Queen’s Economics Department Working Paper No. 937

The Power of Bootstrap Tests

Russell Davidson
GREQAM, Queen’s University

James G. MacKinnon
Queen’s University

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

09-1996

The Power 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

Electronic mail: **russell@ehess.cnrs-mrs.fr**

and

James G. MacKinnon

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

Electronic mail: **jgm@qed.econ.queensu.ca**

Abstract

Bootstrap tests are tests for which the significance level is calculated using some variant of the bootstrap, which may be parametric or nonparametric. We show that the power of a bootstrap test will generally be very close to the power of the asymptotic test on which it is based, provided that both tests are properly adjusted to have the correct size. We also discuss the loss of power that can occur when the number of bootstrap samples is relatively small. Some Monte Carlo results for two forms of omitted variable test in logit models are presented. These illustrate the theoretical results of the paper and demonstrate that the size-adjusted power of asymptotic tests can vary greatly depending on the method used for size adjustment.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Joel Horowitz and numerous seminar participants for comments on earlier work. Much of the paper was written while the second author was visiting GREQAM.

September, 1996

1. Introduction

In many situations, the bootstrap can be used to perform hypothesis tests that are more reliable in finite samples than tests based on asymptotic theory. In econometrics, the use of the bootstrap for this purpose has been advocated by Horowitz (1994), Hall and Horowitz (1996), Davidson and MacKinnon (1996), and others. If the bootstrap is to work well, the original test statistic must be asymptotically pivotal. In other words, its asymptotic distribution must not depend on any unknown features of the process that generated the data. For asymptotically pivotal test statistics, the bootstrap will yield more accurate inferences than asymptotic theory, in the sense that the errors it makes will be of lower order in the sample size n. The errors committed by using the bootstrap are generally lower by a factor of either $n^{-1/2}$ or n^{-1} than the errors committed by relying on asymptotic theory; see Hall (1992) and Davidson and MacKinnon (1996).

There are several ways in which the bootstrap can be used for hypothesis testing. One approach, which in our view is the simplest and most satisfactory, is to use the bootstrap to compute P values. We first compute a test statistic, say $\hat{\tau}$, in the usual way. Using estimates of the model under the null hypothesis, we then draw B bootstrap samples. Each of these is used to compute a bootstrap test statistic τ_j^* in exactly the same way that $\hat{\tau}$ was computed from the real sample. For a one-tailed test with a rejection region in the upper tail, the bootstrap P value may then be estimated by

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* \geq \hat{\tau}), \tag{1}$$

where $I(\cdot)$ is the indicator function. As $B \to \infty$, it is clear that the estimated bootstrap P value $\hat{p}^*(\hat{\tau})$ will tend to the (true) bootstrap P value $p^*(\hat{\tau})$, which is defined as

$$p^*(\hat{\tau}) \equiv \Pr_{\hat{\mu}}(\tau \geq \hat{\tau}),$$

where $\hat{\mu}$ denotes the *bootstrap DGP* that is used to generate the bootstrap samples. For the purposes of this paper, we will define an (ideal) *bootstrap test* as a test that is based on the bootstrap P value $p^*(\hat{\tau})$ and a *feasible bootstrap test* as one that is based on the estimated bootstrap P value $\hat{p}^*(\hat{\tau})$. Thus a bootstrap test will reject the null hypothesis at level α if $p^*(\hat{\tau}) < \alpha$, and a feasible bootstrap test will reject it if $\hat{p}^*(\hat{\tau}) < \alpha$.

Of course, $\hat{\mu}$ will be a function of the data, and so it will generally differ from the true (but unknown) DGP μ_0. As a consequence, unless the test statistic τ is pivotal, the bootstrap P value will generally differ from the true P value, which is defined as

$$p(\hat{\tau}) \equiv \Pr_{\mu_0}(\tau \geq \hat{\tau}),$$

causing the size of the bootstrap test to be incorrect. However, since the difference between $p^*(\hat{\tau})$ and $p(\hat{\tau})$ is at most $O(n^{-1})$, and is in many cases of lower order, any size distortion of the bootstrap test is often (but not always) very small. For a detailed analysis of what determines the size distortion of bootstrap tests, see Davidson and MacKinnon (1996), hereafter referred to as DM96.

It is natural to ask whether bootstrapping a test has any effect on its power. There are two reasons for which it might have an effect. The first is that $p^*(\hat{\tau})$ might be systematically larger or smaller than $p(\hat{\tau})$ when the DGP does not belong to the null hypothesis. As we shall demonstrate in Section 3, this is not the case. On the contrary, as we shall prove, the difference between the power of the bootstrap test and the power of the asymptotic test on which it is based, when either both tests or just the asymptotic test are correctly adjusted to have the right size under the null, may be of either sign. Moreover, this difference is of the same order in n as the size distortion of the bootstrap test, and so it may reasonably be expected to be very small in many cases.

The second reason for which bootstrapping a test might affect its power is that $\hat{p}^*(\hat{\tau})$ might be less than $p^*(\hat{\tau})$ when the null hypothesis is false. This is in fact the case, as we will discuss in Section 4. However, provided B is chosen appropriately, the power loss from bootstrapping should normally be very small.

