $1/B$; see Jöckel (1986) and Davidson and MacKinnon (2000).

The second problem is that, when B is small, the results of the test can depend nontrivially on the particular sequence of random numbers used to generate the bootstrap test statistics, and most investigators find this unsatisfactory. Since $\hat{p}^*$ is just a frequency, the variance of $\hat{p}^*$ is $p^*(1-p^*)/B$. Thus, when $p^* = .05$, the standard error of $\hat{p}^*$ is 0.0219 for $B = 99$, 0.0069 for $B = 999$, and 0.0022 for $B = 9999$. This suggests that, if computational cost is not a serious concern, it might be dangerous to use a value of B less than 999, and it would not be unreasonable to use $B = 9999$.

When computational cost is a concern, but it is not extremely high, it is often possible to obtain reliable results for a small value of B by using an iterative procedure proposed in Davidson and MacKinnon (2000). The idea is to start with a small value of B, decide whether the outcome of the test would almost certainly have been the same if B had been ∞, and then increase B if not. This process is then repeated until either an unambiguous result is obtained or it is clear that $\hat{p}^*(\hat{\tau})$ is very, very close to α. For example, if $B = 19$, and 5 or more of the τ_j^* are greater than $\hat{\tau}$, we can safely decide not to reject at the .05 level, because the probability of obtaining that many values of τ_j^* larger than $\hat{\tau}$ by chance if $p^* = .05$ is very small (it is actually .00024). However, if fewer than 5 of the τ_j^* are greater than $\hat{\tau}$, we need to generate some more bootstrap samples and calculate a new P value. Much of the time, especially when the null hypothesis is true, this procedure stops when B is under 100.

When computational cost is extremely high, two useful procedures have recently been proposed. Racine and MacKinnon (2007a) proposes a very simple method for performing Monte Carlo tests that does not require $\alpha(B+1)$ to be an integer. However, this procedure may lack power. Racine and MacKinnon (2007b) proposes a more complicated method for calculating bootstrap P values that is based on kernel smoothing. The P value depends on the actual values of $\hat{\tau}$ and the τ_j^*, not just on the rank of $\hat{\tau}$ in the sorted list of all the test statistics. This method does not quite yield exact tests, but it can substantially increase power when B is very small. In some cases, one can reliably reject the null at the .05 level using fewer than 20 bootstrap samples.

Quite a few popular specification tests in econometrics are pivotal if we condition on the regressors and the distribution of the error terms is assumed to be known. These include any test that just depends on ordinary least squares residuals and on the matrix of regressors and that does not depend on the variance of the error terms. Examples include the Durbin-Watson test and several other tests for serial correlation, as well as popular tests for heteroskedasticity, skewness, and kurtosis. Tests for heteroskedasticity will be discussed further in the next section.

It is very easy to perform this sort of Monte Carlo test. After we calculate $\hat{\tau}$ from the residual n–vector $\hat{\boldsymbol{u}}$ and, possibly, the regressor matrix $\boldsymbol{X}$, we generate B bootstrap samples. We can do this by generating B vectors of length n from the standard normal distribution, or possibly from some other assumed distribution, and regressing each of them on $\boldsymbol{X}$ so as to obtain B vectors of bootstrap residuals $\hat{\boldsymbol{u}}_j^*$. We then calculate each of the bootstrap statistics τ_j^* from $\hat{\boldsymbol{u}}_j^*$ in the same way that we calculated $\hat{\tau}$ from $\hat{\boldsymbol{u}}$. Provided we have chosen B correctly, the bootstrap P value (2) will then provide us with an exact test.

Even when τ is not pivotal, using the bootstrap to compute a P value like (2) is asymptotically valid. Moreover, this type of bootstrap testing can in many cases yield more accurate results than using an asymptotic distribution to compute a P value. This subject will be discussed in the next section.

When we wish to perform a two-tailed test, we cannot use equation (2) to compute a bootstrap P value. If we are willing to assume that τ is symmetrically distributed around zero, we can use the **symmetric bootstrap *P* value**

$$\hat{p}_{\mathrm{s}}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\big(|\tau_j^*| > |\hat{\tau}|\big), \tag{3}$$

which effectively converts a two-tailed test into a one-tailed test. If we are not willing to make this assumption, which can be seriously false for t statistics based on parameter estimates that are biased, we can instead use the **equal-tail bootstrap *P* value**

$$\hat{p}_{\mathrm{et}}^*(\hat{\tau}) = 2\min\!\bigg(\frac{1}{B}\sum_{j=1}^{B}\mathrm{I}(\tau_j^* \le \hat{\tau}),\ \frac{1}{B}\sum_{j=1}^{B}\mathrm{I}(\tau_j^* > \hat{\tau})\bigg). \tag{4}$$

Here we actually perform two tests, one against values in the lower tail of the distribution and the other against values in the upper tail. The factor of 2 is necessary to take account of this. Without it, $\hat{p}_{\mathrm{et}}^*$ would lie between 0 and 0.5. If the mean of the τ_j^* is far from zero, $\hat{p}_{\mathrm{s}}^*$ and $\hat{p}_{\mathrm{et}}^*$ may be very different, and tests based on them may have very different properties under both null and alternative hypotheses.

Of course, equation (3) can only apply to test statistics that can take either sign, such as t statistics. For test statistics that are always positive, such as ones that are asymptotically chi-squared, only equation (2) is usually applicable. But we could use (4) if we wanted to reject for small values of the test statistic as well as for large ones.

Equations (2), (3), and (4) imply that the results of a bootstrap test are invariant to monotonically increasing transformations of the test statistic. Applying the same transformation to all the test statistics does not affect the rank of $\hat{\tau}$ in the sorted list of $\hat{\tau}$ and the τ_j^*, and therefore it does not affect the bootstrap P value. One of the implications of this result is that, for linear and nonlinear regression models, a bootstrap F test and a bootstrap likelihood ratio test based on the same bootstrap DGP must yield the same outcome.

3. Finite-Sample Properties of Bootstrap Tests

Even when B is infinite, bootstrap tests will generally not be exact when τ is not pivotal. Their lack of exactness arises from the difference between the true distribution characterized by the CDF $F(\tau)$ and the bootstrap distribution characterized by the CDF $F^*(\tau)$. When more than one sort of bootstrap DGP can be used, we should always use the one that makes $F^*(\tau)$ as close as possible to $F(\tau)$ in the neighborhood of the critical value $F_{1-\alpha}(\tau)$. Unfortunately, this is easier said than done. Either very sophisticated econometric theory or extensive simulation experiments may be needed to determine which of several bootstrap DGPs leads to the most reliable tests.

A very important result, which may be found in Beran (1988), shows that, when a test statistic is **asymptotically pivotal**, bootstrapping yields what is generally called an **asymptotic refinement**. A test statistic is asymptotically pivotal if its asymptotic distribution does not depend on anything that is unknown. If a test statistic has a conventional asymptotic distribution such as standard normal or chi-squared, then it must be asymptotically pivotal. But a test statistic can be asymptotically pivotal without having a known asymptotic distribution. In this context, what is meant by an asymptotic refinement is that the **error in rejection probability**, or **ERP**, of the bootstrap test is of a lower order in the sample size n than the ERP of an asymptotic test based on the same test statistic.

A serious treatment of asymptotic refinements is well beyond the scope of this paper. Rigorous discussions are generally based on Edgeworth expansions of the distributions of test statistics; see Beran (1988) and Hall (1992). Davidson and MacKinnon (1999) take a different approach based on what they call the **rejection probability function**, or **RPF**, which relates the probability that a test will reject the null to one or more nuisance parameters. Strictly speaking, this approach applies only to parametric bootstrap tests, but it helps to illuminate other cases as well.

Consider the case in which the DGP depends on a single nuisance parameter. If the RPF is flat, as it must be when a test statistic is pivotal, then a parametric bootstrap test will be exact, because the value of the nuisance parameter does not matter. When the RPF is not flat, as is commonly the case, a bootstrap test will generally not be exact, because the estimated parameter of the bootstrap DGP will differ from the (unknown) parameter of the true DGP. How well a bootstrap test performs in such a case depends on the slope and curvature of the RPF and the bias and precision of the estimated nuisance parameter.

Although the actual behavior of bootstrap tests in simulation experiments does not accord well with theory in every case, the literature on the finite-sample properties of bootstrap tests has produced several theoretical results of note:

- When a test statistic is asymptotically pivotal, bootstrapping it will generally yield a test with an ERP of lower order in the sample size than that of an asymptotic test based on the same statistic.
- When a test statistic is not asymptotically pivotal, bootstrapping it will generally yield an asymptotically valid test, but the ERP of the bootstrap test will not be of lower order in the sample size than that of an asymptotic test.

- In many cases, the reduction in ERP due to bootstrapping is $O(n^{-1/2})$ for one-tailed tests and $O(n^{-1})$ for two-tailed tests that assume symmetry around the origin. Note that, when a test statistic is asymptotically chi-squared, a test that rejects when the statistic is in the upper tail has the properties of a two-tailed test from the point of view of this theory. In contrast, a test based on the equal-tail P value (4) has the properties of a one-tailed test.
- To minimize the Type I errors committed by bootstrap tests, we should attempt to estimate the bootstrap DGP as efficiently as possible. This generally means imposing the null hypothesis whenever it is possible to do so.

In general, bootstrap tests can be expected to perform well under the null whenever the bootstrap DGP provides a good approximation to those aspects of the true DGP to which the distribution of the test statistic is sensitive. Since different test statistics may be sensitive to different features of the DGP, it is quite possible that a particular bootstrap DGP may work well for some tests and poorly for others.

It is not always easy to impose the null hypothesis on a bootstrap DGP without also imposing parametric assumptions that the investigator may not be comfortable with. Various partly or wholly nonparametric bootstrap DGPs, some of which impose the null and some of which do not, are discussed in Section 5. Martin (2007) discusses how to impose the null in a nonparametric way in certain cases of interest.

Although there is always a modest loss of power due to bootstrapping when B is small, bootstrapping when B is large generally has little effect on power. Comparing the power of tests that do not have the correct size is fraught with difficulty; see Davidson and MacKinnon (2006a). It is shown in that paper that, if bootstrap and asymptotic tests based on the same test statistic are size-corrected in a sensible way, then any difference in power should be modest.

Since Monte Carlo tests are exact, and bootstrap tests are generally not exact, it may seem attractive to use the former whenever possible. The problem is that, in order to do so, it is generally necessary to make strong distributional assumptions. For concreteness, consider tests for heteroskedasticity. Dufour *et al.* (2004) show that many popular test statistics for heteroskedasticity in the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta}+\boldsymbol{u}$ are pivotal when the regressors are treated as fixed and the distribution of the error terms is known up to a scale factor. These statistics all have the form $\tau(\boldsymbol{Z}, \boldsymbol{M_X u})$. That is, they simply depend on a matrix $\boldsymbol{Z}$ of regressors that is treated as fixed and on the residual vector $\boldsymbol{M_X u}$, where the projection matrix $\boldsymbol{M_X}$ yields residuals from a regression on $\boldsymbol{X}$. Moreover, they are invariant to σ^2, the variance of the error terms.

One particularly popular procedure, proposed in Koenker (1981), involves taking the vector of residuals $\boldsymbol{M_X y}$, squaring each element, regressing the squared residuals on a constant and the matrix $\boldsymbol{Z}$, and then calculating the usual F statistic for all slope coefficients to be zero. The Koenker procedure is asymptotically invariant to the distribution of the error terms. It was originally proposed as a modification to the LM test of Breusch and Pagan (1979), which is asymptotically valid only when the error terms are normally distributed.

