The Size Distortion of Bootstrap Tests

by

Russell Davidson

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

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

russell@ehess.cnrs-mrs.fr

and

James G. MacKinnon

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

jgm@qed.econ.queensu.ca

Abstract

We provide a theoretical framework in which to study the accuracy of bootstrap P values, which may be based on a parametric or nonparametric bootstrap. In the parametric case, the accuracy of a bootstrap test will depend on the shape of what we call the critical value function. We show that, in many circumstances, the error in rejection probability of a bootstrap test will be one whole order of magnitude smaller than that of the corresponding asymptotic test. We also propose a simulation method for estimating this error that requires the calculation of only two test statistics per replication.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. Earlier versions were presented at Universidad Carlos III de Madrid, Universidad Complutense de Madrid, Cambridge University, INSEE-CREST (Paris), CORE (Louvain-la-Neuve), the Tinbergen Institute (Amsterdam), the University of Geneva, the European University Institute (Florence), the ESRC Econometrics Conference (Bristol), the 1996 Berkeley Symposium on the Bootstrap, the Canadian Econometric Study Group, Queen's University, Laval University, and the University of Texas (Austin). We are grateful to many seminar participants and to three anonymous referees for comments. We are especially grateful to Joel Horowitz, not only for comments, but also for his probing questions that led us to clarify the paper. The first draft of the paper was written while the second author was visiting GREQAM.

June, 1998

1. Introduction

Although testing hypotheses is a central concern of econometrics, the distributions of the most frequently used test statistics are known only asymptotically. Thus inference on the basis of them can be risky. There are two approaches to solving this problem. The first is to modify a test statistic analytically so that it approaches its asymptotic distribution more rapidly, as in Attfield (1995), or to modify the critical values so that the true rejection probability of the test approaches its nominal value more rapidly, as in Rothenberg (1984). This approach often requires algebraic derivations that are very far from trivial, and in many cases it seems to be infeasible.

The second approach, which is becoming popular because of the dramatic increase in the speeds of computers in recent years, is to employ some form of bootstrap test. The basic idea of bootstrap testing is to draw a large number of "bootstrap samples" from a distribution which obeys the null hypothesis and is constructed in such a way that the bootstrap samples, as far as possible, resemble the real sample. The test statistics from the bootstrap samples can then be used to calculate a P value for the observed test statistic. Beran (1986) and Hall and Titterington (1989) were among the first papers to study bootstrap testing formally.

There has been some discussion of bootstrap testing methodology in the econometrics literature. Notable papers include Horowitz (1994), who applies the bootstrap to information matrix tests, and Hall and Horowitz (1996), in which bootstrap methods are described for dynamic GMM models. These and other studies have generally found that bootstrap tests are much more reliable than asymptotic tests. If the sample size is n, the error in rejection probability committed by an asymptotic test is, in general, of order $n^{-1/2}$ for a one-tailed test and of order n^{-1} for a two-tailed test. Use of the bootstrap can reduce this order by a factor of $n^{-1/2}$, n^{-1}, or even more; see Hall (1992) for a very full discussion, based on Edgeworth expansions, of the extent to which asymptotic refinements are available in different contexts.

Even when it is known theoretically that the bootstrap provides refinements of a given order, the size distortion of a bootstrap test may vary considerably according to circumstances. We use the term "size distortion" here in the sense in which econometricians would usually understand it, to mean the difference between the nominal level of the test and its actual rejection probability. Since this terminology is not standard outside econometrics, however, we will henceforth speak of the "error in rejection probability," or ERP, of a test. In this paper, we analyze the ERP of bootstrap tests in terms of what we call "critical value functions" and "rejection probability functions." We show how the slope and curvature of these functions determine the extent of the ERP of bootstrap tests. In the parametric case, graphs of the functions can depict clearly those regions in the parameter space where the bootstrap will behave more or less well. Thus inspection of such graphs can yield valuable intuition concerning the performance of bootstrap tests.

Our theoretical framework treats all test statistics in approximate P value form. This device allows us to abstract from the variety of possible asymptotic distributions that test statistics may have, and thereby provides a single theoretical framework in which

the determinants of the bootstrap refinements can be studied. This framework yields two major results. First, it shows that extra refinements, known to be available in a variety of specific cases that are not obviously related, are more generally available if the bootstrapped test statistic is asymptotically independent of the bootstrap distribution. This can almost always be achieved in econometric applications. Second, it provides a new way to estimate by simulation the ERP of bootstrap tests that obey the asymptotic independence condition. This method requires the calculation of only two test statistics per replication, and it is thus very much cheaper than doing a Monte Carlo experiment in which every replication involves a bootstrap test.

In the next section, we explain our terminology and notation and introduce several important concepts. The principal theoretical result of the paper is proved in Section 3, and the implications of this result are explored in Section 4. A simulation method for estimating the ERP of bootstrap tests is proposed in Section 5, and we provide evidence that it can perform very well indeed.

2. Basic Concepts and Notation

In this paper, a model $\mathbb{M}$ will be a set of data-generating processes (DGPs) for one or more dependent variables. By the term DGP we mean a complete stochastic specification, with enough information that it provides a unique recipe for simulation for any specified sample size. A fully parametric model is one for which all the DGPs that belong to it are in one-one correspondence with the elements of a finite-dimensional parameter space. When a model is not fully parametric, many DGPs correspond to each point in the parameter space, differing with regard to things like the distribution of the error terms. For instance, a regression model with normal errors is fully parametric, while the model given by the same regression equation, but with errors that are merely specified to be IID with mean zero and finite variance, is not. A generic element, or DGP, of a model $\mathbb{M}$ will be denoted as μ.

A test statistic τ, computed as a function of data generated by some DGP, is said to be (asymptotically) pivotal for a model $\mathbb{M}$ if the (asymptotic) distribution of τ is the same for each DGP $\mu \in \mathbb{M}$. For $\mathbb{M}$ to represent the null hypothesis tested by τ, it is usually necessary that τ be asymptotically pivotal for $\mathbb{M}$. Any statistic for which the asymptotic distribution is known and the same for all $\mu \in \mathbb{M}$ is automatically asymptotically pivotal. The importance of pivotalness in bootstrap methodology has been insisted on by many authors. Hall (1992) gives extensive bibliographical notes on this point; see especially the ends of his Chapters 1 and 3.