In the next section, we introduce some basic concepts and notation. In Section 3, we use these to prove the principal theoretical results of the paper. In Section 4, we discuss the power loss from employing feasible rather than ideal bootstrap tests. In Section 5, we discuss some important issues that arise whenever one attempts to compare the power of two or more tests in a Monte Carlo experiment when at least one of the tests does not have the correct size under the null hypothesis. These issues arise whether or not bootstrapping is involved, but they are often neglected. Finally, in Section 6, we present some Monte Carlo results for tests of omitted variables in a logit model. These results demonstrate just how useful the theoretical results of Sections 3 and 5 can be.

2. Some Basic Concepts

In order to discuss the finite-sample performance of bootstrap and asymptotic tests, we need to establish some basic concepts and notation. Many of these were first proposed in DM96, and more detailed discussion may be found there. It is very convenient to redefine the test statistic τ so that its nominal asymptotic distribution is uniform on $[0, 1]$. Thus, if the asymptotic distribution function for τ were $F(\tau)$, the original test statistic $\hat{\tau}$ would be replaced by its asymptotic P value, which would be $1 - F(\hat{\tau})$ for a one-tailed test and $1 - F(|\hat{\tau}|) + F(-|\hat{\tau}|)$ for a two-tailed test. If τ is thus redefined, an asymptotic test will reject the null hypothesis at level α whenever $\hat{\tau} < \alpha$. In the remainder of the paper, we shall for simplicity consider only tests that are in this form.

We can characterize the performance of a test based on $\hat{\tau}$ by two functions that depend on the level of the test α and on the DGP μ. The first of these is the P *value function*, or PVF. It is defined as

$$S(\alpha, \mu) \equiv \Pr_\mu(\hat{\tau} \leq \alpha). \tag{2}$$

The PVF $S(\alpha, \mu)$ measures the true size of the test that has nominal size α. For fixed μ, $S(\alpha, \mu)$ is just the c.d.f. of $\hat{\tau}$ evaluated at α. The inverse of the P value function is the *critical value function*, or CVF, which is implicitly defined by the equation

$$\Pr_\mu\big(\hat{\tau} \leq Q(\alpha, \mu)\big) = \alpha. \tag{3}$$

It is clear from (3) that $Q(\alpha, \mu)$ is the α quantile of the distribution of $\hat{\tau}$ under μ. If an asymptotic test were exact, then we would have $S(\alpha, \mu) = \alpha$ and $Q(\alpha, \mu) = \alpha$ for all α and μ. More commonly, asymptotic tests will either overreject or underreject in finite samples. If, for the DGP μ, a test overrejects at level α, then $S(\alpha, \mu) > \alpha$ and $Q(\alpha, \mu) < \alpha$.

The difference between $S(\alpha, \mu)$ and α will be referred to as the P *value discrepancy function* for the asymptotic test. It is implicitly defined by the equation

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

where the integer $l \geq 1$ is defined so that $s(\alpha, \mu)$ will be $O(1)$. In most cases, we expect that $l = 1$ or $l = 2$. Analogously to (4), we can define the *critical value discrepancy function* by the equation

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

Once again, the integer l is chosen so that $q(\alpha, \mu)$ is $O(1)$, and it will be the same as the l in (4).

The bootstrap critical value for $\hat{\tau}$ is $Q(\alpha, \hat{\mu})$. This is a random variable which will be asymptotically nonrandom and equal to α. Any size distortion of the bootstrap test arises from the possibility that $Q(\alpha, \hat{\mu}) \neq Q(\alpha, \mu_0)$. Although these two CVFs will generally differ whenever the test is not pivotal, there are good reasons to believe that they will often not differ by much. If $\hat{\tau}$ is nearly pivotal, then $Q(\alpha, \mu)$ does not depend much on μ. Moreover, if the bootstrap DGP is estimated with reasonable precision, $\hat{\mu}$ will generally be close to μ_0. It is therefore convenient to define a new random variable γ, of order unity as $n \to \infty$, as follows:

$$Q(\alpha, \hat{\mu}) = Q(\alpha, \mu_0) + n^{-k/2} \gamma, \tag{5}$$

where k is an integer chosen to make (5) true. It is difficult to say precisely what the value of k will be. However, as we discuss in DM96, $k = l + 1$ for the parametric bootstrap under standard regularity conditions. This will also be true for the nonparametric bootstrap in all cases to which the Edgeworth expansion theory of Hall (1992) applies. Since it is normally the case that $l = 1$ or $l = 2$, the most common values for k will be 2 and 3.