Performing a Monte Carlo test for heteroskedasticity simply involves drawing B error vectors $\boldsymbol{u}_j^*$ from an assumed distribution, using each of them to calculate a bootstrap statistic $\tau(\boldsymbol{Z}, \boldsymbol{M_X u}_j^*)$, and then calculating a P value by (2). But what if the assumed distribution is incorrect? As Godfrey, Orme, and Santos Silva (2005) show, when the distribution from which the $\boldsymbol{u}_j^*$ are drawn differs from the true distribution of $\boldsymbol{u}$, Monte Carlo tests for heteroskedasticity can be seriously misleading. In particular, Monte Carlo versions of Breusch-Pagan tests can overreject or underreject quite severely in samples of any size, while Monte Carlo versions of Koenker tests can suffer from modest size distortions when the sample size is small. In contrast, nonparametric bootstrap tests in which the $\boldsymbol{u}_j^*$ are obtained by resampling residuals (this sort of bootstrap DGP will be discussed in Section 5) always yield valid results for large samples and generally perform reasonably well even in small samples.

None of these results is at all surprising. In order to obtain valid results in large samples, we either need to use a test statistic, such as the F statistic of Koenker (1981), that is asymptotically invariant to the distribution of the error terms, or we need to use a bootstrap DGP that adapts itself asymptotically to the distribution of the error terms, or preferably both. Using a Monte Carlo test based on the wrong distributional assumption together with a test statistic that is not asymptotically invariant to the distribution of the error terms is a recipe for disaster.

Another way to get around this sort of problem, instead of using a nonparametric bootstrap, is proposed in Dufour (2006). This paper introduces **maximized Monte Carlo tests**. In principle, these can be applied to any sort of test statistic where the null distribution depends on one or more nuisance parameters. The idea is to perform a (possibly very large) number of simulation experiments, each for a different set of nuisance parameters. Using some sort of numerical search algorithm, the investigator searches over the nuisance parameter(s) in an effort to maximize the bootstrap P value. The null hypothesis is rejected only if the maximized P value is less than the predetermined level of the test. In the context of testing for heteroskedasticity, it is necessary to search over a set of possible error distributions. An application of maximized Monte Carlo tests to financial economics may be found in Beaulieu, Dufour, and Khalaf (2007).

The idea of maximized Monte Carlo tests is elegant and ingenious, but these tests can be very computationally demanding. Moreover, their actual rejection frequency may be very much less than the level of the test, and they may in consequence be severely lacking in power. This can happen when the RPF is strongly dependent on the value(s) of one or more nuisance parameters, and there exist parameter values (perhaps far away from the ones that actually generated the data) for which the rejection probability under the null is very high. The maximized Monte Carlo procedure will then assign a much larger P value than the one that would have been obtained if the true, but unknown, values of the nuisance parameters had been used.

4. Double Bootstrap and Fast Double Bootstrap Tests

It seems plausible that, if bootstrapping a test statistic leads to an asymptotic refinement, then bootstrapping a quantity that has already been bootstrapped will lead to a further refinement. This is the basic idea of the **iterated bootstrap**, of which a special case is the **double bootstrap**, proposed in Beran (1987, 1988).

There are at least two quite different types of double bootstrap test. The first type potentially arises whenever we do not have an asymptotically pivotal test statistic to start with. Suppose, for example, that we obtain a vector of parameter estimates $\hat{\boldsymbol{\theta}}$ but no associated covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\theta}})$, either because it is impossible to estimate $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ at all or because it is impossible to obtain a reasonably accurate estimate. We can always use the bootstrap to estimate $\mathrm{Var}(\hat{\boldsymbol{\theta}})$. If $\hat{\boldsymbol{\theta}}^*_j$ denotes the estimate from the j^{th} bootstrap sample, and $\bar{\boldsymbol{\theta}}^*$ denotes the average of the bootstrap estimates, then

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) \equiv \frac{1}{B}\sum_{j=1}^{B}(\hat{\boldsymbol{\theta}}^*_j - \bar{\boldsymbol{\theta}}^*)(\hat{\boldsymbol{\theta}}^*_j - \bar{\boldsymbol{\theta}}^*)^{\top}$$

provides a reasonable way to estimate $\mathrm{Var}(\hat{\boldsymbol{\theta}})$. Note that whatever bootstrap DGP is used here should not impose any restrictions on $\boldsymbol{\theta}$.

We can easily construct a variety of Wald statistics using $\hat{\boldsymbol{\theta}}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$. The simplest would just be an asymptotic t statistic for some element of $\boldsymbol{\theta}$ to equal a particular value. Creating a t statistic from a parameter estimate, or a Wald statistic from a vector of parameter estimates, is sometimes called **prepivoting**, because it turns a quantity that is not asymptotically pivotal into one that is.

Even though test statistics of this sort are asymptotically pivotal, their asymptotic distributions may not provide good approximations in finite samples. Thus it seems natural to bootstrap them. Doing so is conceptually easy but computationally costly. The procedure works as follows:

1. Obtain the estimates $\hat{\boldsymbol{\theta}}$.
2. Generate B_2 bootstrap samples from a bootstrap DGP that does not impose any restrictions, and use them to estimate $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$.
3. Calculate whatever test statistic $\hat{\tau} \equiv \tau\big(\hat{\boldsymbol{\theta}}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\big)$ is of interest.
4. Generate B_1 bootstrap samples using a bootstrap DGP that imposes whatever restrictions are to be tested. Use each of them to calculate $\hat{\boldsymbol{\theta}}^*_j$.
5. For each of the B_1 bootstrap samples, perform steps 2 and 3 exactly as before. This yields B_1 bootstrap test statistics $\tau^*_j \equiv \tau\big(\hat{\boldsymbol{\theta}}^*_j, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}^*_j)\big)$.
6. Calculate the bootstrap P value for $\hat{\tau}$ using whichever of the formulas (2), (3), or (4) is appropriate, with B_1 playing the role of B.

This procedure is conceptually simple, and, so long as the procedure for computing $\hat{\boldsymbol{\theta}}$ is reliable, it should be straightforward to implement. The major problem is computational cost, which can be formidable, because we need to obtain no less than

$(B_1+1)(B_2+1)$ estimates of $\boldsymbol{\theta}$. For example, if $B_1 = 999$ and $B_2 = 500$, we need to obtain 501,000 sets of estimates. It may therefore be attractive to utilize either the method of Davidson and MacKinnon (2000) or the one of Racine and MacKinnon (2007b), or both together, because they can allow B_1 to be quite small.

The second type of double bootstrap test arises when we do have an asymptotically pivotal test statistic τ to start with. It works as follows:

1. Obtain the test statistic $\hat{\tau}$ and whatever estimates are needed to generate bootstrap samples that satisfy the null hypothesis.

2. Generate B_1 bootstrap samples that satisfy the null, and use each of them to compute a bootstrap statistic τ_j^* for $j = 1, \ldots, B_1$.

3. Use $\hat{\tau}$ and the τ_j^* to calculate the first-level bootstrap P value $\hat{p}^*(\hat{\tau})$ according to, for concreteness, equation (2).

4. For each of the B_1 first-level bootstrap samples, generate B_2 second-level bootstrap samples, and use each of them to compute a second-level bootstrap test statistic τ_{jl}^{**} for $l = 1, \ldots, B_2$.

5. For each of the B_1 first-level bootstrap samples, compute the second-level bootstrap P value

$$\hat{p}_j^{**} = \frac{1}{B_2} \sum_{l=1}^{B_2} \mathrm{I}(\tau_{jl}^{**} > \tau_j^*).$$

   Observe that this formula is very similar to equation (2), but we are comparing τ_j^* with the τ_{jl}^{**} instead of comparing $\hat{\tau}$ with the τ_j^*.

6. Compute the **double-bootstrap *P* value** as the proportion of the $\hat{p}_j^{**}$ that are smaller (i.e., more extreme) than $\hat{p}^*(\hat{\tau})$:

$$\hat{p}^{**}(\hat{\tau}) = \frac{1}{B_1} \sum_{j=1}^{B_1} \mathrm{I}\big(\hat{p}_j^{**} < \hat{p}^*(\hat{\tau})\big). \tag{5}$$

   In order to avoid the possibility that $\hat{p}_j^{**} = \hat{p}^*(\hat{\tau})$, which would make the strict inequality here problematical, it is desirable that $B_2 \neq B_1$.

This type of double bootstrap simply treats the single bootstrap P value as a test statistic and bootstraps it. Observe that (5) is just like (2), but with the inequality reversed. Like the first type of double bootstrap, this one is computationally demanding, as we need to calculate $1 + B_1 + B_1B_2$ test statistics, and it may therefore be attractive to employ methods that allow B_1 and/or B_2 to be small.

The computational cost of performing this type of double bootstrap procedure can be substantially reduced by utilizing one or more ingenious stopping rules proposed in Nankervis (2005). The idea of these stopping rules is to avoid unnecessarily calculating second-level bootstrap test statistics that do not affect the decision on whether or not to reject the null hypothesis.

The asymptotic refinement that we can expect from the second type of double bootstrap test is greater than we can expect from the first type. For the first type, the ERP will normally be of the same order in the sample size as the ERP of an ordinary (single) bootstrap test. For the second type, it will normally be of lower order. Of course, this does not imply that double bootstrap tests of the second type will always work better than double bootstrap tests of the first type.

Davidson and MacKinnon (2007) proposes a procedure for computing **fast double bootstrap**, or **FDB**, P values for asymptotically pivotal test statistics. It is much less computationally demanding than performing a true double bootstrap, even if the procedures of Nankervis (2005) are employed, because there is just one second-level bootstrap sample for each first-level one, instead of B_2 of them. Steps 1, 2, and 3 of the procedure just given are unchanged, except that B replaces B_1. Steps 4, 5, and 6 are replaced by the following ones:

4. For each of the B bootstrap samples, generate a single dataset using a second-level bootstrap sample, and use it to compute a second-level test statistic τ_j^{**}.
5. Calculate the $1-\hat{p}^*$ quantile of the τ_j^{**}, denoted by $\hat{Q}_B^{**}(1-\hat{p}^*)$ and defined implicitly by the equation

$$\frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\Big(\tau_j^{**} < \hat{Q}_B^{**}\big(1-\hat{p}^*(\hat{\tau})\big)\Big) = 1-\hat{p}^*(\hat{\tau}). \tag{6}$$

   Of course, for finite B, there will be a range of values of Q_B^{**} that satisfy (6), and we must choose one of them somewhat arbitrarily.
6. Calculate the FDB P value as

$$\hat{p}_F^{**}(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\Big(\tau_j^{*} > \hat{Q}_B^{**}\big(1-\hat{p}^*(\hat{\tau})\big)\Big).$$

   Thus, instead of seeing how often the bootstrap test statistics are more extreme than the actual test statistic, we see how often they are more extreme than the $1-\hat{p}^*$ quantile of the τ_j^{**}.

The great advantage of this procedure is that it involves the calculation of just $2B+1$ test statistics, although B should be reasonably large to avoid size distortions on the order of $1/B$. However, for the FDB to be asymptotically valid, τ must be asymptotically independent of the bootstrap DGP. This is a reasonable assumption for the parametric bootstrap when the parameters of the bootstrap DGP are estimated under the null hypothesis, because a great many test statistics are asymptotically independent of all parameter estimates under the null; see Davidson and MacKinnon (1999).[1]

[1] For the linear regression model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$ with normal errors, it is easy to show that the F statistic for $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is independent of the restricted estimates $\tilde{\boldsymbol{\beta}}_1$, because the latter depend on the projection of $\boldsymbol{y}$ onto the subspace spanned by the columns of $\boldsymbol{X}_1$, and the former depends on its orthogonal complement. For more general types of model, the same sort of independence holds, but only asymptotically.