In this paper, we deal only with statistics that can be approximated with error at most $O(n^{-2})$ by a smooth function of sample moments. This assumption puts us in the framework of the "smooth function model" considered by Hall (1992). Regularity conditions are not our main concern here, and so we choose to avoid explicit reference to them by remaining throughout in this framework, in order to rely on Hall's regularity conditions.

Let the model $\mathbb{M}$ represent the null hypothesis we wish to test, and let τ be a test statistic that is asymptotically pivotal for $\mathbb{M}$, where the test rejects for large values of the statistic. We denote by $\hat{\tau}$ the realization of τ calculated from data generated by some unknown DGP μ_0. If μ_0 satisfies the null hypothesis, that is, if $\mu_0 \in \mathbb{M}$, the ideal P value for inference based on $\hat{\tau}$ is

$$p(\hat{\tau}) \equiv \Pr_{\mu_0}(\tau \geq \hat{\tau}). \tag{1}$$

The distribution of the random variable $p(\tau)$, of which $p(\hat{\tau})$ is a realization, is uniform on $[0, 1]$. But, since μ_0 is unknown, we cannot compute (1). The "bootstrap P value" $p^*(\hat{\tau})$ is the bootstrap estimate of $p(\hat{\tau})$. It is obtained by replacing μ_0 in (1) by a bootstrap DGP, $\hat{\mu}$. Thus

$$p^*(\hat{\tau}) \equiv \Pr_{\hat{\mu}}(\tau \geq \hat{\tau}). \tag{2}$$

The bootstrap DGP may be either parametric or nonparametric. In the former case, $\mathbb{M}$ must be a fully parametric model, and $\hat{\mu}$ belongs to $\mathbb{M}$ itself, being defined in terms of estimates of the parameters of $\mathbb{M}$. In the latter case, $\mathbb{M}$ need not be fully parametric, and $\hat{\mu}$ will be based, at least in part, on some sort of empirical distribution function, although it often requires parameter estimates as well. In either case, the bootstrap DGP $\hat{\mu}$ will depend on the data used to obtain $\hat{\tau}$. In practice, $p^*(\hat{\tau})$ usually cannot be computed analytically. Instead, a large number of bootstrap samples is drawn from $\hat{\mu}$, and the proportion of the bootstrap samples for which the bootstrap statistic is no smaller than $\hat{\tau}$ is used to estimate $p^*(\hat{\tau})$.

If τ is an exact pivot, it is well known that, in the case of the parametric bootstrap, $p^*(\hat{\tau}) = p(\hat{\tau})$. This follows immediately from the facts that $\hat{\mu} \in \mathbb{M}$, and so satisfies the null hypothesis, and that the distribution of τ is invariant in $\mathbb{M}$. When τ is only asymptotically pivotal, the distribution of τ depends on $\mu \in \mathbb{M}$, but the asymptotic distribution does not. Let this asymptotic distribution have CDF F. At nominal level α, the test rejects if the asymptotic P value $1 - F(\hat{\tau}) < \alpha$. We introduce the "rejection probability function," or RPF, as a measure of the true rejection probability:

$$R(\alpha, \mu) \equiv \Pr_{\mu}\big(1 - F(\tau) < \alpha\big). \tag{3}$$

This function gives the (true) rejection probability under μ of a test at nominal level α. Clearly, it also depends on the sample size n, but it avoids notational clutter to leave that dependence implicit. It can be seen that $R(\cdot, \mu)$ is the CDF of the asymptotic P value under μ. The ERP of the asymptotic test at nominal level α is defined implicitly by the equation

$$R(\alpha, \mu) = \alpha + n^{-l/2} r(\alpha, \mu), \tag{4}$$

where $l \geq 1$ is defined so that $r(\alpha, \mu)$ will be $O(1)$. Results discussed in Hall (1992) imply that the value of l will be different in different cases. In the case of a one-sided test based on an asymptotically $N(0, 1)$ statistic, $l = 1$. In the case of a two-sided test based on an asymptotically $N(0, 1)$ statistic and the case of an asymptotically χ^2 statistic, $l = 2$.

If each DGP $\mu \in \mathbb{M}$ is fully characterized by a parameter vector $\boldsymbol{\theta}$, the rejection probability function can be written as $R(\alpha, \boldsymbol{\theta})$. Suppose that a data set actually generated by a DGP characterized by $\boldsymbol{\theta}_0$ yields estimates $\hat{\boldsymbol{\theta}}$. Then the parametric bootstrap DGP is characterized by $\hat{\boldsymbol{\theta}}$. The probability of rejection by the asymptotic test can be calculated under the DGPs corresponding to both $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. From (3) and (4), the difference between the two probabilities is

$$R(\alpha, \hat{\boldsymbol{\theta}}) - R(\alpha, \boldsymbol{\theta}_0) = n^{-l/2}\big(r(\alpha, \hat{\boldsymbol{\theta}}) - r(\alpha, \boldsymbol{\theta}_0)\big). \tag{5}$$

If r is differentiable, it can be Taylor expanded around $\boldsymbol{\theta}_0$ to obtain

$$R(\alpha, \hat{\boldsymbol{\theta}}) - R(\alpha, \boldsymbol{\theta}_0) \stackrel{a}{=} n^{-l/2} \boldsymbol{r}_{\boldsymbol{\theta}}^{\top}(\alpha, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{6}$$

where $\boldsymbol{r}_{\boldsymbol{\theta}}(\alpha, \boldsymbol{\theta}_0)$ is the vector of first derivatives of $r(\alpha, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}_0$. If $\hat{\boldsymbol{\theta}}$ is root-n consistent, the right-hand side of (6) is of order $n^{-(l+1)/2}$. Thus the bootstrap approximation to the distribution of $\hat{\tau}$ is in error only at order $n^{-(l+1)/2}$, better than the error of the asymptotic distribution by a factor of $n^{-1/2}$. To the best of our knowledge, this analysis first appeared, in a more general form not limited to the parametric bootstrap, in Beran (1988). See also Hall (1992).