The P value function $S(\alpha, \mu)$ can be interpreted as the c.d.f. of $\hat{\tau}$ under μ. In order to describe the joint distribution of $\hat{\tau}$ and γ, we also need the distribution of γ conditional on $\hat{\tau}$. Let us denote by $g(\gamma \,|\, \tau)$ the density of γ conditional on $\hat{\tau} = \tau$ under the DGP μ_0. The true size of the bootstrap test is the probability under μ_0 that $\hat{\tau} \leq Q(\alpha, \hat{\mu})$. By (5), this true size is

$$\int_{-\infty}^{\infty} d\gamma \int_0^{Q + n^{-k/2}\gamma} dS(\tau)\, g(\gamma \,|\, \tau), \tag{6}$$

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

In DM96, we showed that the size of the bootstrap test, expression (6), is equal to α plus a discrepancy that vanishes asymptotically. This discrepancy can be written as

$$n^{-k/2} \int_{-\infty}^{\infty} d\gamma\, \gamma\, g(\gamma \,|\, \alpha) + O(n^{-(k+l)/2}). \tag{7}$$

The leading-order term in (7) has a simple interpretation. It is the expectation, conditional on $\hat{\tau} = \alpha$, of $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$. Thus, to leading order, it is the bias, conditional on $\hat{\tau} = \alpha$, of the bootstrap estimate of the size-α critical value. When this bias is nonzero, it is responsible for the size distortion of the bootstrap test to leading order. Assuming that $k = l + 1$, as will generally be the case, we see that the size distortion of the bootstrap test will be at most $O(n^{-1})$ when $l = 1$ and at most $O(n^{-3/2})$ when $l = 2$. When the bootstrap estimate of the size-α critical value is unbiased through $O(n^{-k/2})$, the leading-order term in (7) vanishes, and the size distortion of the bootstrap test will be of even lower order.

3. The Power of Bootstrap and Asymptotic Tests

In this section, we characterize the difference between the power of bootstrap tests and the power of the asymptotic tests on which they are based. The key difference between the analysis of this section and the analysis of DM96, which was discussed in the previous section, is that the DGP is now assumed to be μ_1, which is not a member of the null hypothesis. Therefore, the test statistic τ will no longer have a distribution close to uniform $[0, 1]$, at least not if the test has any reasonable power. In fact, if μ_1 were a fixed DGP, independent of the sample size, then τ would be asymptotically concentrated on zero, since any consistent test would asymptotically reject the null hypothesis with probability one. So that asymptotic theory will give sensible results, we therefore assume that μ_1 is a *drifting DGP*, that is, one which is determined by a DGP belonging to the null hypothesis plus a perturbation that is usually $O(n^{-1/2})$; see Davidson and MacKinnon (1993, Chapter 12) for a detailed discussion of drifting DGPs. Precisely how μ_1 is constructed will not matter for the results of this section.

For any given sample size, the c.d.f. of τ under μ_1 can be written as

$$P(\alpha, \mu_1) \equiv \Pr_{\mu_1}(\tau \leq \alpha). \tag{8}$$

This definition of $P(\alpha, \mu_1)$ is very similar to the definition of $S(\alpha, \mu)$ in (2). The notation has changed only because we are now concerned with power rather than with size. For α different from 0 or 1, $P(\alpha, \mu_1)$ will tend neither to 0 nor to 1 as $n \to \infty$, because μ_1 drifts toward the null hypothesis at an appropriate rate.

Unlike μ_1, the bootstrap DGP $\hat{\mu}$ must belong to the null hypothesis, or at least, in the nonparametric case, it must be close to it in the appropriate sense. It is a random distribution determined by μ_1 rather than by some μ_0 in the null hypothesis, but it is just the same sort of distribution as it would have been if it had been determined by such a μ_0. For a parametric bootstrap, it is determined by estimates of the parameters of the null hypothesis. For a nonparametric bootstrap, it is determined by residuals or similar quantities obtained by estimating the null hypothesis. The α quantile of $\hat{\mu}$ will still be asymptotically nonrandom and equal to α. It will be expressible, as in (5), in terms of $Q(\alpha, \mu_0)$, for some μ_0, and an asymptotically mean zero random variable that we can still write as γ.

It is not entirely clear just what μ_0 to use in equation (5) when the actual DGP is μ_1. Essentially, we want μ_0 to be as close as possible to μ_1 while still satisfying the null hypothesis. From the theoretical point of view, μ_0 must simply be such that $Q(\alpha, \hat{\mu})$ is given by (5). However, that equation is not quite enough to determine μ_0 uniquely, since changing μ_0 by an amount that affects $Q(\alpha, \mu_0)$ only by a quantity of order $O(n^{-(k+1)/2})$ is clearly compatible with all the requirements on k and on γ. For Monte Carlo experiments, the precise choice of μ_0 does matter, and this issue will be discussed further in Section 5.