The assumption makes sense in a number of other cases as well, including the residual and wild bootstraps when they use parameter estimates under the null.

There is no guarantee that double bootstrap tests will always work well. Experience suggests that, if the improvement in ERP from using a single-level bootstrap test is modest, then the further gain from using either a double bootstrap test of the second type or an FDB test is likely to be even more modest.

5. Bootstrap Data Generating Processes

From the discussion in Section 3, it is evident that the choice of bootstrap DGP is absolutely critical. Just what choices are available depend on the model being estimated and on the assumptions that the investigator is willing to make.

At the heart of many bootstrap DGPs is the idea of **resampling**, which was the key feature of the earliest bootstrap methods proposed in Efron (1979, 1982). Suppose we are interested in some quantity $\theta(\boldsymbol{y})$, where $\boldsymbol{y}$ is an n–vector of data with typical element y_t. What is meant by resampling is that each element of every bootstrap sample $\boldsymbol{y}_j^*$ is drawn randomly from the EDF of the y_t, which assigns probability $1/n$ to each of the y_t. Thus each element of every bootstrap sample can take on n possible values, namely, the values of the y_t, each with probability $1/n$. Each bootstrap sample therefore contains some of the y_t just once, some of them more than once, and some of them not at all.

Resampling is conceptually simple and computationally efficient. It works in theory, at least asymptotically, because the EDF consistently estimates the distribution of the y_t. For regression models, and other models that involve averaging, it often works very well in practice. However, in certain respects, the bootstrap samples $\boldsymbol{y}_j^*$ differ fundamentally from the original sample $\boldsymbol{y}$. For example, the largest element of $\boldsymbol{y}_j^*$ can never exceed the largest element of $\boldsymbol{y}$ and will be less than it nearly 37% of the time. Thus resampling is likely to yield very misleading results if we are primarily interested in the tails of a distribution.

Bootstrap samples based on resampling are less diverse than the original sample. When the y_t are drawn from a continuous distribution, the n elements of $\boldsymbol{y}$ are all different. But the n elements of each bootstrap sample inevitably include a number of duplicates, because the probability that any element of the original sample will not appear in a particular bootstrap sample approaches $1/e = 0.3679$ as $n \to \infty$. This can cause bootstrap DGPs based on resampling to yield invalid inferences, even asymptotically. An important example is discussed in Abadie and Imbens (2006), which shows that certain bootstrap methods based on resampling are invalid for matching estimators.

Resampling can be used in a variety of ways. One of the most general and widely used bootstrap DGPs is the **pairs bootstrap**, or **pairwise bootstrap**, which dates back to Freedman (1981, 1984). The idea is simply to resample the data, keeping the dependent and independent variables together in pairs. In the context of the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, this means forming the matrix $[\boldsymbol{y}\;\; \boldsymbol{X}]$ with typical row

$[y_t \;\; \boldsymbol{X}_t]$ and resampling the rows of this matrix. Each observation of a bootstrap sample is $[y_t^* \;\; \boldsymbol{X}_t^*]$, a randomly chosen row of $[\boldsymbol{y} \;\; \boldsymbol{X}]$. This method implicitly assumes that each observation $[y_t \;\; \boldsymbol{X}_t]$ is an independent random drawing from a multivariate distribution, which may or may not be a reasonable assumption.

The pairs bootstrap can be applied to a very wide range of models, not merely regression models. It is most natural to use it with cross-section data, but it is often used with time-series data as well. When the regressors include lagged dependent variables, we simply treat them the same way as any other column of $\boldsymbol{X}$. The pairs bootstrap does not require that the error terms be homoskedastic. Indeed, error terms do not explicitly appear in the bootstrap DGP at all. However, a serious drawback of this method is that it does not condition on $\boldsymbol{X}$. Instead, each bootstrap sample has a different $\boldsymbol{X}^*$ matrix. This can lead to misleading inferences in finite samples when the distribution of a test statistic depends strongly on $\boldsymbol{X}$.

In the context of bootstrap testing, the pairs bootstrap is somewhat unsatisfactory. Because it is completely nonparametric, the bootstrap DGP does not impose any restrictions on the parameters of the model. If we are testing such restrictions, as opposed to estimating covariance matrices or standard errors, we need to modify the bootstrap test statistic so that it is testing something which is true in the bootstrap DGP. Suppose the actual test statistic takes the form of a t statistic for the hypothesis that $\beta = \beta^0$:

$$\hat{\tau} = \frac{\hat{\beta} - \beta^0}{s(\hat{\beta})}\,.$$

Here $\hat{\beta}$ is the unrestricted estimate of the parameter β that is being tested, and $s(\hat{\beta})$ is its standard error. The bootstrap test statistic cannot test the hypothesis that $\beta = \beta^0$, because that hypothesis is not true for the bootstrap DGP. Instead, we must use the bootstrap test statistic

$$\tau_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}}{s(\hat{\beta}_j^*)}\,, \tag{7}$$

where $\hat{\beta}_j^*$ is the estimate of β from the j^{th} bootstrap sample, and $s(\hat{\beta}_j^*)$ is its standard error, calculated by whatever procedure was employed to calculate $s(\hat{\beta})$ using the actual sample. Since the estimate of β from the bootstrap samples should, on average, be equal to $\hat{\beta}$, at least asymptotically, the null hypothesis tested by τ_j^* is "true" for the pairs bootstrap DGP.

As an aside, when $s(\hat{\beta})$ is difficult to compute, it is very tempting to replace $s(\hat{\beta}_j^*)$ in (7) by $s(\hat{\beta})$. This temptation should be resisted at all costs. If we were to use the same standard error to compute both $\hat{\tau}$ and τ_j^*, then we would effectively be using the nonpivotal quantity $\hat{\beta} - \beta^0$ as a test statistic. Bootstrapping it would be asymptotically valid, but it would offer no asymptotic refinement. We can always use the bootstrap to compute $s(\hat{\beta})$ and the $s(\hat{\beta}_j^*)$, which will lead to the first type of double bootstrap test discussed in the previous section.

In the case of a restriction like $\beta = \beta^0$, it is easy to modify the bootstrap test statistic, as in (7), so that bootstrap testing yields valid results. But this may not be so easy to do when there are several restrictions and some of them are nonlinear. Great care must be taken when using the pairs bootstrap in such cases.

Many simulation results suggest that the pairs bootstrap is never the procedure of choice for regression models. Consider the nonlinear regression model

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{8}$$

where $x_t(\boldsymbol{\beta})$ is a regression function, in general nonlinear in the parameter vector $\boldsymbol{\beta}$, that implicitly depends on exogenous and predetermined variables. It is assumed either that the sample consists of cross-section data or that the regressand and all regressors are stationary. Note that the error terms u_t are assumed to be independent and identically distributed.

A good way to bootstrap this sort of model is to use the **residual bootstrap**. The first step is to estimate (8) under the null hypothesis, obtaining parameter estimates $\tilde{\boldsymbol{\beta}}$ and residuals $\tilde{u}_t$. If the model (8) does not have a constant term or the equivalent, then the residuals may not have mean zero, and they should be recentered. Unless the test statistic to be bootstrapped is invariant to the variance of the error terms, it is advisable to rescale the residuals so that they have the correct variance. The simplest type of rescaled residual is

$$\ddot{u}_t \equiv \left(\frac{n}{n - k_1}\right)^{1/2} \tilde{u}_t, \tag{9}$$

where k_1 is the number of parameters estimated under the null hypothesis. The first factor here is the inverse of the square root of the factor by which $1/n$ times the sum of squared residuals underestimates σ^2. A somewhat more complicated method uses the diagonals of the **hat matrix**, that is, $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, to rescale each residual by a different factor. It may work a bit better than (9) when some observations have high leverage. Note that this method is a bit more complicated than the analogous one for the wild bootstrap that is described below, because we need to ensure that the rescaled residuals have mean zero; see Davidson and MacKinnon (2006b).

The residual bootstrap DGP may be written as

$$y_t^* = x_t(\tilde{\boldsymbol{\beta}}) + u_t^*, \quad u_t^* \sim \text{EDF}(\ddot{u}_t). \tag{10}$$

In other words, we evaluate the regression function at the restricted estimates and then resample from the rescaled residuals. There might be a slight advantage in terms of power if we were to use unrestricted rather than restricted residuals in (10). This would involve estimating the unrestricted model to obtain residuals $\hat{u}_t$ and then replacing $\tilde{u}_t$ by $\hat{u}_t$ and k_1 by k in (9). What is important is that we use $\tilde{\boldsymbol{\beta}}$ rather than $\hat{\boldsymbol{\beta}}$. Doing so allows us to compute the bootstrap test statistics in exactly the same way as the actual one. Moreover, since $\tilde{\boldsymbol{\beta}}$ will, in most cases, be more precisely estimated than $\hat{\boldsymbol{\beta}}$, the bootstrap DGP (10) will more closely approximate the true DGP than it would if we used $\hat{\boldsymbol{\beta}}$.

When $x_t(\boldsymbol{\beta})$ includes lagged values of the dependent variable, the residual bootstrap should be modified so that the y_t^* are generated recursively. For example, if the restricted model were $y_t = \beta_1 + \beta_2 z_t + \beta_3 y_{t-1} + u_t$, the bootstrap DGP would be

$$y_t^* = \tilde{\beta}_1 + \tilde{\beta}_2 z_t + \tilde{\beta}_3 y_{t-1}^* + u_t^*, \quad u_t^* \sim \mathrm{EDF}(\ddot{u}_t). \tag{11}$$

In most cases, the observed pre-sample value(s) of y_t are used to start the recursion. It is important that the bootstrap DGP should be stationary, which in this case means that $|\tilde{\beta}_3| < 1$. If this condition is not satisfied naturally, it should be imposed on the bootstrap DGP.

The validity of the residual bootstrap depends on the strong assumption that the error terms are independent and identically distributed. We cannot use it when they are actually heteroskedastic. If the form of the heteroskedasticity is known, we can easily modify the residual bootstrap by introducing a skedastic function, estimating it, using feasible weighted least squares (linear or nonlinear), and then resampling from standardized residuals. But if the form of the heteroskedasticity is unknown, the best method that is currently available, at least for tests on $\boldsymbol{\beta}$, seems to be the **wild bootstrap**, which was originally proposed in Wu (1986).

For a restricted version of a model like (8), with independent but possibly heteroskedastic error terms, the wild bootstrap DGP is

$$y_t^* = x_t(\tilde{\boldsymbol{\beta}}) + f(\tilde{u}_t) v_t^*,$$

where $f(\tilde{u}_t)$ is a transformation of the t^{th} residual $\tilde{u}_t$, and v_t^* is a random variable with mean 0 and variance 1. One possible choice for $f(\tilde{u}_t)$ is just $\tilde{u}_t$, but a better choice is

$$f(\tilde{u}_t) = \frac{\tilde{u}_t}{(1 - h_t)^{1/2}}, \tag{12}$$

where h_t is the t^{th} diagonal of the "hat matrix" that was defined just after (9). When the $f(\tilde{u}_t)$ are defined by (12), they would have constant variance if the error terms were homoskedastic.