The information contained in the function $R(\alpha, \mu)$ is also provided by the "critical value function," or CVF, $Q(\alpha, \mu)$, defined implicitly by the equation

$$\Pr_{\mu}\big(\tau \geq Q(\alpha, \mu)\big) = \alpha. \tag{7}$$

$Q(\alpha, \mu)$ is thus the critical value for τ which gives a rejection probability of α if the DGP is μ. The relation between the RPF $R(\alpha, \mu)$ and the CVF $Q(\alpha, \mu)$ is

$$R\big(1 - F(Q(\alpha, \mu)), \mu\big) = \alpha, \tag{8}$$

and the rejection region for the bootstrap test of nominal level α is

$$\tau \geq Q(\alpha, \hat{\mu}). \tag{9}$$

Therefore, the rejection probability of the bootstrap test under any DGP μ is the probability of the event (9) under μ.

In the parametric case, a DGP $\mu \in \mathbb{M}$ is completely characterized by a parameter vector $\boldsymbol{\theta}$, and so we can, at least in principle, graph the CVF as a function of $\boldsymbol{\theta}$. As an illustration, Figure 1 shows the CVF, for $\alpha = .05$, for a hypothetical nonpivotal test statistic, asymptotically distributed as the absolute value of $N(0, 1)$, for a model the DGPs of which are characterized by a single parameter θ. It follows from (9) that the rejection region of the bootstrap test can be defined in the space of $\hat{\theta}$ and τ, as the region above the graph of the CVF.

The rectangle above the horizontal line marked $Q(.05, 0)$ in the figure shows all $(\hat{\tau}, \hat{\theta})$ pairs that would lead to rejection at nominal level .05 when $\theta_0 = 0$ if the ideal P value (1) could be used. In contrast, the area above the CVF shows all

pairs that actually *will* lead to rejection with a bootstrap test. The effect of the difference between the two rejection regions on the performance of the bootstrap test depends on the joint distribution of τ and $\hat{\theta}$. For comparison, the rectangle above the dotted line shows all pairs that lead to rejection with the asymptotic critical value $Q^{\infty}(.05) = 1.96$. Clearly, the bootstrap test will work much better than the asymptotic test.

From Figure 1, we see that, when $\theta_0 = 0$, the bootstrap test may either overreject or underreject. For values of $\hat{\theta}$ reasonably near 0, it will overreject when $\hat{\theta} < 0$. For those values of $\hat{\theta}$, the CVF is below $Q(.05, 0)$, and the bootstrap critical value will consequently be too small. Similarly, the bootstrap test will underreject whenever $\hat{\theta} > 0$. If $\hat{\theta}$ is approximately unbiased and not very variable, these two types of errors should tend to offset each other, since the CVF is approximately linear near $\theta = 0$. In contrast, near the minimum θ_1 and the maximum θ_2, all the errors will be of the same sign. Thus we would expect the bootstrap test to underreject when $\theta_0 = \theta_1$ and to overreject when $\theta_0 = \theta_2$.

Figure 2 graphs the RPF $R(.05, \theta)$ for exactly the same one-parameter case as the CVF in Figure 1. Both functions evidently convey essentially the same information.

3. The Rejection Probabilities of Bootstrap Tests

In this section, we obtain an approximate expression for the ERP of bootstrap tests. The expression we obtain is useful for at least three reasons. First, it provides a clear intuitive explanation of the source of the ERP, based on the joint distribution of the test statistic τ and the bootstrap DGP $\hat{\mu}$. Second, it shows that an additional asymptotic refinement that is known to exist for certain specific cases, beyond the usual one in (6) noted by Beran, is available more generally. Third, it allows us to develop a relatively inexpensive simulation method for estimating the ERP of bootstrap tests.

It is convenient for our analysis to replace the statistic τ by the asymptotic P value $1 - F(\tau)$, which we henceforth denote simply by τ. The sign of the inequality in (7) must be changed, because one rejects when a P value is less, rather than greater, than a given value. For statistics τ that are asymptotic P values, the CVF is defined by

$$\Pr_{\mu}\big(\tau \leq Q(\alpha, \mu)\big) = \alpha \tag{10}$$

rather than by (7). Clearly, $Q(\alpha, \mu)$ is now the α quantile of τ under μ. Similarly, (3) simplifies to $R(\alpha, \mu) = \Pr_{\mu}(\tau \leq \alpha)$, and (8) reduces to

$$R\big(Q(\alpha, \mu), \mu\big) = \alpha, \tag{11}$$

from which it is clear that $Q(\alpha, \mu)$ is the inverse function of $R(\alpha, \mu)$ for given μ. Since both R and Q are increasing in their first arguments, (11) implies that $Q\big(R(\alpha, \mu), \mu\big) = \alpha$. Analogously to (4), we have

$$Q(\alpha, \mu) = \alpha + n^{-l/2} q(\alpha, \mu), \tag{12}$$

with the function q of order unity. The l in (12) is the same as the l in (4).

Equations like (4) and (12) do not look very much like the Edgeworth expansions about the standard normal density used in, among others, Hall and Titterington (1989) and Hall (1992). They are, however, fundamentally equivalent, the differences being due to our use of statistics that are asymptotically uniform on $[0, 1]$, rather than standard normal. The uniform density turns out to be better adapted to the study of ERP.

The bootstrap critical value for τ at nominal level α is $Q(\alpha, \hat{\mu})$, a random variable that is asymptotically nonrandom and equal to α. For given α, it is convenient to define a new random variable γ, of order unity as $n \to \infty$, as follows:

$$Q(\alpha, \hat{\mu}) = Q(\alpha, \mu_0) + n^{-k/2}\gamma, \tag{13}$$

where k is chosen to make (13) true. For the parametric bootstrap with root-n consistent estimates, Beran's argument in the last section implies that $k = l + 1$. This is also true for the nonparametric bootstrap whenever Hall's Edgeworth expansion theory based on the smooth function model applies. In both these cases, (6) shows that $R(\alpha, \hat{\mu}) - R(\alpha, \mu_0)$, and so also $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$, is $O(n^{-(l+1)/2})$. Thus, since $l \geq 1$, we can be confident that $k \geq 2$ in most cases of interest.

The RPF $R(\cdot, \mu_0)$ is the marginal CDF of τ. In order to calculate the rejection probability of the bootstrap test under the DGP μ_0, we need the joint distribution under μ_0 of τ and γ. Thus, let $g(\gamma \mid \tau)$ be the density of γ conditional on τ under μ_0. With this specification, we can compute the bootstrap rejection probability as the probability under μ_0 that $\tau \leq Q(\alpha, \hat{\mu})$. By (13), this is

$$\int_{-\infty}^{\infty} d\gamma \int_0^{Q+n^{-k/2}\gamma} dR(\tau)\, g(\gamma \mid \tau), \tag{14}$$

where, for ease of notation, we have set $Q = Q(\alpha, \mu_0)$ and $R(\tau) = R(\tau, \mu_0)$.