Let us suppose then that, under the drifting DGP μ_1, equation (5) is satisfied for some μ_0 and for some γ that asymptotically has mean zero. As in DM96, we need the joint density of τ and γ under μ_1. By (8), the marginal density of τ is $P'(\tau)$, and we may denote the density of γ conditional on τ by $g_1(\gamma \,|\, \tau)$. The power of the bootstrap test based on τ at nominal size α is the probability under μ_1 of rejecting the null hypothesis, that is, the probability that $\tau \leq Q(\alpha, \hat{\mu})$. As in (6), the power can be expressed as

$$\int_{-\infty}^{\infty} d\gamma \int_{0}^{Q \,+\, n^{-k/2}\gamma} dP(\tau)\; g_1(\gamma \,|\, \tau). \tag{9}$$

This can be split into two parts. The first part is

$$\int_{0}^{Q} dP(\tau) \int_{-\infty}^{\infty} d\gamma\; g_1(\gamma \,|\, \tau) = P\big(Q(\alpha, \mu_0), \mu_1\big), \tag{10}$$

where the equality follows from the fact that $g_1(\gamma\,|\,\tau)$ is a density and therefore integrates to unity. The right-hand side of (10) is simply the power of the asymptotic test based on τ, where the critical value is such that the test has true size α under μ_0. It can therefore be interpreted as the size-corrected power of the asymptotic test at level α.

The second part into which (9) can be split is

$$\int_{-\infty}^{\infty} d\gamma \int_{0}^{n^{-k/2}\gamma} d\tau\; P'\big(Q(\alpha,\mu_0)+\tau\big)\, g_1\big(\gamma\,|\,Q(\alpha,\mu_0)+\tau\big). \tag{11}$$

To leading order, this is just

$$n^{-k/2}\, P'(\alpha) \int_{-\infty}^{\infty} d\gamma\; \gamma\, g_1(\gamma\,|\,\alpha) + O(n^{-(k+l)/2}); \tag{12}$$

compare (7). The first term here is, through $O(n^{-k/2})$, the difference in power between the bootstrap test at *nominal* level α and the asymptotic test at *true* level α.

From (7), we have seen that the size distortion of the bootstrap test is at most $O(n^{-k/2})$, the same order as the power difference given by (12). Therefore, we can be certain that the power of the bootstrap test at true level α differs from the power of the asymptotic test at true level α by at most $O(n^{-k/2})$. However, (12) does not provide an explicit expression for this difference. We can obtain such an expression if we replace α in (10) and (11) by the nominal level which corresponds to true level α for the bootstrap test. The appropriate correction is given by (7), and so α is to be replaced by

$$\alpha - n^{-k/2} \int_{-\infty}^{\infty} d\gamma\; \gamma\, g(\gamma\,|\,\alpha) + O(n^{-(k+l)/2}).$$

To the order we are considering, this replacement affects only (10), which becomes to leading order

$$P\big(Q(\alpha,\mu_0),\mu_1\big) - n^{-k/2}\, P'(\alpha) \int_{-\infty}^{\infty} d\gamma\; \gamma\, g(\gamma\,|\,\alpha).$$

From this and from (12), we conclude that the difference between the power of the bootstrap and asymptotic tests at true level α, conditional on μ_1, is

$$n^{-k/2}\, P'(\alpha) \int_{-\infty}^{\infty} d\gamma\; \gamma\, \big(g_1(\gamma\,|\,\alpha) - g(\gamma\,|\,\alpha)\big) + O(n^{-(k+l)/2}). \tag{13}$$

From expression (13), we obtain two important and very simple results. Firstly, we see that, in general, the bootstrap and asymptotic tests will have power that differs, on a size-corrected basis, only at $O(n^{-k/2})$. In the case of the parametric bootstrap, this means that any discrepancy is at most $O(n^{-(l+1)/2})$, a result that is essentially the same as one obtained by Horowitz (1994). Our result applies to the nonparametric bootstrap as well. Secondly, we see from the first term in (13) that, to highest order, any power difference is due solely to the possibility that γ may have a different distribution under the null and under the nonnull DGPs. In other words, the power difference arises from the possibility that $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$ may differ under μ_0 and μ_1. For a pivotal test, there can be no power difference, and for a test that is reasonably close to being pivotal, the difference should usually be quite small.

In some cases, it is possible to obtain an even stronger result than (13). When the bootstrap DGP is based on restricted parameter estimates under the null hypothesis, γ and $\hat{\tau}$ will be asymptotically independent. In this case, as we showed in DM96, expression (7) for the size distortion of the bootstrap test can be replaced by a similar expression in which the leading order term is $O(n^{-(k+j)/2})$ instead of $O(n^{-k/2})$, where $j \geq 1$. If we were to repeat the analysis that led to (13) for this case, the result would be similar to (13), but with the leading-order term being $O(n^{-(k+j)/2})$. We will not bother to go through this exercise, because, although the result is of some theoretical interest, it does not change the basic insight given by (13). The power difference between the bootstrap and asymptotic tests might be of even smaller order, but it would still arise from the possibility that $Q(\alpha, \hat{\mu}) - Q(\alpha, \mu_0)$ may differ under μ_0 and μ_1.

4. Sampling Variability and Power Loss

It is obvious that using the estimated P value $\hat{p}^*$ instead of the genuine bootstrap P value p^* will have some effect on the power of bootstrap tests, and it seems plausible that feasible bootstrap tests will be less powerful than ideal ones because of the sampling variability of $\hat{p}^*$. This issue has been explored in the literature on *Monte Carlo testing*, which is older than the literature on the bootstrap. In this section, we briefly discuss some results from the Monte Carlo testing literature, show how they apply to bootstrap testing, and present the results of some simulation experiments which illustrate just how much power can be lost in practice.