There are various ways to specify the distribution of the v_t^*. The simplest, but not the most popular, is the **Rademacher distribution**,

$$v_t^* = 1 \text{ with probability } \tfrac{1}{2}; \quad v_t^* = -1 \text{ with probability } \tfrac{1}{2}. \tag{13}$$

Thus each bootstrap error term can take on only two possible values. Davidson and Flachaire (2008) have shown that wild bootstrap tests based on (13) usually perform better than wild bootstrap tests which use other distributions when the conditional distribution of the error terms is approximately symmetric. When this distribution is sufficiently asymmetric, however, it may be better to use another two-point distribution, which is the one that is most commonly used in practice:

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

In either case, since the expectation of the square of $\tilde{u}_t$ is approximately the variance of u_t, the wild bootstrap error terms will, on average, have about the same variance as the u_t. In many cases, this seems to be enough for the wild bootstrap DGP to mimic the essential features of the true DGP.

Although it is most natural to use the wild bootstrap with cross-section data, it can also be used with at least some types of time-series model, provided the error terms are uncorrelated. See Gonçalves and Kilian (2004). The wild bootstrap can also be used with clustered data. In this case, the entire vector of residuals for each cluster is multiplied by v_t^* for each bootstrap sample so as to preserve any within-cluster relationships among the error terms. See Cameron, Gelbach, and Miller (2008).

Bootstrap methods may be particularly useful in the case of multivariate regression models, because standard asymptotic tests often overreject severely; see Stewart (1997) and Dufour and Khalaf (2002). When there are g dependent variables, a multivariate nonlinear regression model can be written, using notation similar to that of (8), as

$$y_{ti} = x_{ti}(\boldsymbol{\beta}) + u_{ti}, \quad t = 1, \ldots, n, \;\; i = 1, \ldots, g. \tag{14}$$

If we arrange the y_{ti}, x_{ti}, and u_{ti} into g–vectors $\boldsymbol{y}_t$, $\boldsymbol{x}_t$, and $\boldsymbol{u}_t$, respectively, the model (14) can be written more compactly as

$$\boldsymbol{y}_t = \boldsymbol{x}_t(\boldsymbol{\beta}) + \boldsymbol{u}_t, \quad \mathrm{E}(\boldsymbol{u}_t\boldsymbol{u}_t^\top) = \boldsymbol{\Sigma}. \tag{15}$$

Here the conventional assumption that the error terms are correlated across equations, but homoskedastic and serially uncorrelated, is made explicit. The model (15) is typically estimated by either feasible GLS or maximum likelihood. Doing so under the null hypothesis yields parameter estimates $\tilde{\boldsymbol{\beta}}$ and residuals $\tilde{\boldsymbol{u}}_t$.

The residual bootstrap DGP for the model (15) is simply

$$\boldsymbol{y}_t^* = \boldsymbol{x}_t(\tilde{\boldsymbol{\beta}}) + \boldsymbol{u}_t^*, \quad \boldsymbol{u}_t^* \sim \mathrm{EDF}(\tilde{\boldsymbol{u}}_t), \tag{16}$$

which is analogous to (10). We evaluate the regression functions at the restricted estimates $\tilde{\boldsymbol{\beta}}$ and resample from vectors of residuals, as in the pairs bootstrap. This resampling preserves the joint empirical distribution of the residuals, and in particular the correlations among them, without imposing any distributional assumptions. This approach is simpler and less restrictive than drawing the $\boldsymbol{u}_t^*$ from a multivariate normal distribution with covariance matrix estimated from the $\tilde{\boldsymbol{u}}_t$, which is a procedure that is also sometimes used. Of course, we may wish to impose the normality assumption since, in some cases, doing so makes it possible to perform Monte Carlo tests on multivariate linear regression models; see Dufour and Khalaf (2002). Alternatively, if we wished to allow for heteroskedasticity, we could use a wild bootstrap.

A particularly important type of multivariate regression model is the **simultaneous equations model**. Often, we are only interested in one equation from such a model, and it is estimated by generalized instrumental variables (also called two-stage least squares). A model with one structural equation and one or more reduced form equations may be written as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta}_1 + \boldsymbol{Y}\boldsymbol{\beta}_2 + \boldsymbol{u} \tag{17}$$

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\mathit{\Pi}} + \boldsymbol{V}. \tag{18}$$

Here $\boldsymbol{y}$ is an $n \times 1$ vector of the endogenous variable of interest, $\boldsymbol{Y}$ is an $n \times g$ matrix of other endogenous variables, $\boldsymbol{Z}$ is an $n \times k$ matrix of exogenous variables, and $\boldsymbol{W}$ is an $n \times l$ matrix of instruments, which must include all the exogenous variables. In finite samples, inferences about $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ based on asymptotic theory can be very unreliable, so it is natural to think about bootstrapping.

However, bootstrapping this model requires some effort. Even if we are only interested in the structural equation (17), the bootstrap DGP must generate both $\boldsymbol{y}^*$ and $\boldsymbol{Y}^*$. We could just use the pairs bootstrap (Freedman, 1984), but it often does not work very well. In order to use the residual bootstrap, we need estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\mathit{\Pi}}$, as well as a way to generate the bootstrap error terms $\boldsymbol{u}^*$ and $\boldsymbol{V}^*$. It is natural to use 2SLS estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ from (17) under the null hypothesis, along with OLS estimates of $\boldsymbol{\mathit{\Pi}}$ from (18). The bootstrap error terms may then be obtained by resampling from the residual vectors $[\hat{u}_t, \hat{\boldsymbol{V}}_t]$, as in (16).

Although the residual bootstrap DGP just described is asymptotically valid and generally provides less unreliable inferences than using asymptotic tests, it often does not work very well. But its finite-sample properties can be greatly improved if we use OLS estimates $\hat{\boldsymbol{\mathit{\Pi}}}'$ from the system of regressions

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\mathit{\Pi}} + \hat{\boldsymbol{u}}\boldsymbol{\delta}^\top + \text{residuals} \tag{19}$$

rather than the usual OLS estimates $\hat{\boldsymbol{\mathit{\Pi}}}$ from (18). In (19), $\hat{\boldsymbol{u}}$ is the vector of 2SLS residuals from estimation of (17) under the null hypothesis, and $\boldsymbol{\delta}$ is a g–vector of coefficients to be estimated. We also use $\hat{\boldsymbol{V}}' \equiv \boldsymbol{Y} - \boldsymbol{W}\hat{\boldsymbol{\mathit{\Pi}}}'$ instead of $\hat{\boldsymbol{V}}$ when we resample the bootstrap error terms.

It was shown in Davidson and MacKinnon (2008) that, in the case in which there is just one endogenous right-hand-side variable, t statistics for $\beta_2 = 0$ are far more reliable when the bootstrap DGP uses $\hat{\boldsymbol{\mathit{\Pi}}}'$ and $\hat{\boldsymbol{V}}'$ than when it uses $\hat{\boldsymbol{\mathit{\Pi}}}$ and $\hat{\boldsymbol{V}}$. This is especially true when the instruments are weak, which is when asymptotic tests tend to be seriously unreliable. A wild bootstrap variant of this procedure, which allows for heteroskedasticity of unknown form, was proposed in Davidson and MacKinnon (2009).

Any sort of resampling requires independence. Either the data must be treated as independent drawings from a multivariate distribution, as in the pairs bootstrap, or the error terms must be treated as independent drawings from either univariate or multivariate distributions. In neither case can serial dependence be accommodated. Several bootstrap methods that do allow for serial dependence have been proposed, however. Surveys of these methods include Bühlmann (2002), Horowitz (2003), Politis (2003), and Härdle, Horowitz, and Kreiss (2003).

One way to generalize the residual bootstrap to allow for serial correlation is to use what is called the **sieve bootstrap**, which assumes that the error terms follow an unknown, stationary process with homoskedastic innovations. The sieve bootstrap, which was given its name in Bühlmann (1997), attempts to approximate this stationary process, generally by using an AR(p) process, with p chosen either by some sort of model selection criterion or by sequential testing. One could also use MA or ARMA processes, and this appears to be preferable in some cases; see Richard (2007).

After estimating the regression model, say (8), under the null hypothesis, we retain the residuals and use them to select the preferred AR(p) process and estimate its coefficients, making sure that it is stationary. Of course, this two-step procedure would not be valid if (8) were a dynamic model. Then the bootstrap DGP is

$$\begin{aligned} u_t^* &= \sum_{i=1}^{p} \hat\rho_i u_{t-i}^* + \varepsilon_t^*, \quad t = -m, \ldots, 0, 1, \ldots, n, \\ y_t^* &= x_t(\tilde{\boldsymbol{\beta}}) + u_t^*, \quad t = 1, \ldots, n, \end{aligned}$$

where the $\hat\rho_i$ are the estimated parameters, and the ε_t^* are resampled from the (possibly rescaled) residuals of the AR(p) process. Here m is an arbitrary number, such as 100, chosen so that the process can be allowed to run for some time before the sample period starts. We set the initial values of u_{t-i}^* to zero and discard the u_t^* for $t < 1$.

The sieve bootstrap is somewhat restrictive, because it rules out GARCH models and other forms of conditional heteroskedasticity. Nevertheless, it is quite popular. It has recently been applied to unit root testing in Park (2003) and Chang and Park (2003), and it seems to work quite well in many cases.

Another approach that is more in the spirit of resampling and imposes fewer assumptions is the **block bootstrap**, originally proposed in Künsch (1989). The idea is to divide the quantities that are being resampled into blocks of b consecutive observations. These quantities may be either rescaled residuals or $[\boldsymbol{y}, \boldsymbol{X}]$ pairs. The blocks, which may be either overlapping or nonoverlapping and may be either fixed or variable in length, are then resampled. Theoretical results due to Lahiri (1999) suggest that the best approach is to use overlapping blocks of fixed length. This is called the **moving-block bootstrap**.

For the moving-block bootstrap, there are $n - b + 1$ blocks. The first contains observations 1 through b, the second contains observations 2 through $b + 1$, and the last contains observations $n-b+1$ through n. Each bootstrap sample is then constructed by resampling from these overlapping blocks. Unless n/b is an integer, one or more of the blocks will have to be truncated to form a sample of length n.

A variant of the moving-block bootstrap for dynamic models is the **block-of-blocks** bootstrap (Politis and Romano, 1992), which is analogous to the pairs bootstrap. Consider the dynamic regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \gamma y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

If we define $\boldsymbol{Z}_t$ as $[y_t, y_{t-1}, \boldsymbol{X}_t]$, then we can construct $n-b+1$ overlapping blocks as

$$\boldsymbol{Z}_1 \ldots \boldsymbol{Z}_b; \;\; \boldsymbol{Z}_2 \ldots \boldsymbol{Z}_{b+1}; \;\; \ldots\ldots \;\; ; \boldsymbol{Z}_{n-b+1} \ldots \boldsymbol{Z}_n.$$

These overlapping blocks are then resampled. Note that the block-of-blocks bootstrap can be used with any sort of dynamic model, not just regression models.

Because of the way the blocks overlap in a moving-block bootstrap, not all observations appear with equal frequency in the bootstrap samples. Observations 1 and n each occur in only one block, observations 2 and $n-1$ each occur in only two blocks, and so on. This can have implications for the way one should calculate the bootstrap test statistics. See Horowitz *et al.* (2006).

Block bootstrap methods inevitably suffer from two problems. The first problem is that the bootstrap samples cannot possibly mimic the original sample, because the dependence is broken at the start of each new block; this is often referred to as the join-point problem. The shorter the blocks, the more join points there are. The second problem is that the bootstrap samples look too much like the particular sample we started with. The longer the blocks, the fewer the number of different blocks in any bootstrap sample, and hence the less opportunity there is for each of the bootstrap samples to differ from the original one.