The integral over τ in (14) can be split into two parts, as follows:

$$\int_0^{Q} dR(\tau) \int_{-\infty}^{\infty} d\gamma\, g(\gamma \mid \tau) + \int_{-\infty}^{\infty} d\gamma \int_0^{n^{-k/2}\gamma} d\tau\, R'(Q+\tau)\, g(\gamma \mid Q+\tau), \tag{15}$$

where R' is the derivative of $R(\tau)$. Because g is a density, the integral over γ in the first term of (15) equals 1, and so the whole first term equals α, by (11). The ERP is therefore given by the second term in (15). By (4), $R'(Q+\tau) = 1 + O(n^{-l/2})$. Thus, if g is smooth enough, we may write the second term of (15) as

$$\begin{aligned} & n^{-k/2} \int_{-\infty}^{\infty} d\gamma\, \gamma \left(g(\gamma \mid \alpha) + O(n^{-k/2})\right)\left(1 + O(n^{-l/2})\right) \\ = {} & n^{-k/2} \int_{-\infty}^{\infty} d\gamma\, \gamma\, g(\gamma \,|\, \alpha) + O(n^{-(k+l)/2}). \end{aligned} \tag{16}$$

Expression (16) is the ERP of the bootstrap test. The first term is $O(n^{-k/2})$, in accord with Beran's (1988) analysis for asymptotic pivots. We now go beyond his analysis by exploiting the explicit expression in (16) for the leading-order distortion.

The leading-order term in (16) is the expectation, conditional on $\tau = \alpha$, of $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$. Thus it is the bias, conditional on $\tau = \alpha$, of the bootstrap critical value for nominal level α. When this bias is nonzero and of order $n^{-k/2}$, it is responsible for the leading-order ERP of the bootstrap test. But if it is zero or of lower order, the ERP is of lower order, because, in that case, those $\hat{\mu}$ which overestimate the critical value will on average be balanced to leading order by those $\hat{\mu}$ which underestimate it. Specifically, if

$$\int_{-\infty}^{\infty} d\gamma \, \gamma \, g(\gamma \mid \alpha) = O(n^{-i/2}),$$

for $i \leq k$, the distortion (16) is of order $n^{-(k+i)/2}$.

The actual value of k in specific testing situations can often be found in the existing literature on bootstrap confidence intervals, because the limits of these intervals are defined in terms of quantiles of the bootstrap distribution, and that is precisely what $Q(\alpha, \hat{\mu})$ is. For instance, in Section 3.6 of Hall (1992), it is shown that, for a symmetric confidence interval based on an asymptotically standard normal statistic, the exact and bootstrap critical values differ only at order $n^{-3/2}$. In our terms, $k = 3$. Hall goes on to show that the coverage error of a bootstrap confidence interval is of still lower order, n^{-2} in general. Examination of Hall's derivation of this result reveals that the additional refinement is due, as (16) would suggest, precisely to the fact that the *bias*, or expectation of the difference between the true and bootstrap critical values, is of lower order than the difference itself.

In fact, the interpretation of the leading-order term of (16) as a conditional expectation makes it obvious why it vanishes for a symmetric two-tailed test. The condition that $\tau = \alpha$ corresponds to two possibilities for the underlying asymptotically standard normal statistic: it may be equal to either the positive critical value or its negative. Under the null, because the test is symmetric, these events are equally probable to leading order. If γ and the asymptotically standard normal statistic have an approximate bivariate normal distribution, as they do in Hall's demonstration, then the expectations of γ conditional on the two critical values, positive and negative, are equal and opposite in sign. Thus the expectation of γ conditional on $\tau = \alpha$ vanishes to leading order.

4. A Further Refinement

In general, without any requirement of a symmetric two-sided test, the first term in (16) will vanish to leading order if a further condition is satisfied. This condition is that γ and τ should be asymptotically independent.

The asymptotic independence of γ and τ can often be achieved by using the fact that parameter estimates under the null are asymptotically independent of the statistics

associated with tests of that null in a wide variety of circumstances. If $\hat{\boldsymbol{\theta}}$ is an extremum estimator that satisfies first-order conditions in the interior of the parameter space of the null hypothesis, the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ will be asymptotically independent of any classical test statistic. For the case of the classical test statistics based on maximum likelihood estimation, a detailed proof of this may be found in Davidson and MacKinnon (1987). The proof can be extended in regular cases to NLS, GMM, and other forms of extremum estimation.

Thus, for the parametric bootstrap, the condition of asymptotic independence of τ and γ will always be satisfied, provided the parameters are estimated under the null and have the usual asymptotic properties. This is because $Q(\alpha, \hat{\mu})$, and hence γ, is a smooth function of the parameter estimates, and is thus asymptotically independent of the test statistic τ. This argument clearly also applies to cases in which only a conditional model for the dependent variable is fully parametric.

Asymptotic independence can often be achieved with little trouble for the nonparametric bootstrap as well. As one example, consider bootstrapping a test statistic in a linear regression model that includes a constant term by resampling from the residuals. Here, in order to achieve asymptotic independence, the bootstrap DGP would be based on estimating the model under the null. The bootstrap regression function would be given by the fitted values, which, as before, are asymptotically independent of any classical test statistic, and the bootstrap error terms would be independent drawings from the empirical distribution of the residuals, u_t. Now consider a t statistic on a variable not included under the null. Asymptotically, it will be a linear combination of the residuals, and so not independent of the bootstrap distribution. To leading order asymptotically, it will have the form

$$\tau \equiv n^{-1/2} \sum_{t=1}^{n} x_t u_t, \quad \text{where } \sum_{t=1}^{n} x_t = 0 \text{ and } \sum_{t=1}^{n} x_t^2 = 1. \tag{17}$$