The idea of a Monte Carlo test is generally attributed to Barnard (1963). Suppose that a test statistic τ is known to be pivotal but does not have a distribution that is readily computable. For a test at level α, we calculate B artificial samples, which we may still refer to as bootstrap samples, and from them we compute B test statistics, τ_j^*, $j = 1, \ldots, B$. It is essential to choose B so that $\alpha(B+1)$ is an integer. We sort all $B+1$ test statistics, $\hat{\tau}$ and the B bootstrap ones, so that the test statistic which would lead to the most decisive rejection of the null hypothesis is ranked first. Thus if we want to reject when $\hat{\tau}$ is large, we would sort the test statistics from largest to smallest. On the other hand, if all the test statistics were in P value form, we would sort them from smallest to largest. We then see where $\hat{\tau}$ lies in the sorted list. If the rank of $\hat{\tau}$ is less than or equal to αB, we reject the null hypothesis at level α; otherwise, we do not reject it. For example, if $\alpha = .05$ and $B = 999$, we would reject the null whenever the rank of $\hat{\tau}$ is no greater than 50.

There are now two sources of randomness, the randomness of the data, which causes $\hat{\tau}$ to be random, and the randomness of the bootstrap samples. Nevertheless, this test procedure has been shown to be exact: Provided τ is pivotal, the probability of rejecting a true null hypothesis really is α; see Marriott (1979). However, the additional randomness does lead to a loss of power. This issue was first investigated for a rather special case by Hope (1968). Subsequently, Jöckel (1986) obtained some fundamental theoretical results for a fairly wide class of Monte Carlo tests. Because using a Monte Carlo test of size α based on B bootstrap samples is equivalent to calculating $\hat{p}^*$ using (1) and rejecting the null hypothesis whenever $\hat{p}^* < \alpha$, these results also apply to bootstrap tests, but only when the test statistic is pivotal.

For any test and any fixed alternative, we can define the *size-power function* $\eta(\alpha)$ as the probability that the test will reject the null when its true size is α. This function implicitly depends on the DGP, but for notational simplicity this dependence is not made explicit. This function is precisely what we estimate when we plot a size-power curve using simulation results; see Davidson and MacKinnon (1995). Because $\eta(0) = 0$ and $\eta(1) = 1$, and we need $\eta(\alpha) > \alpha$ for $0 < \alpha < 1$ if the test is to have power greater than its size, we expect the size-power function to be concave. For tests that follow standard noncentral distributions, such as the noncentral χ^2 and the noncentral F, the size-power function is indeed concave. However, it will not necessarily be concave for every test, regardless of what DGP generated the data.

Let the size-power function for a bootstrap test based on B bootstrap samples be denoted by $\eta^*(\alpha, m)$, where $\alpha(B + 1) = m$. When the original test statistic is assumed to be pivotal, $\eta(\alpha) \equiv \eta^*(\alpha, \infty)$. Under this condition, Jöckel proves the following two results:

(i) If $\eta(\alpha)$ is concave, then so is $\eta^*(\alpha, m)$.

(ii) Assuming that $\eta(\alpha)$ is concave, $\eta^*(\alpha, m + 1) > \eta^*(\alpha, m)$.

Thus increasing the number of bootstrap samples will always increase the power of the test. Just how much it will do so depends in a fairly complicated way on the shape of the size-power function $\eta(\alpha)$, and Jöckel's theoretical results on this point are not at all easy to interpret.

Although the assumption of pivotalness is essential if we are to compare the power of an asymptotic test with the power of a feasible bootstrap test, it is not needed if we just wish to compare the power of an ideal bootstrap test with the power of a feasible bootstrap test that corresponds to it. We simply have to interpret $\eta(\alpha)$ as the size-power curve for the ideal bootstrap test. Provided it is concave, Jöckel's two results apply. Thus we conclude that the feasible bootstrap test will be less powerful than the ideal bootstrap test whenever the size-power function for the ideal test is concave.

To see just how the choice of B affects test power, we conducted a small simulation experiment. We generated artificial test statistics from the very simple model

$$y_t = \gamma + u_t, \quad u_t \sim N(0, 1), \quad t = 1, \ldots, 4,$$

where the null hypothesis is that $\gamma = 0$. These test statistics were then converted to P values, either by using the c.d.f. of the $t(3)$ distribution so as to obtain p^*, or by drawing bootstrap test statistics from the $t(3)$ distribution and using equation (1) to obtain $\hat{p}^*$ for various values of B. There were 400,000 replications.