For any block bootstrap method, the choice of b is very important. If b is too small, the join-point problem may well be severe. If b is too large, the bootstrap samples may well be insufficiently random. In theory, b must increase with n, and the rate of increase should often be proportional to $n^{1/3}$. Of course, since actual sample sizes are generally fixed, it is not clear what this means in practice. It often requires quite a large sample size for it to be possible to choose b so that the blocks are neither too short nor too long.

Block bootstrap methods have many variants designed for different types of problem. For example, Paparoditis and Politis (2003) proposes a **residual-based block bootstrap** designed for unit root testing. Andrews (2004) proposes an ingenious **block-block bootstrap** that involves modifying the original test statistic, so that its join-point features resemble those of the bootstrap statistics.

The finite-sample properties of block bootstrap methods are generally not very good. In some cases, it can require substantial effort just to show that they are asymptotically valid; see Gonçalves and White (2004). In theory, these methods frequently offer higher-order accuracy than asymptotic methods, but the rate of improvement is generally quite small; see Hall, Horowitz, and Jing (1995) and Andrews (2002, 2004).

6. Multiple Test Statistics

Whenever we perform two or more tests at the same time, we cannot rely on ordinary P values, or ordinary critical values, because the probability of obtaining an unusually large test statistic by chance increases with the number of tests we perform. However, controlling the overall significance level of several tests without sacrificing power is difficult to do using classical procedures. In order to ensure that the overall significance level is no greater than α, we must use a significance level α' for each test that is smaller than α. If there are m tests, the well-known **Bonferroni inequality** tells us to set $\alpha' = \alpha/m$. A closely related approach is to set $\alpha' = 1-(1-\alpha)^{1/m}$. This will yield a very similar, but slightly less conservative, answer. If we are using P values, the Bonferroni P value is simply m times the smallest P value for each of the individual tests. Thus, when performing five tests, we would need one of them to have a P value of less than .01 before we could reject at the .05 level.

When some of the test statistics are positively correlated, the Bonferroni approach can be much too conservative. In the extreme case in which all of the statistics were perfectly correlated, it would be appropriate just to use α as the significance level for each of the tests. One attractive feature of bootstrap testing is that it is easily adapted to situations in which there is more than one test statistic. Westfall and Young (1993) provides an extensive discussion of how to test multiple hypotheses using bootstrap methods. However, with the recent exception of Godfrey (2005), there seems to have been little work on this subject in econometrics.

Suppose we have m test statistics, τ_1 through τ_m. If these all follow the same distribution under the null hypothesis, then it makes sense to define

$$\tau_{\max} \equiv \max(\tau_l), \quad l = 1, \ldots, m, \tag{20}$$

and treat $\tau_{\max}$ like any other test statistic for the purpose of bootstrapping. We simply need to calculate all m test statistics for the actual sample and for each bootstrap sample and use them to compute $\hat{\tau}_{\max}$ and B realizations of $\tau^*_{\max}$. However, if the τ_l follow different distributions, perhaps because the statistics have different numbers of degrees of freedom, or perhaps because the discrepancies between their finite-sample and asymptotic distributions differ, then the value of $\tau_{\max}$ will be excessively influenced by whichever of the τ_l tend to be most extreme under the null hypothesis.

In most cases, therefore, it make sense to base an overall test on the minimum of the P values of all the individual tests,

$$p_{\min} \equiv \min\bigl(p(\tau_l)\bigr), \quad l = 1, \ldots, m. \tag{21}$$

This can be thought of as a test statistic to be bootstrapped, where we reject when $p_{\min}$ is in the lower tail of the empirical distribution of the $p^*_{\min}$. In equation (21), $p(\tau_l)$

denotes a P value associated with the test statistic τ_l. It could be either an asymptotic P value or a bootstrap P value; the latter may or may not be more reliable. If it is a bootstrap P value, then bootstrapping $p_{\min}$ will require a double bootstrap. Such a double bootstrap scheme has been proposed in Godfrey (2005) for dealing with multiple diagnostic tests for linear regression models.

In many cases, it should not be necessary to use a double bootstrap. If asymptotic P values are reasonably reliable, then using a single bootstrap with $p_{\min}$ can be expected to work well. Whenever the test statistics τ_1 through τ_m have a joint asymptotic distribution that is free of unknown parameters, then $\tau_{\max}$, and by extension $p_{\min}$, must be asymptotically pivotal. Thus bootstrapping either of them will provide an asymptotic refinement. However, this is a stronger condition than assuming that each of τ_1 through τ_m is asymptotically pivotal, and it may not always hold. When it does not, a (single) bootstrap test based on $\tau_{\max}$ or $p_{\min}$ will still be asymptotically valid, but it will not offer any asymptotic refinement. Of course, one can then apply a double bootstrap scheme like the second one discussed in Section 4, and doing so will offer an asymptotic refinement; see Godfrey (2005).

Test statistics like (20) arise naturally when testing for structural change with an unknown break point. Here each of the τ_l is an F, Wald, or LR statistic that tests for a break at a different point in time. Such tests are often called "sup" (short for "supremum") tests. Thus, in the case of a regression model where each of the τ_l is an F statistic, $\tau_{\max}$ is referred to as a **supF** statistic. The number of test statistics is the number of possible break points, which is typically between $0.6n$ and $0.9n$, where n is the sample size. Various authors, notably Andrews (1993), Andrews and Ploberger (1994), and Hansen (1997), have calculated asymptotic critical values for many cases, but these values may not be reliable in finite samples.

It is natural to apply the the bootstrap to this sort of problem. Diebold and Chen (1996) and Hansen (2000) seem to have been among the first papers to do so. The two **fixed regressor bootstrap** procedures suggested in Hansen (2000) are essentially variants of the residual bootstrap and the wild bootstrap. They are called "fixed regressor" because, when the regressors include lagged dependent variables, the latter are treated as fixed rather than generated recursively as in (11). This seems unsatisfactory, but it allows the proposed procedures to work for nonstationary as well as stationary regressors. Note that, since $\tau_{\max}$ in this case has an asymptotic distribution, it is evidently asymptotically pivotal, and so bootstrapping should, in principle, yield an asymptotic refinement.

The supF test for structural change will be examined further in the next section, partly because it provides a simple example of a procedure that is conceptually easier to perform as a bootstrap test than as an asymptotic one, and partly because it illustrates the importance of how the bootstrap DGP is chosen.

Another case in which it would be very natural to use the bootstrap with multiple test statistics is when performing point-optimal tests. Suppose there is some parameter, say θ, that equals θ_0 under the null hypothesis. The idea of a point-optimal test is to construct a test statistic that is optimal for testing $\theta = \theta_0$ against a specific alternative, say $\theta = \theta_1$. Such a test may have substantially more power when θ is in the neighborhood of θ_1 than conventional tests that are locally optimal. Classic papers on point-optimal tests for serial correlation include King (1985) and Dufour and King (1991). Elliott, Rothenberg, and Stock (1996) applies the idea to unit root testing.

The problem with point-optimal tests is that, if the actual value of θ happens to be far from θ_1, the test may not be very powerful. To guard against this, it is natural to calculate several test statistics against different, plausible values of θ. This introduces the usual problems associated with performing multiple tests, but they are easily overcome by using the bootstrap.

In the case of tests for first-order serial correlation in the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ with normal errors, point-optimal tests for ρ_0 against ρ_1 have the form

$$\tau(\rho_1, \rho_0) = \frac{\hat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}_0^{-1} \hat{\boldsymbol{u}}}{\tilde{\boldsymbol{u}}^\top \boldsymbol{\Sigma}_1^{-1} \tilde{\boldsymbol{u}}}, \tag{22}$$

where $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are $n \times n$ positive definite matrices proportional to the covariance matrices of the error terms under the null and alternative hypotheses, respectively, and $\hat{\boldsymbol{u}}$ and $\tilde{\boldsymbol{u}}$ are vectors of GLS residuals corresponding to those two error covariance matrices. When the null hypothesis is that the errors are serially uncorrelated, so that $\rho_0 = 0$, $\boldsymbol{\Sigma}_0$ is an identity matrix, and $\hat{\boldsymbol{u}}$ is a vector of OLS residuals.

Although the test statistic (22) does not follow any standard distribution, it is not hard to show that it is asymptotically pivotal if the distribution of the error terms is known. With fixed regressors as well as error terms that follow a known distribution, it will be exactly pivotal, and Monte Carlo testing will be exact. More generally, since the distribution of the error terms evidently matters for the distribution of $\tau(\rho_1, \rho_0)$, even asymptotically, one would expect bootstrap testing to be asymptotically valid but not to offer any asymptotic refinement. Nevertheless, since the EDF of the residuals should provide a good estimate of the distribution of the error terms in large samples, it seems plausible that bootstrap testing will work reasonably well if the sample size is large enough.

Bootstrapping this sort of test statistic is much more convenient than looking up nonstandard critical values even if we are just performing a single point-optimal test. It is even more convenient when we wish to perform several such tests. For example, if there are m alternative values of ρ, we just need to calculate

$$\hat{\tau}_{\max} \equiv \max\bigl(\hat{\tau}(\rho_1, \rho_0), \hat{\tau}(\rho_2, \rho_0), \ldots, \hat{\tau}(\rho_m, \rho_0)\bigr)$$

and bootstrap it in the usual way. Of course, the bootstrap P value associated with $\hat{\tau}_{\max}$ will inevitably be larger than the one associated with any individual point-optimal test statistic. Thus a test based on $\hat{\tau}_{\max}$ must have less power than one based on $\hat{\tau}(\rho_1, \rho_0)$ for ρ sufficiently close to ρ_1.

Yet another case in which it would be natural to use the bootstrap with multiple test statistics is when testing more than two nonnested hypotheses. Consider a set of nonnested and possibly nonlinear regression models, H_0 through H_m:

$$H_l\colon \quad \boldsymbol{y} = \boldsymbol{x}_l(\boldsymbol{\beta}_l) + \boldsymbol{u}_l, \;\; l = 0, \ldots, m,$$

where $\boldsymbol{y}$ is an n–vector of observations on a dependent variable, and each of the $\boldsymbol{x}_l(\boldsymbol{\beta}_l)$ is a vector of regression functions, which may be linear or nonlinear in $\boldsymbol{\beta}_l$. The object is to test the specification of one of these models by using the evidence obtained by estimating the others. See Davidson and MacKinnon (2004, Chapter 15).

The easiest way to test H_0 against any one of the other models, say H_1, is to use the P test proposed in Davidson and MacKinnon (1981), which reduces to the better-known J test, proposed in the same paper, when H_0 is a linear regression model. The P test is based on the Gauss-Newton regression

$$\boldsymbol{y} - \boldsymbol{x}_0(\hat{\boldsymbol{\beta}}_0) = \boldsymbol{X}_0(\hat{\boldsymbol{\beta}}_0)\boldsymbol{b} + \alpha\big(\boldsymbol{x}_1(\hat{\boldsymbol{\beta}}_1) - \boldsymbol{x}_0(\hat{\boldsymbol{\beta}}_0)\big) + \text{residuals}. \tag{23}$$

Here $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ are vectors of least squares estimates, $\boldsymbol{x}_0(\hat{\boldsymbol{\beta}}_0)$ and $\boldsymbol{x}_1(\hat{\boldsymbol{\beta}}_1)$ are vectors of fitted values, and $\boldsymbol{X}_0(\hat{\boldsymbol{\beta}}_0)$ is a matrix of the derivatives of $\boldsymbol{x}_0(\boldsymbol{\beta}_0)$ with respect to $\boldsymbol{\beta}_0$, evaluated at $\hat{\boldsymbol{\beta}}_0$. The P statistic is simply the t statistic for $\alpha = 0$. Under the assumption of homoskedasticity, this t statistic would be based on the usual OLS covariance matrix for regression (23). Under the weaker assumption of heteroskedasticity of unknown form, it would be based on a heteroskedasticity-consistent covariance matrix.