The quantiles of the bootstrap distribution are determined by the empirical distribution of the residuals. Consider the asymptotic covariance of τ and the empirical distribution function evaluated at z. It is

$$\lim_{n\to\infty} \operatorname{Cov}\Biggl(n^{-1/2} \sum_{t=1}^{n} x_t u_t,\ n^{-1/2} \sum_{t=1}^{n} \Bigl(I(u_t < z) - E\bigl(I(u_t < z)\bigr)\Bigr)\Biggr), \tag{18}$$

where $I(\cdot)$ denotes the indicator function, which is equal to 1 when its argument is true and 0 otherwise. Since the residuals are asymptotically IID, (18) becomes

$$\lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} x_t E\Biggl(u_t \Bigl(I(u_t < z) - E\bigl(I(u_t < z)\bigr)\Bigr)\Biggr) = a \lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} x_t = 0,$$

where $a \equiv E\bigl(u_t\bigl(I(u_t < z) - E(I(u_t < z))\bigr)\bigr)$ is independent of t. Since both τ and the empirical CDF are asymptotically normal, they are asymptotically independent

because their asymptotic covariance is zero. This kind of result can be obtained much more generally, of course.

Suppose now that γ and τ are asymptotically independent. Then the conditional density of γ becomes

$$g(\gamma \mid \tau) = g(\gamma)\big(1 + n^{-j/2} f(\gamma, \tau)\big), \tag{19}$$

where $g(\gamma)$ is the asymptotic marginal density of γ, and $j \geq 1$ is chosen so that $f(\gamma, \tau)$ is of order unity as $n \to \infty$.

For any valid bootstrap procedure, $q(\alpha, \hat{\mu})$ must be a consistent estimator of $q(\alpha, \mu_0)$, so that

$$\int_{-\infty}^{\infty} d\gamma \; \gamma \, g(\gamma) = 0,$$

for otherwise $q(\alpha, \hat{\mu})$ would be asymptotically biased. Then (16) becomes

$$n^{-(k+j)/2} \int_{-\infty}^{\infty} d\gamma \; \gamma \, g(\gamma) \, f(\gamma, \alpha) + O(n^{-(k+j+1)/2}). \tag{20}$$

The interpretation of (20) is the same as that of (16). The first term is the bias, conditional on $\tau = \alpha$, of the bootstrap critical value at nominal level α. When $k = 2$ and $j = 1$, this term will be $O(n^{-3/2})$.

Under asymptotic independence of the statistic τ and the bootstrap DGP $\hat{\mu}$, the extra refinement seen in (20) can be derived within the context of Hall's Edgeworth expansion approach, although with considerably more difficulty than that involved in deriving (20); see the Appendix.

5. Estimating the ERP of Bootstrap Tests

In the study of the performance of bootstrap tests, Monte Carlo experiments play a vital role. However, such experiments can be extremely time-consuming, because each replication requires the calculation of $B+1$ test statistics if B bootstrap samples are used. In this section, we show that, when τ and $\hat{\mu}$ are asymptotically independent, it is possible to estimate the ERP of a bootstrap test by calculating only two statistics per replication.

With asymptotic independence, the first term in (16) is, to leading order, just the unconditional expectation of $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$. From (4) and (11), it follows that, to order $n^{-l/2}$,

$$Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0) = R(\alpha, \mu_0) - R(\alpha, \hat{\mu}). \tag{21}$$

The unconditional expectation of this is approximately equal to the ERP when α is the nominal level of the *asymptotic* test. However, since the nominal level of an asymptotic test is often far from its actual rejection probability, it is of greater interest to study the ERP when α is the actual rejection probability. To do so, we need to replace α in (21) by $Q(\alpha, \mu_0)$, because, by (10), $Q(\alpha, \mu_0)$ is the critical value

that yields a rejection probability of α. Thus, by (11), the ERP of the bootstrap test is approximately equal to

$$E_{\mu_0}\Big(R\big(Q(\alpha,\mu_0),\mu_0\big) - R\big(Q(\alpha,\mu_0),\hat{\mu}\big)\Big) = \alpha - E_{\mu_0}\Big(R\big(Q(\alpha,\mu_0),\hat{\mu}\big)\Big). \tag{22}$$

Now consider a random variable which can be simulated as follows. First, a random bootstrap DGP $\hat{\mu}$ is drawn from μ_0. Then the random variable is obtained by drawing a statistic τ from $\hat{\mu}$. The CDF of this variable evaluated at α is

$$E_{\mu_0}\big(\Pr_{\hat{\mu}}(\tau < \alpha)\big) = E_{\mu_0}\big(R(\alpha,\hat{\mu})\big). \tag{23}$$

This suggests that we can estimate (22) as follows. For each of M replications indexed by m, draw a sample from μ_0. Use this sample to compute $\hat{\tau}_m$ and the bootstrap DGP $\hat{\mu}_m$. Now draw another sample from $\hat{\mu}_m$, and use this second sample to compute $\hat{\tau}^*_m$. In this way, $\hat{\tau}^*_m$ is by construction a drawing from the distribution (23), and $\hat{\tau}_m$ is a drawing from the distribution with CDF $R(\cdot,\mu_0)$. Thus we can estimate $Q(\alpha,\mu_0)$ as the α quantile of the $\hat{\tau}_m$, and then estimate $E_{\mu_0}\big(R(Q(\alpha,\mu_0),\hat{\mu})\big)$ as the proportion of the $\hat{\tau}^*_m$ that are smaller than the estimate of $Q(\alpha,\mu_0)$.

Let the estimate of $E_{\mu_0}\big(R(Q(\alpha,\mu_0),\hat{\mu})\big)$ be denoted b. Then, by (22), the estimate of the bootstrap ERP is $\alpha - b$, and the estimate of the rejection probability of the bootstrap test is $2\alpha - b$. For the last estimate to lie between 0 and 1, it is necessary that $2\alpha - 1 \leq b \leq 2\alpha$. For $\alpha < 1/2$, the first inequality is automatically satisfied. If the second is not, it means that more than twice as many of the $\hat{\tau}^*_m$ as of the $\hat{\tau}_m$ are smaller than the estimate of $Q(\alpha,\mu_0)$. In such a case, the bootstrap ERP is so great that the present method is likely to be inapplicable.