The values of B that were used were 19, 24, 39, 49, 79, 99, 199, and 399. These values were chosen because they yield valid Monte Carlo tests for commonly encountered values of α. The smallest value of B that yields a valid Monte Carlo test for $\alpha = .05$ is 19, for $\alpha = .04$, 24, for $\alpha = .025$, 39, and so on. Although B does not have to be chosen so that $\alpha(B + 1)$ is an integer for any commonly encountered values of α, it is a very good idea to do so. When this is not the case, the estimated bootstrap P value $\hat{p}^*(\hat{\tau})$ will provide a biased estimate of the true bootstrap P value $p^*(\hat{\tau})$ when the latter is equal to α. As a consequence, estimates of test size and power in a simulation experiment will not vary smoothly with α. In order to eliminate this type of bias, the size-power curves for our experiments were plotted only for values of α such that $\alpha(B+1)$ was an integer. Because the distribution of the ideal bootstrap test is known for the simple case we are examining, and because B is chosen so that test size is exact for all values of α that are plotted, we can simply plot the observed power of each test against its nominal size without having to correct for any possible size distortion.

Figure 1 presents the results of two out of the five experiments that we performed. The horizontal axis shows test size and the vertical axis shows the difference between the power of the test based on $B = \infty$ and the power of the test based on some finite value of B. For $B \leq 79$, the points at which power loss is evaluated are shown as balls. In the top panel of the figure, $\gamma = 4$. This value of γ implies that the test based on $B = \infty$ is reasonably powerful; it has power 0.757 for a test at the .05 level. We see that, in this case, power loss can be quite substantial. For $B = 19$, the loss can be nearly .15 at the .05 level, and it would clearly be even larger than this at smaller test sizes if we could perform valid tests using such a small value of B. Even for $B = 399$, power loss is greater than .01 for all test sizes less than .05. In the bottom panel of the figure, $\gamma = 2$. This value of γ implies that the test based on $B = \infty$ is not very powerful; it has power 0.290 for a test at the .05 level. In this second case, the power loss is also much smaller, at least for the small test sizes that are normally of interest. For large test sizes, however, the power loss is actually greater in this case.

The results shown in Figure 1, along with unreported results for $\gamma = 1$, $\gamma = 3$, and $\gamma = 5$, make it clear that the power loss from bootstrapping depends in a complicated way on the shape of the size-power function. It is relatively large when the ideal bootstrap test is quite powerful and the size-power function displays a lot of curvature. On the other hand, it can be very small when the size-power curve is straight, either because the test has a great deal of power and the curve is nearly horizontal, or because the test has little power and the curve is close to the 45° line.

In view of these results, it would seem advisable to use a fairly large number of bootstrap samples. Because it is the easiest approach, most investigators will simply want to pick B in advance and do a single set of experiments. When this approach is used, we recommend that B be at least 399. However, if computational cost is not a serious problem, 999 would be a better choice, 1999 would be even better, and even 9999 might not be excessive. When computational cost is a serious problem, it may be best to start with a small value of B and then increase it if the results are inconclusive.

5. Measuring Test Power in Monte Carlo Experiments

The theoretical results of Section 3 did not require us to say anything about the specification of μ_1 or the choice of a μ_0 that corresponds to μ_1. However, if we wish to study the power of any test, whether asymptotic or bootstrap, by using Monte Carlo experiments, we will need to specify μ_1 explicitly. Moreover, unless we confine our attention to power unadjusted for size, which is not a very interesting thing to study unless all the tests always perform extremely well under the null, we will need to choose a μ_0 that corresponds to each μ_1 in order to compute size-corrected power. Unfortunately, there is no unique way to choose such a μ_0, and there is therefore no unique way to measure size-corrected power in a Monte Carlo experiment.

Consider a model with two parameters, θ and δ, where the null hypothesis is that $\delta = 0$. If μ_1 has parameters (θ_1, δ_1), the obvious DGP to use for μ_0 is the one with parameters $(\theta_1, 0)$. We shall call this DGP the *naive null.* As the name we have given it implies, this choice for μ_0 is not very satisfactory. Indeed, there are at least two difficulties with the naive null. The first is that it may be a long way from the actual DGP, much further than many other DGPs that also satisfy the null hypothesis. The second is that the naive null depends on the way the alternative model is parametrized. For instance, if instead of θ and δ we were to use $\theta + \delta$ and δ as parameters, then the naive null, in the old parametrization, would be the DGP with parameters $\theta_1 + \delta_1$ and 0. It would clearly be preferable to choose a DGP that satisfies the null in a parametrization-independent fashion. Thus it appears that the size of the test under the naive null is not what we want to use to compute size-corrected power. A better choice, which we now discuss, is the *pseudo-true null.*

Asymptotically at least, the closest null to a given fixed DGP is the null DGP characterized by the *pseudo-true values*, in the sense of White (1982), that correspond to the fixed DGP. The vector of pseudo-true values is defined as the probability limit of the quasi-maximum likelihood estimator of the null hypothesis under the fixed DGP. White shows that the pseudo-true values are the parameters of the DGP in the null hypothesis that minimize the Kullback-Leibler Information Criterion (KLIC) with respect to the fixed DGP. In practice, it is convenient simply to define the closest DGP in the null to be the one that minimizes the KLIC. In most cases of interest, although the KLIC formally depends on sample size, it turns out that the parameters of the KLIC-minimizing DGP are independent of the sample size. Note that the KLIC is a quantity defined purely in terms of two DGPs, quite independently of how those DGPs may be parametrized.