Under the null hypothesis that the data are generated by H_0 and a suitable assumption about the error terms, either sort of P statistic is asymptotically distributed as standard normal. However, in samples of moderate size, the P statistic is often far from its asymptotic distribution. It generally has a positive expectation, which can be quite large, and, in consequence, it tends to overreject severely. Thus, in a great many cases, it is not safe to compare the P statistic to the standard normal distribution.

However, at least for homoskedastic errors, bootstrapping the P statistic using the residual bootstrap generally works very well. Davidson and MacKinnon (2002a) provides a detailed analysis for the case in which H_0 and H_1 are both linear models. Even when the asymptotic test rejects more than half the time at the .05 level, the bootstrap test typically overrejects quite modestly. There are some, generally unrealistic, cases in which the ordinary bootstrap test performs much better than the asymptotic test but still overrejects noticeably. In such cases, using the fast double bootstrap generally results in a substantial further improvement, so that the overrejection often becomes negligible; see Davidson and MacKinnon (2002b).

It is customary to test H_0 separately against each of H_1 through H_m, although one could also test H_0 against all the other models jointly by generalizing (23) in the obvious way and using an F statistic with m numerator degrees of freedom. When we compute m separate P statistics, there is the usual problem of how to make inferences. If we denote the m test statistics by $\hat{\tau}_1$ through $\hat{\tau}_m$, we could simply calculate $\hat{\tau}_{\max}$

as in (20), but this is probably not a good thing to do, because the τ_l may have quite different finite-sample distributions under the null hypothesis.

It would be better to use a double bootstrap, in which we first assign a bootstrap P value $\hat{p}^*(\hat{\tau}_l)$ to each of the $\hat{\tau}_l$. This should be an equal-tail P value based on equation (4) if we wish to perform a two-tailed test, because the distribution of the τ_l is often far from being symmetric around the origin. We would then bootstrap the overall test statistic

$$\hat{p}^*_{\min} = \min\bigl(\hat{p}^*(\hat{\tau}_l)\bigr), \quad l = 1, \ldots, m.$$

Under the assumption of homoskedasticity, it would be appropriate to employ the ordinary t statistic for $\alpha = 0$ in (23) and the residual bootstrap. Under the weaker assumption of heteroskedasticity of unknown form, it would be appropriate to employ a heteroskedasticity-robust t statistic and the wild bootstrap.

7. Finite-Sample Properties of Bootstrap SupF Tests

Although bootstrap tests can work very well indeed, they certainly do not always perform particularly well. Moreover, the precise form of the bootstrap DGP is often very important. In this section, these two points are illustrated by examining the performance of several bootstrap tests based on the supF statistic introduced in the previous section.

The supF statistic may be calculated whenever we have a linear regression model estimated using time series data. Such a model may be written as

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

where $\boldsymbol{X}_t$ is a $1 \times k$ vector of observations on regressors that may include lagged dependent variables. To calculate the statistic, we first estimate this model by OLS to obtain the sum of squared residuals, or SSR. Let π be a number greater than 0 and, in most cases, substantially less than 0.50 (typical values are 0.10, 0.15, and 0.20), and let n_1 be the integer closest to πn. We next run the regression again for every pair of subsamples

$$\begin{array}{rcl} 1, \ldots, n_1 & \text{and} & n_1 + 1, \ldots, n; \\ 1, \ldots, n_1 + 1 & \text{and} & n_1 + 2, \ldots, n; \\ & \cdots\cdots\cdots & \\ 1, \ldots, n - n_1 & \text{and} & n - n_1 + 1, \ldots, n. \end{array}$$

Let the sum of squared residuals from the first of each pair of regressions be denoted SSR_1, and the one from the second be denoted SSR_2. Then, for the l^{th} pair of subsample regressions, we can calculate the usual F, or Chow, statistic for the coefficients to be the same in both subsamples:

$$F_l(k, n - 2k) = \frac{(\text{SSR} - \text{SSR}_1 - \text{SSR}_2)/k}{(\text{SSR}_1 + \text{SSR}_2)/(n - 2k)}. \tag{24}$$

The supF statistic is just the maximum of the statistics (24) over the $n-2n_1+1$ pairs of subsamples. The asymptotic distribution of this statistic depends on k and π; see Andrews (1993) and Hansen (1997). The programs from Hansen (1997) are actually in terms of k times supF, since the same asymptotic distribution applies to a much wider class of tests for structural stability than just the supF test, and the test statistic is, in general, the maximum of $n-2n_1+1$ statistics that each have an asymptotic $\chi^2(k)$ distribution.

Results for three sets of simulation experiments are reported. In the first set, there are no lagged dependent variables. The vector $\boldsymbol{X}_t$ simply consists of a constant term and $k-1$ independent, normally distributed random variables, and the error term u_t is also IID normal. Of course, this is not a very realistic model for time-series data, but it makes it easy to focus on the role of k, the number of regressors. There are four sets of experiments, for sample sizes $n = 50$, 100, 200, and 400. In all of them, $\pi = 0.15$. The number of regressors, k, varies from 2 to 7 when $n = 50$ and from 2 to 12 for the remaining values of n. Note that, when $n = 50$, $\pi n = 7.5$, and therefore $n_1 = 8$. This is why $k \leq 7$ when $n = 50$.

Four different tests are investigated. The first is an asymptotic test based on the approximate asymptotic distribution functions of Hansen (1997). The others are bootstrap tests using the residual bootstrap, as in equation (10), the wild bootstrap (Davidson-Flachaire version), and the pairs bootstrap, respectively. Note that, for the supF test, the pairs bootstrap does impose the null hypothesis, because all the $[y_t, \boldsymbol{X}_t]$ pairs are implicitly assumed to come from the same joint distribution, and so we do not need to modify the test statistic.

Every experiment has 100,000 replications, and all bootstrap methods use $B = 199$. This is a much smaller value of B than would normally be used in practice, but errors tend to cancel out across replications in a simulation experiment like this one. The value of B is small because these experiments are very expensive, especially for the larger sample sizes. Each replication involves running $(B+1)(2n-4n_1+3)$ linear regressions.

The results of the experiments are shown in Figure 1. There are four panels, which correspond to the four different sample sizes. Note that the vertical axis in the two top panels, for $n = 50$ and $n = 100$, is not the same as in the two bottom panels, for $n = 200$ and $n = 400$. This choice was made because, as expected, all the tests perform better for larger sample sizes.

Although the asymptotic test works very well for $k = 2$, it overrejects more and more severely as k increases. In contrast, the residual bootstrap seems to perform very well for all values of k, and the wild bootstrap underrejects just a little for large values of k and small values of n. The pairs bootstrap performs very differently from, but no better than, the asymptotic test; it underrejects more and more severely as k increases. In this case, since the errors are IID, there is no need to use either the wild or the pairs bootstrap. In practice, however, one might well want to guard against heteroskedasticity by using one of these methods. The cost of using the wild bootstrap appears to be modest, but the pairs bootstrap should always be avoided. Of course, it would also be desirable to use a heteroskedasticity-robust test statistic if heteroskedasticity were suspected.

In the second set of experiments, there are only two regressors, a constant term and a lagged dependent variable. Thus the model is really the AR(1) process

$$y_t = \beta_1 + \beta_2 y_{t-1} + u_t, \quad u_t \sim \mathrm{NID}(0, \sigma^2). \tag{25}$$

There are 19 cases for each sample size, corresponding to the following values of β_2: $-0.9, -0.8, \ldots, 0.8, 0.9$. All results are invariant to the values of β_1 and σ.

This time, five different tests are investigated. Once again, the first is an asymptotic test, and the others are bootstrap tests. There are two versions of the residual bootstrap. One is the fixed regressor bootstrap of Hansen (2000), which is calculated exactly the same way as the residual bootstrap in the first experiment, incorrectly treating the $\boldsymbol{X}_t$ as fixed. The other is a recursive residual bootstrap, in which the model is assumed to be known and estimates of β_1 and β_2 from the entire sample are used, along with resampled rescaled residuals, to generate the bootstrap samples. This is the type of bootstrap that was studied in Diebold and Chen (1996) for the same model. The value of $|\beta_2|$ is constrained to be less than 0.99 to ensure stationarity. When $\hat{\beta}_2$ is greater than 0.99 or less than -0.99, it is set to 0.99 or -0.99, as appropriate. This happened just twice in all the experiments when $n = 100$ and $\beta_2 = -0.9$, and a few hundred times when $n = 50$ and $|\beta_2| = 0.9$.

The other two bootstrap methods are block-of-blocks variants of the moving block bootstrap. These are essentially generalizations of the pairs bootstrap. The length of the blocks is set to the smallest integer greater than $n^{1/3}$ in one case, and to twice that in the other. The actual block lengths are indicated in Figure 2, which shows rejection frequencies as a function of β_2 for all five tests.

From Figure 2, we can see that the performance of all the tests is sensitive to the value of β_2. For the largest sample sizes, the recursive residual bootstrap performs quite well for all values of β_2. For the smaller sample sizes, it underrejects slightly for negative and small positive values of β_2, and it overrejects noticeably for large positive values. None of the other methods performs at all well. They all underreject for most values of β_2 and overreject, often severely, for positive ones that are sufficiently large. The fixed regressor bootstrap performs almost the same as the asymptotic test. The two moving block bootstraps always reject less often than the asymptotic test, both when it underrejects and when it overrejects. This is more true for the variant that uses longer blocks, which overrejects only for the largest positive values of β_2.

Notice that every one of the rejection frequency curves in Figure 2 is below 0.05 for some values of β_2 and above it for others. Thus we can always find a value of β_2 for which any of the tests, whether bootstrap or asymptotic, happens to work perfectly. Nevertheless, with the exception of the recursive residual bootstrap, it is fair to say that, overall, none of these methods works well.

Although the recursive residual bootstrap works reasonably well, it is interesting to see whether it can be improved upon, especially for extreme values of β_2. Two variants of this procedure are therefore examined. The first is the fast double bootstrap, or FDB, which was discussed in Section 4. This method is easy to implement, although the probability of obtaining estimates of β_2 greater than 0.99 in absolute value is much greater than for the single bootstrap, because of the random variation in the values of β_2 used in the second-level bootstrap DGPs. All such estimates were replaced by either 0.99 or -0.99, as appropriate.

The third method that is studied is cruder but less computationally intensive than the FDB. One obvious problem with the recursive residual bootstrap is that the OLS estimate of β_2 is biased. A simple way to reduce this bias is to generate B bootstrap samples, calculate the average value $\bar{\beta}_2^*$ over them, and then obtain the estimator

$$\hat{\beta}_2' \equiv \hat{\beta}_2 - (\bar{\beta}_2^* - \hat{\beta}_2) = 2\hat{\beta}_2 - \bar{\beta}_2^*.$$

The term in parentheses here is an estimate of the bias; see MacKinnon and Smith (1998). The idea is simply to use $\hat{\beta}_2'$ instead of $\hat{\beta}_2$ in the bootstrap DGP. This sort of bias correction has been used successfully when bootstrapping confidence intervals for impulse response functions from vector autoregressions; see Kilian (1998). When $|\hat{\beta}_2'| > 0.99$, something which occurs not infrequently when the absolute value of β_2 is large, $\hat{\beta}_2'$ is replaced by a value halfway between $\hat{\beta}_2$ and either 1 or -1. This value is in turn replaced by 0.99 or -0.99, if necessary.