We have investigated the performance of this Monte Carlo procedure in the context of three different bootstrap tests, both parametric and nonparametric. In all three cases, it worked remarkably well. As an illustration, we report the results of one set of experiments here. These experiments concern a Wald test for linear restrictions on the parameters of a tobit model. There are 10 regressors, including a constant term, and 7 restrictions. See Davidson and MacKinnon (1999) for discussion of the model and the test statistic. The number of restrictions is relatively large, because the performance of the bootstrap and asymptotic tests deteriorates as this number increases, and we wanted to investigate a case in which the bootstrap test does not work perfectly.

The results of our experiments are summarized in Figure 3. We generated 1,000,000 samples for each of $n = 40, 42, 44, \ldots, 100$ and calculated the rejection probabilities for the asymptotic test at nominal level .05. For each sample, we also calculated $\hat{\tau}^*_m$ and used it to estimate the rejection probability for the bootstrap test, as described above. The number of replications is very large because we did not want the figure to suffer from too much experimental error. Sample sizes less than 40 were not investigated to avoid too many replications for which ML estimates of the tobit model do not exist. The rejection probabilities, actual in the case of the asymptotic

test and approximate in the case of the bootstrap test, are shown respectively as solid and dotted curves. As the theory suggests, the bootstrap test works very much better than the asymptotic test for this example.

It would have been very costly indeed to investigate the actual performance of bootstrap tests for as many sample sizes, or as many replications, as we used to estimate its performance based on (22). We contented ourselves with experiments of $500,000$ replications for tests based on 99 bootstrap samples, for $n = 40, 50, 60, \ldots, 100$. The results are displayed as large dots in Figure 3. These seven large dots required far more computing time than the two curves. It can be seen that the estimates of the ERP of the bootstrap test based on (22) are almost identical to those from the brute-force estimates based on actual bootstrapping; any differences can be attributed to experimental error.

These results appear to be extremely promising. It would probably be unwise to conduct a Monte Carlo experiment on the performance of a bootstrap test without ever estimating the latter directly. However, it will often be possible to use the approximate method proposed here to investigate most cases, with only a few brute force experiments to verify that the approximate results are valid. It also seems plausible that (22) can be made use of in practice to reduce the ERP of bootstrap tests whenever τ and $\hat{\mu}$ are asymptotically independent. This conjecture is the subject of ongoing research.

6. Summary and Conclusion

In this paper, we have provided a simple theoretical framework for understanding a number of existing results on the higher-order refinements provided by the bootstrap. It is well known that, when an asymptotically pivotal statistic is used for testing, the bootstrap test yields a refinement of at least order $n^{-1/2}$ over the asymptotic test. That further refinements, beyond $n^{-1/2}$, are sometimes available, as in the case of tests that are asymptotically chi-squared, is also known, but the source of these refinements is not immediately obvious. In Section 3, we explained precisely where they come from.

Our analysis shows that the extra refinement, which in most cases will be of order $n^{-1/2}$, is not limited to specific cases, but is quite generally available if the test statistic is asymptotically independent of the bootstrap DGP. This condition is satisfied for the parametric bootstrap if the parameter estimates are extremum estimators obtained under the null, and in many cases for the nonparametric bootstrap as well. Together with existing results, this new result suggests that bootstrap tests can, in many circumstances, be more accurate than asymptotic tests by a full order of n^{-1}.

The asymptotic independence condition, when it is satisfied, makes it possible to estimate the ERP of bootstrap tests by computing only two test statistics on each replication, instead of the $B + 1$ test statistics that would be needed to calculate a bootstrap P value. This result is potentially of great utility for Monte Carlo studies

of the bootstrap. Although we do not claim that it will work well in all cases, we have demonstrated that it can work very well indeed in practice.

References

Attfield, C.L.F. (1995) A Bartlett adjustment to the likelihood ratio test for a system of equations. *Journal of Econometrics* 66, 207–223.

Beran, R. (1986) Simulated power functions. *Annals of Statistics* 14, 151–173.

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

Davidson, R. & J.G. MacKinnon (1987) Implicit alternatives and the local power of test statistics. *Econometrica* 55, 1305–1329.

Davidson, R. & J.G. MacKinnon (1999) Bootstrap testing in nonlinear models. *International Economic Review* 40, forthcoming.

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

Hall, P. & J. L. Horowitz (1996) Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* 64, 891–916.

Hall, P. & D.M. Titterington (1989) The effect of simulation order on level accuracy and power of Monte-Carlo tests. *Journal of the Royal Statistical Society* Series B 51, 459–467.

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

Rothenberg, T.J. (1984) Hypothesis testing in linear models when the error covariance matrix is nonscalar. *Econometrica* 52, 827–842.

Appendix

In this appendix, we demonstrate that the extra refinement seen in (20) can also be obtained within Hall's Edgeworth expansion approach. We consider a simple model and two different parametric bootstrap procedures, one for which the test statistic and the bootstrap DGP are asymptotically independent, and another for which they are not. In accord with the results from our approach, we find that the order of the bootstrap ERP is smaller in the former case.

The model has just two parameters, θ_1 and θ_2, which can be estimated by maximum likelihood. The test statistic is a Wald test of the hypothesis that $\theta_2 = 0$, computed as an asymptotic t statistic. In order not to be confused by the extra refinement given by a two-tailed test, we consider a one-tailed test against the alternative that $\theta_2 < 0$. We consider two possible parametric bootstraps. For the first, the bootstrap DGP is the DGP with parameters $(\tilde{\theta}_1, 0)$, where $\tilde{\theta}_1$ is the ML estimate of θ_1 subject to the constraint $\theta_2 = 0$. By standard ML theory, $\tilde{\theta}_1$ is asymptotically independent of any classical test statistic for the hypothesis that $\theta_2 = 0$. The second bootstrap DGP uses the parameters $(\hat{\theta}_1, 0)$, where $\hat{\theta}_1$ is the unconstrained ML estimate from a procedure in which θ_1 and θ_2 are estimated jointly. The Wald statistic, based on $\hat{\theta}_2$, will not, in general, be asymptotically independent of this second bootstrap DGP.

Without loss of generality, we suppose that the true value of θ_1 is zero. The ML estimator is asymptotically normal, and so we can represent the asymptotic distribution of $n^{1/2}(\hat{\theta}_1, \hat{\theta}_2)$ as bivariate normal, with mean zero, and covariance matrix

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

of order unity as $n \to \infty$. Then, by standard results on constrained ML estimation, we have that, asymptotically,

$$n^{1/2}\tilde{\theta}_1 = n^{1/2}\hat{\theta}_1 - \frac{\rho\sigma_1}{\sigma_2} n^{1/2}\hat{\theta}_2. \tag{A.1}$$

In what follows, we ignore the fact that σ_1, σ_2, and ρ will usually have to be estimated, since this makes no difference asymptotically. Similarly, the Wald statistic is asymptotically just $n^{1/2}\hat{\theta}_2/\sigma_2$.