If we start from a given DGP μ_1 for a given sample size n, the drifting DGP through μ_1 suitable for power analysis has an end point in the null, μ_0^1, which minimizes the KLIC to it from μ_1. We will define the end point μ_0^1 as the *pseudo-true null.* Since we cannot perform a size correction of a nonpivotal test without choosing a specific null DGP, it appears that the pseudo-true null μ_0^1 is the most reasonable one to choose. While this choice is inevitably somewhat arbitrary, it has the advantages of being defined in a parametrization-independent manner and of introducing no unnecessary dependence on the sample size. However, it has the disadvantage that it may not always be easy to compute analytically, especially for dynamic models.

Since the pseudo-true null is an asymptotic concept, it is not entirely clear that it is the best μ_0 to use in finite samples. However, Horowitz (1995) shows that a bootstrap test is asymptotically equivalent to an exact test of a simple null hypothesis consisting of just one DGP, namely the pseudo-true null. At least for bootstrap tests, this is another indication that the pseudo-true null is the most appropriate DGP to use for size correction even in finite samples. As we shall see in the next section, the choice of which μ_0 to use for correcting the size of an asymptotic test can have very substantial effects on size-corrected power.

6. Testing for Omitted Variables in a Logit Model

In this section, we present the results of some Monte Carlo experiments. These illustrate the principal theoretical results of Section 3 and highlight the importance of using the pseudo-true null rather than the naive null when size-correcting asymptotic tests. The experiments deal with Lagrange multiplier tests for omitted variables in the logit model. We chose to examine the logit model because it is not a regression model, and because the results of Horowitz (1994) and Davidson and MacKinnon (1995) suggest that, for information matrix tests in the closely related probit model, bootstrapping may greatly improve the finite-sample properties of one form of the LM test.

The logit model that we are dealing with may be written as

$$E(y_t) = F_t(\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Z}_t\boldsymbol{\gamma}) \equiv \left(1 + e^{-\boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{Z}_t\boldsymbol{\gamma}}\right)^{-1}, \tag{14}$$

where y_t is an observation on a dependent variable that is either 0 or 1, $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$ are, respectively, a $1 \times k$ vector and a $1 \times r$ vector of regressors, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are, respectively, a $k \times 1$ vector and an $r \times 1$ vector of unknown parameters. Under the null hypothesis, $\boldsymbol{\gamma} = \mathbf{0}$. This null hypothesis may be tested in many ways. Two of the easiest are to use tests based on artificial regressions. We shall consider two such tests. The first is the outer product of the gradient, or OPG, variant of the LM test, and the second is the efficient score, or ES, variant. No sensible person would use the OPG variant in preference to the ES variant without bootstrapping, since the asymptotic form of the OPG variant has considerably worse finite-sample properties under the null (Davidson and MacKinnon, 1984). However, in the related context of information matrix tests for binary response models, Horowitz (1994) found that the

OPG variant worked well when bootstrapped, although he did not compare it with the ES variant.

Suppose that we estimate the logit model (14) under the null hypothesis, obtain restricted ML estimates $\tilde{\boldsymbol{\beta}}$, and use them to calculate $\tilde{F}_t \equiv F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$ and $\tilde{f}_t \equiv f(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$, where $f(\cdot)$ is the first derivative of $F(\cdot)$. Then the OPG test statistic is n minus the sum of squared residuals from the artificial regression with typical observation

$$1 = \frac{\tilde{f}_t(y_t - \tilde{F}_t)}{\tilde{F}_t(1-\tilde{F}_t)}\left(\sum_{i=1}^{k} X_{ti}b_i + \sum_{i=1}^{r} Z_{ti}g_i\right) + \text{residual}. \tag{15}$$

and the ES test statistic is the explained sum of squares from the artificial regression with typical observation

$$\frac{y_t - \tilde{F}_t}{\big(\tilde{F}_t(1-\tilde{F}_t)\big)^{1/2}} = \frac{\tilde{f}_t}{\big(\tilde{F}_t(1-\tilde{F}_t)\big)^{1/2}}\left(\sum_{i=1}^{k} X_{ti}b_i + \sum_{i=1}^{r} Z_{ti}g_i\right) + \text{residual}. \tag{16}$$

In regressions (15) and (16), the b_i and g_i are parameters to be estimated, and the first factors on the right-hand side are weights that multiply all the regressors. In both regressions, the regressors with parameters b_i are orthogonal to the regressand; the tests are really asking whether the regressors with parameters g_i have any explanatory power.