In the third set of experiments, the model is once again (25), but there are now 23 values of β_2, with -0.95, -0.85, 0.85, and 0.95 added to the values considered previously so as to obtain more information about what happens near the edge of the stationarity region. Because the recursive residual bootstrap works very well indeed for $n = 400$, the sample sizes are now 25, 50, 100, and 200. Also, because the FDB suffers from size distortions when B is small, B is now 999 instead of 199.

Figure 3 shows rejection frequencies for the recursive residual bootstrap test and the bias-corrected and FDB variants of it as functions of β_2. It is clear that bias correction, at least as it is implemented here, generally does not improve the performance of the recursive residual bootstrap test. In contrast, the FDB typically does improve its performance. For the smaller sample sizes, this improvement is noticeable for most values of β_2. For the larger sample sizes, it is noticeable only for large, positive values of β_2, the values for which both bootstrap and asymptotic tests perform worst.

8. Conclusion

Tests based on asymptotic theory can be difficult to implement and do not always work well in finite samples. In such cases, bootstrap testing may be very attractive. There are at least four ways to compute bootstrap P values. We may wish to reject in the upper tail, the lower tail, or both tails. In the latter case, we may or may not wish to impose the assumption that the distribution of the test statistic is symmetric around the origin. Equations (3) and (4) are both perfectly valid ways of computing P values for two-tailed tests, but they may yield quite different results.

In certain circumstances, which generally involve rather strong distributional assumptions, bootstrap tests can actually be exact, and they are called Monte Carlo tests. However, the idea that bootstrap tests always perform well is quite false, as the results in Figures 1 and 2 illustrate. Procedures such as the double bootstrap and fast double bootstrap may help matters, but this is by no means guaranteed.

More importantly, there are often many ways to specify a bootstrap DGP. Whether a bootstrap test works well or badly depends on how well the bootstrap DGP mimics the essential features of the true DGP under the null hypothesis. In general, it seems that bootstrap procedures which are at least partly parametric perform better than ones that are fully nonparametric. That is certainly true for the supF test studied in Section 7. The parameters should always be estimated under the null hypothesis, if possible. Of course, we would not expect a parametric bootstrap to work well if the rejection probabilities depended strongly on one or more nuisance parameters and those parameters were hard to estimate reliably.

One of the big advantages of bootstrap testing is that it can be used in situations where asymptotic theory is unavailable or difficult to employ. For example, it can easily be used to assign a P value to the maximum of a possibly large number of test statistics. Examples where this may be useful include nonnested hypothesis tests when there are several alternative hypotheses, point-optimal tests, and tests for structural change.

References

Abadie, A., and G. W. Imbens (2006). "On the failure of the bootstrap for matching estimators," NBER Working Paper No. T0325.

Andrews, D. W. K. (1993). "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821–856.

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

Andrews, D. W. K. (2004). "The block-block bootstrap: Improved asymptotic refinements," *Econometrica*, 72, 673–700.

Andrews, D. W. K. and W. Ploberger (1994). "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica*, 62, 1383–1414.

Beaulieu, M.-C., Dufour, J.-M., and L. Khalaf (2007). "Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: An exact simulation-based approach," *Journal of Business and Economic Statistics*, **25**, 398–410.

Beran, R. (1987). "Prepivoting to reduce level error in confidence sets," *Biometrika*, 74, 457–468.

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

Breusch, T. S., and A. R. Pagan (1979). "A simple test for heteroskedasticity and random coefficient variation," *Econometrica*, 47, 1287–1294.

Bühlmann, P. (1997). "Sieve bootstrap for time series," *Bernoulli*, 3, 123–148.

Bühlmann, P. (2002). "Bootstraps for time series," *Statistical Science*, 17, 52–72.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, **90**, 414–427.

Chang, Y., and J. Y. Park (2003). "A sieve bootstrap for the test of a unit root," *Journal of Time Series Analysis*, 24, 379–400.

Davidson, R., and E. Flachaire (2008). "The wild bootstrap, tamed at last," *Journal of Econometrics*, **146**, 162–169.

Davidson, R., and J. G. MacKinnon (1981). "Several tests for model specification in the presence of alternative hypotheses," *Econometrica*, 49, 781–793.

Davidson, R., and J. G. MacKinnon (1999). "The size distortion of bootstrap tests," *Econometric Theory*, 15, 361–376.

Davidson, R., and J. G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?" *Econometric Reviews*, 19, 55–68.

Davidson, R., and J. G. MacKinnon (2002a). "Bootstrap J tests of nonnested linear regression models," *Journal of Econometrics*, 109, 167–193.

Davidson, R., and J. G. MacKinnon (2002b). "Fast double bootstrap tests of nonnested linear regression models," *Econometric Reviews*, 21, 417–427.

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

Davidson, R., and J. G. MacKinnon (2006a). "The power of bootstrap and asymptotic tests," *Journal of Econometrics*, 133, 421–441.

Davidson, R., and J. G. MacKinnon (2006b). "Bootstrap methods in econometrics," Chapter 23 in *Palgrave Handbook of Econometrics: Volume 1 Theoretical Econometrics*, ed. K. Patterson and T. C. Mills, Basingstoke, Palgrave Macmillan, 812–838.

Davidson, R., and J. G. MacKinnon (2007). "Improving the reliability of bootstrap tests with the fast double bootstrap," *Computational Statistics and Data Analysis*, 51, 3259–3281.

Davidson, R., and J. G. MacKinnon (2008). "Bootstrap inference in a linear equation estimated by instrumental variables," *Ecoonometrics Journal*, 11, 443–477.

Davidson, R., and J. G. MacKinnon (2009). "Wild bootstrap tests for IV regression," *Journal of Business and Economic Statistics*, 27, forthcoming.

Davison, A. C., and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*, Cambridge, Cambridge University Press.

Davison, A. C., D. V. Hinkley, and G. A. Young (2003). "Recent developments in bootstrap methodology," *Statistical Science*, 18, 141–157.

DiCiccio, T. J., and B. Efron (1996). "Bootstrap confidence intervals" (with discussion), *Statistical Science*, 11, 189–228.

Diebold, F. X., and C. Chen (1996). "Testing structural stability with endogenous breakpoint: A size comparison of bootstrap and analytic procedures," *Journal of Econometrics*, 70, 221–241.

Dufour, J.-M. (2006). "Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics," *Journal of Econometrics*, 133, 443–477.

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

Dufour, J.-M., and L. Khalaf (2002). "Simulation based finite and large sample tests in multivariate regressions," *Journal of Econometrics*, 111, 303–322.

Dufour, J.-M., L. Khalaf, J.-T. Bernard, and I. Genest (2004). "Simulation-based finite-sample tests for heteroskedasticity and ARCH effects," *Journal of Econometrics*, 122, 317–347.

Dufour, J.-M., and M. L King (1991). "Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or nonstationary AR(1) errors," *Journal of Econometrics*, 47, 115–143.

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

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

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.

Elliott, G., T. J. Rothenberg, and J. H. Stock (1996). "Efficient tests for an autoregressive unit root," *Econometrica*, 64, 813–836.

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

Freedman, D. A. (1984). "On bootstrapping two-stage least-squares estimates in stationary linear models," *Annals of Statistics*, 12, 827–842.

Godfrey, L. G. (2005). "Controlling the overall significance level of a battery of least squares diagnostic tests," *Oxford Bulletin of Economics and Statistics*, 67, 263–279.

Godfrey, L. G., C. D. Orme, and J. M. Santos Silva (2005). "Simulation-based tests for heteroskedasticity in linear regression models: Some further results," *Econometrics Journal*, 9, 76–97.

Gonçalves, S., and L. Kilian (2004). "Bootstrapping autoregressions with heteroskedasticity of unknown form," *Journal of Econometrics*, 123, 89–120.

Gonçalves, S., and H. White (2004). "Maximum likelihood and the bootstrap for dynamic nonlinear models," *Journal of Econometrics*, 119, 199–219.

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

Hall, P., J. L. Horowitz, and B. Y. Jing (1995). "On blocking rules for the bootstrap with dependent data," *Biometrika*, 82, 561–574.

Hansen, B. E. (1997). "Approximate asymptotic P values for structural change tests," *Journal of Business and Economic Statistics*, 15, 60–67.

Hansen, B. E. (2000). "Testing for structural change in conditional models," *Journal of Econometrics*, 97, 93–115.

Härdle, W., J. L. Horowitz, and J.-P. Kreiss (2003). "Bootstrap methods for time series," *International Statistical Review*, 71, 435–459.

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

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

Horowitz, J. L., I. L. Lobato, J. C. Nankervis, and N. E. Savin (2006). "Bootstrapping the Box-Pierce Q test: A robust test of uncorrelatedness," *Journal of Econometrics*, 133, 841–862.

Jöckel, K.-H. (1986). "Finite sample properties and asymptotic efficiency of Monte Carlo tests," *Annals of Statistics*, 14, 336–347.

Kilian, L. (1998). "Small sample confidence intervals for impulse response functions," *Review of Economics and Statistics*, 80, 218–230.

King, M. L. (1985). "A point optimal test for autoregressive disturbances," *Journal of Econometrics*, 27, 21–37.

Koenker, R. (1981). "A note on Studentizing a test for heteroskedasticity," *Journal of Econometrics*, 17, 107–112.

Künsch, H. R. (1989). "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217–1241.

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

MacKinnon, J. G. (2002). "Bootstrap inference in econometrics," *Canadian Journal of Economics*, 35, 615–645.

MacKinnon, J. G., and A. A. Smith (1998). "Approximate bias correction in econometrics," *Journal of Econometrics*, 85, 205–230.

Martin, M. A. (2007). "Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties," *Computational Statistics and Data Analysis*, **51**, 6321–6342.

Nankervis, J. C. (2005). "Stopping rules for double bootstrap tests," University of Essex, working paper.

Paparoditis, E., and D. N. Politis (2003). "Residual based block bootstrap for unit root testing," *Econometrica*, 71, 813–855.

Park, J. Y. (2003). "Bootstrap unit root tests," *Econometrica*, 71, 1845–1895.

Politis, D. N., and J. P. Romano (1992). "General resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation," *Annals of Statistics*, 20, 1985–2007.

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

Racine, J. S., and J. G. MacKinnon (2007a). "Simulation-based tests that can use any number of simulations," *Communications in Statistics: Simulation and Computation*, **36**, 357–365.

Racine, J. S., and J. G. MacKinnon (2007b). "Inference via kernel smoothing of bootstrap P values," *Computational Statistics and Data Analysis*, **51**, 5949–5957.

Richard, P. (2007). *Sieve Bootstrap Unit Root Tests*, Ph.D. Thesis, Department of Economics, McGill University.

Stewart, K. G. (1997). "Exact testing in multivariate regression," *Econometric Reviews*, 16, 321–352.

Westfall, P. H., and S. Young (1993). *Resampling-Based Multiple Testing*, New York, Wiley.