For tests based on asymptotically standard normal statistics, equation (3.24) of Hall (1992) implies that

$$\hat{v}_\alpha - v_\alpha = n^{-1/2}\big(\hat{q}_{11}(z_\alpha) - q_{11}(z_\alpha)\big) + O(n^{-3/2}), \tag{A.2}$$

where v_α and $\hat{v}_\alpha$ are the α quantiles of the distribution of the test statistic when it is generated by the true DGP and by the bootstrap DGP respectively. The functions q_{11} and $\hat{q}_{11}$ are even polynomials of degree at most 2, and their argument z_α is the α quantile of the standard normal distribution. The coefficients of the polynomials q_{11} and $\hat{q}_{11}$ are smooth functions of the low-order moments of the test statistic under the

true and bootstrap DGPs, respectively. For the parametric bootstrap, these low-order moments are functions of the parameters of the DGP, and so, since the ML estimates of the parameters are root-n consistent, we have that $\hat{q}_{11}(z) - q_{11}(z) = O(n^{-1/2})$, from which it follows that $\hat{v}_\alpha - v_\alpha = O(n^{-1})$.

We used Hall's notation in (A.2). It is more convenient to write the polynomials as explicit functions of the parameter θ_1 that characterizes the DGP. Thus we write $q(\theta_1)$ in place of $q_{11}(z_\alpha)$ and $q(\ddot{\theta}_1)$ in place of $\hat{q}_{11}(z_\alpha)$, where $\ddot{\theta}_1$ denotes either $\tilde{\theta}_1$ or $\hat{\theta}_1$, depending on which bootstrap is used. For convenience, we drop the z_α argument. It follows that

$$n(\hat{v}_\alpha - v_\alpha) = n^{1/2}\big(q(\ddot{\theta}_1) - q(0)\big) + O(n^{-1/2}) = q'(0)\, n^{1/2}\ddot{\theta}_1 + O(n^{-1/2}), \qquad \text{(A.3)}$$

where q' is the derivative of q with respect to θ. Note that, in our notation,

$$Q(\alpha, \mu_0) = \Phi(v_\alpha) \quad \text{and} \quad Q(\alpha, \hat{\mu}) = \Phi(\hat{v}_\alpha),$$

where Φ is the standard normal CDF.

We may now make use of the tool for calculating the ERP in the conventional approach, which is Proposition 3.1 in Hall (1992). This result is expressed in terms of the coverage error of a confidence interval. After some simplification made possible by the simplicity of the model under consideration, we find that, for a one-sided interval with nominal coverage α of the sort Hall calls $\hat{J}_1$, which is what we need here, the coverage error is $-n^{-1}\phi(z_\alpha)a_\alpha z_\alpha$, where ϕ is the standard normal density, and a_α is defined to leading order as minus the covariance of the test statistic and expression (A.3). For the bootstrap that uses $\hat{\theta}_1$, we obtain

$$a_\alpha = -E\Big(n^{1/2}\,\frac{\hat{\theta}_2}{\sigma_2}\,q'(0)\,n^{1/2}\hat{\theta}_1\Big) = -\frac{1}{\sigma_2}\,q'(0)\,\rho\sigma_1\sigma_2 = -\rho\sigma_1 q'(0),$$

while, for the bootstrap that uses $\tilde{\theta}_1$, we find from (A.1) that

$$a_\alpha = -E\Big(n^{1/2}\,\frac{\hat{\theta}_2}{\sigma_2}\,q'(0)\,n^{1/2}\tilde{\theta}_1\Big) = -\frac{1}{\sigma_2}\,q'(0)\Big(\rho\sigma_1\sigma_2 - \frac{\rho\sigma_1}{\sigma_2}\sigma_2^2\Big) = 0.$$

This last result is the additional refinement due to asymptotic independence. If one wishes to establish a closer correspondence with the approach of this paper, it should be noted that the nominal coverage of α corresponds to a test of nominal level $1-\alpha$, and that a negative coverage error corresponds to a positive ERP.