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

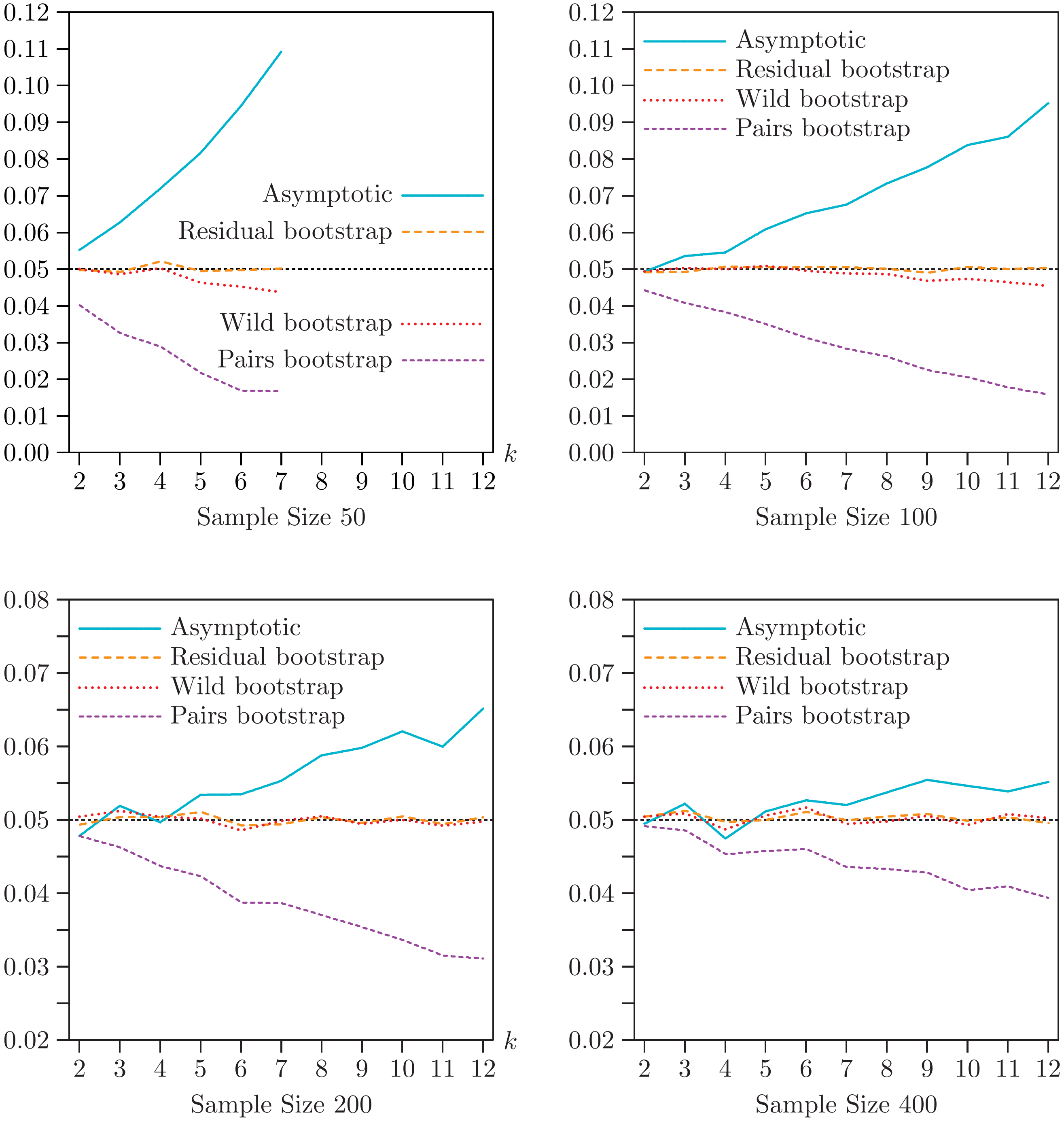

Figure 1. Rejection frequencies for bootstrap and asymptotic tests, static regression model

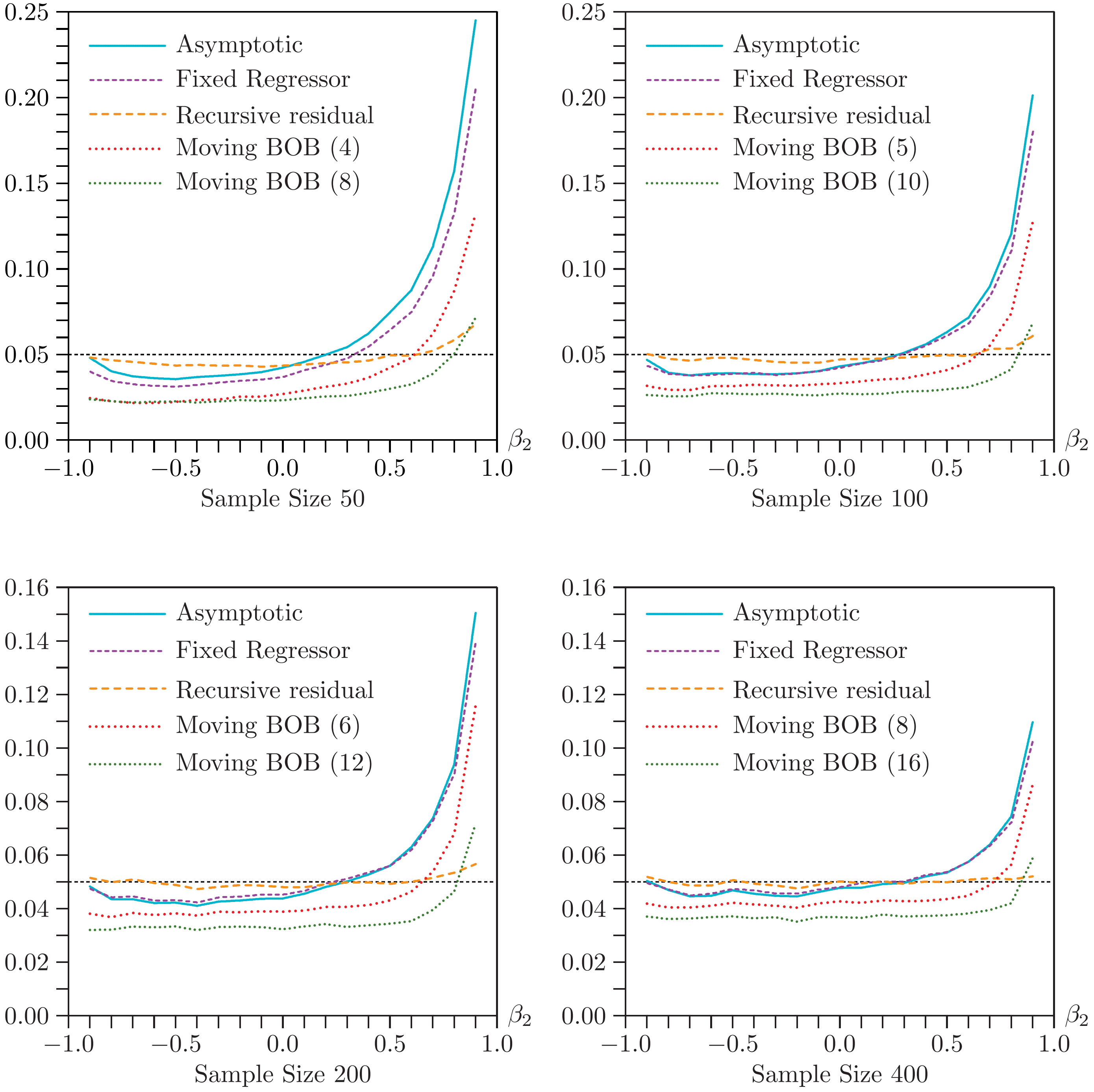

Figure 2. Rejection frequencies for bootstrap and asymptotic tests, AR(1) model

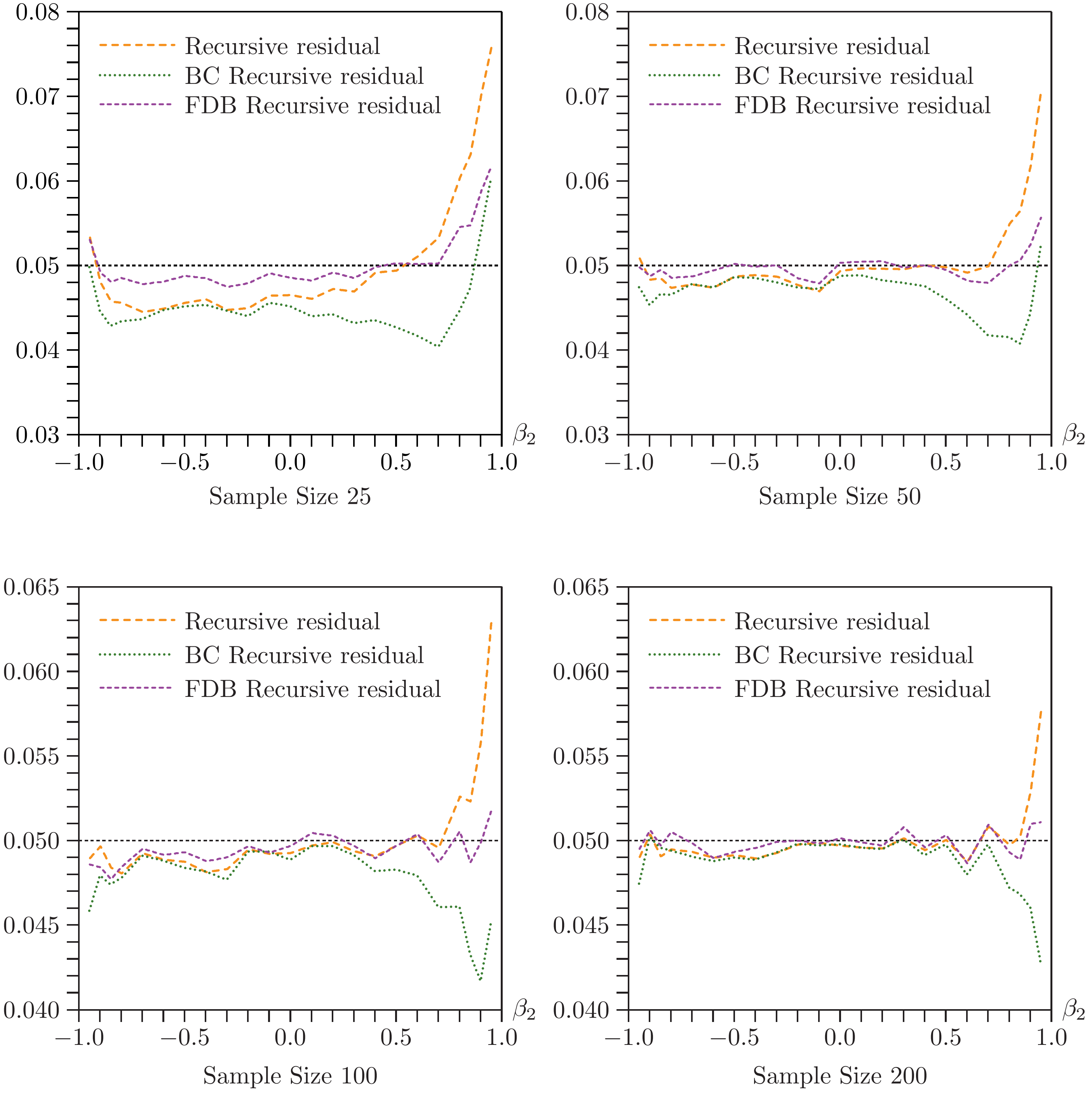

Figure 3. Rejection frequencies for bootstrap tests, AR(1) model