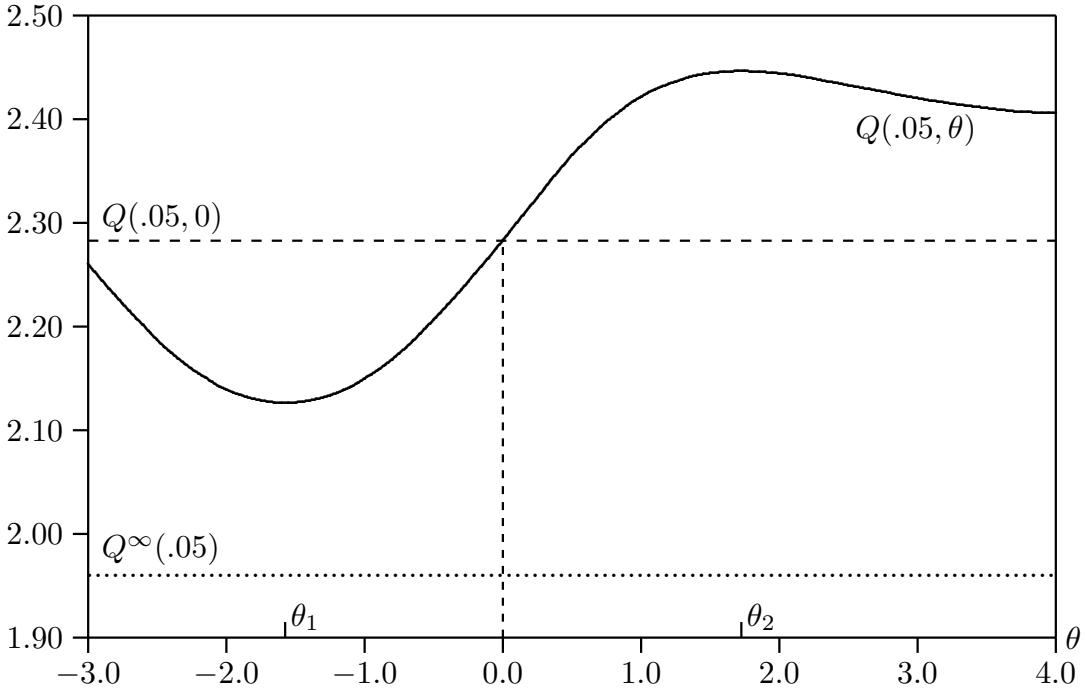

Figure 1. A Critical Value Function

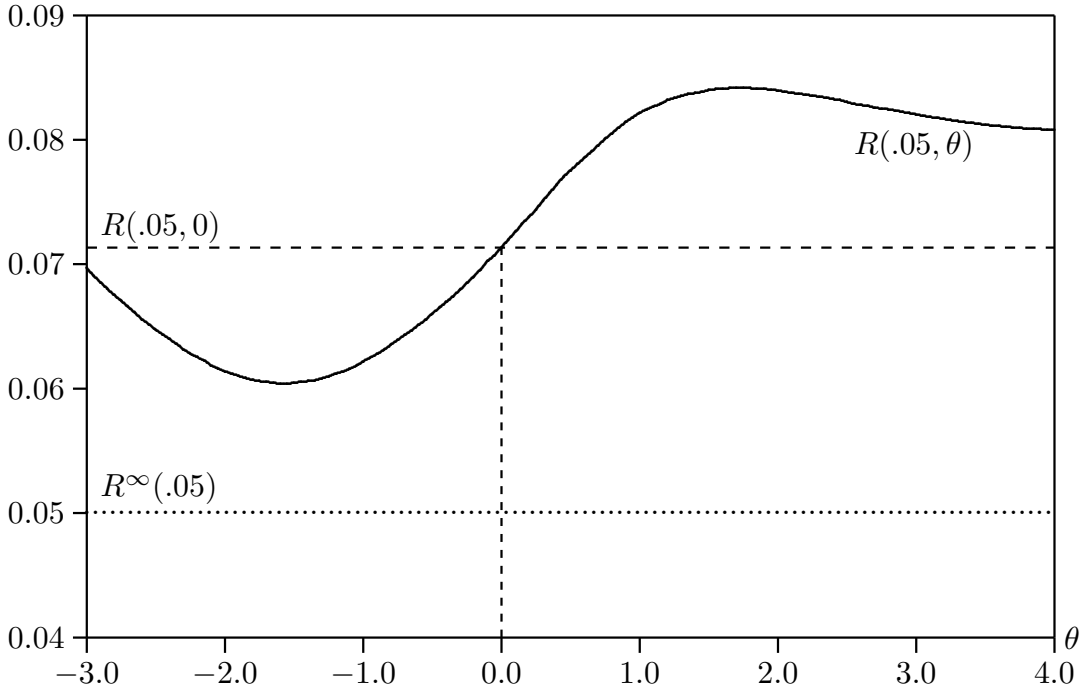

Figure 2. A Rejection Probability Function

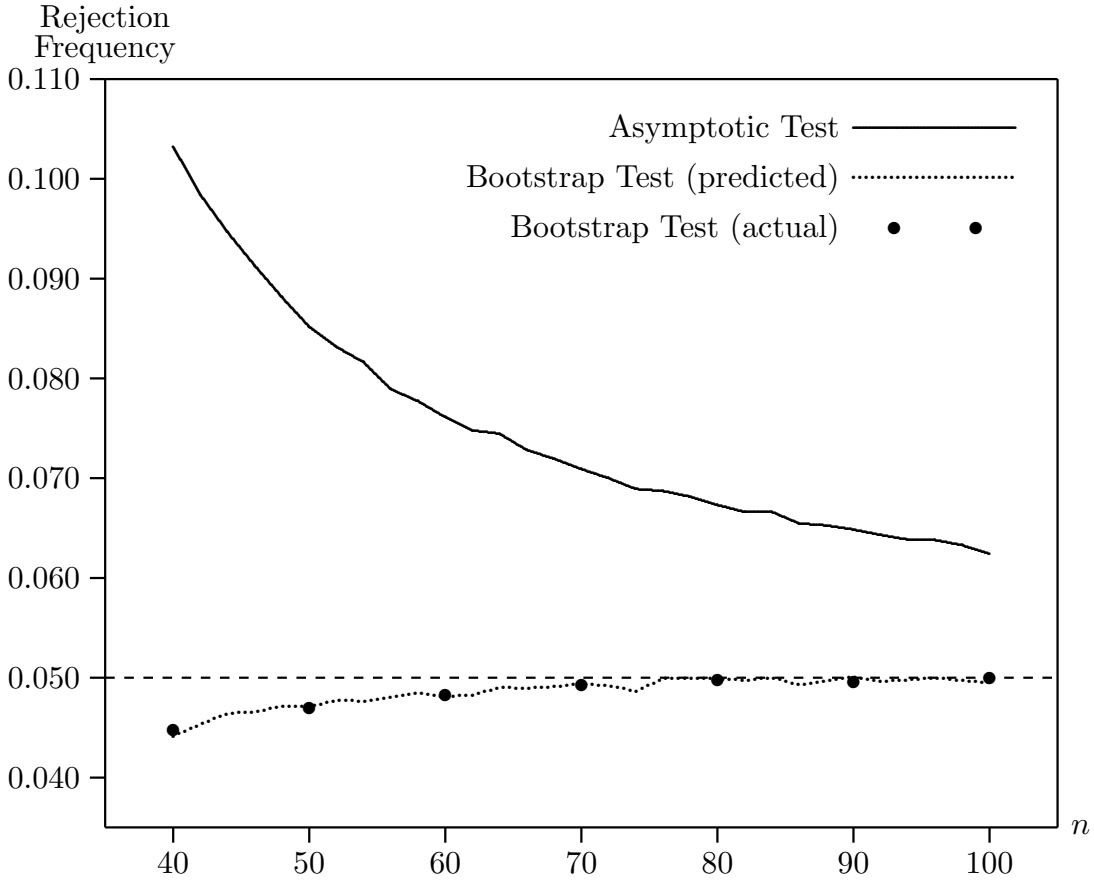

Figure 3. Actual and Predicted Rejection Frequencies at .05 Level