In order to size-correct the asymptotic tests, it is necessary to compute the parameters of the pseudo-true null that corresponds to whatever DGP actually generated the data. If $h(\boldsymbol{y}, \boldsymbol{\beta}_0)$ denotes the joint density of the data $\boldsymbol{y}$ for the pseudo-true model and $g(\boldsymbol{y}, \boldsymbol{\beta}_1, \gamma_1)$ denotes the joint density for the DGP, the KLIC is

$$E\Big(\log\big(g(\boldsymbol{y}, \boldsymbol{\beta}_1, \gamma_1)/h(\boldsymbol{y}, \boldsymbol{\beta}_0)\big)\Big). \tag{17}$$

The parameter vector for the pseudo-true null is the $\boldsymbol{\beta}_0$ that minimizes (17). For the model (14), this is just the vector that maximizes

$$\sum_{t=1}^{n} E\Big(y_t \log(F(\boldsymbol{X}_t\boldsymbol{\beta}_0) + (1-y_t)\log\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta}_0)\big)\Big). \tag{18}$$

Since the only random things here are the y_t, maximizing (18) is equivalent to maximizing

$$\sum_{t=1}^{n}\Big(E(y_t)\log(F(\boldsymbol{X}_t\boldsymbol{\beta}_0) + \big(1 - E(y_t)\big)\log\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta}_0)\big)\Big). \tag{19}$$

The expectations in (19) are to be taken with respect to the DGP characterized by $\boldsymbol{\beta}_1$ and γ_1, and so, from (14), we see that $E(y_t) = F_t(\boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{Z}_t\gamma_1)$. Thus it is quite easy to compute the parameters $\boldsymbol{\beta}_0^1$ of the pseudo-true null that corresponds to any

DGP. We simply need to replace $E(y_t)$ in (19) by $F_t(\boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{Z}_t\boldsymbol{\gamma}_1)$ and maximize it using a logit routine that is able to handle dependent variables which can take on any value in the $(0, 1)$ interval.

We performed two sets of experiments. For case 1, the vector $\boldsymbol{X}_t$ consisted of a constant term and a single regressor that was distributed as $N(0, 1)$. The vector $\boldsymbol{Z}_t$ consisted of 8 regressors that were normally distributed and positively correlated with the single regressor in $\boldsymbol{X}_t$; the squared correlation was 0.5. This correlation ensures that the pseudo-true and naive nulls will generally be quite different, except when the DGP is very close to the naive null. We deliberately made the number of elements of $\boldsymbol{Z}_t$ fairly large, because preliminary experiments suggested that the OPG test, in particular, performed less well as the number of degrees of freedom increased. In order to allow us to plot one-dimensional power functions, we set $\boldsymbol{\gamma}_1 = \delta\boldsymbol{\iota}$, where $\boldsymbol{\iota}$ is a vector of 1s. Thus the only parameter of the DGP that we changed was δ. Because the pseudo-true null depends on the regressors, we used a single set of regressors in all the experiments, and all our experimental results are conditional on them. We set the constant term in $\boldsymbol{\beta}_1$ to 0 and the slope coefficient to 1. Thus, under the null hypothesis, approximately half of the y_t would be 0 and half would be 1.

The sample size for case 1 was 100. This may seem large, but it was necessary to use a reasonably large sample size in order to avoid having to throw out too many replications when the logit routine failed to converge. Nonconvergence, which is almost always caused by perfect classifiers, is more of a problem for smaller sample sizes and for larger values of δ. In order to study tests with reasonable power, it is necessary to consider larger values of δ the smaller is the sample size. Thus, for two reasons, the nonconvergence problem would have been much more severe if the sample size had been smaller. Even for the bootstrap samples, there were no cases of nonconvergence for $-0.6 \leq \delta \leq 0.2$. In the worst case that we bootstrapped, $\delta = 1.4$, less than 0.05% of the logit estimations failed to converge.

Figures 2 and 3 present the principal results of our experiments for case 1 in the form of power functions for asymptotic and bootstrap tests at the .05 level. Figure 2 shows power functions for the ES tests, and Figure 3 shows them for the OPG tests. The power functions for the asymptotic tests are based on 100,000 replications for a large number of values of δ between -1.5 and 1.5; the specific values are not shown. The unadjusted power function (the solid line) shows test power at the nominal .05 level. The two adjusted power functions (the dotted lines) show test power at the "true" .05 level, calculated in two different ways. For all values of δ, the naive adjustment method uses test performance for $(\boldsymbol{\beta}_0, \mathbf{0})$ as a benchmark. In contrast, the pseudo-true adjustment method uses test performance for $(\boldsymbol{\beta}_0^1, \mathbf{0})$ as a benchmark. We therefore had to perform two experiments for every value of δ except $\delta = 0$. In one experiment, the data were generated by the DGP with parameters $(\boldsymbol{\beta}_1, \delta\boldsymbol{\iota})$, and in the other they were generated by the DGP with parameters $(\boldsymbol{\beta}_0^1, \mathbf{0})$.

The adjusted power functions were calculated from estimates of the empirical distribution functions (EDFs) of the test statistic, in P value form, under the alternative and under either the naive or the pseudo-true null. We estimated power as a function of true size, using local polynomial regressions on values of the EDFs in the neighborhood of true size .05, and then used the fitted value at .05 as the estimate of adjusted power. This is essentially equivalent to plotting the two EDFs against each other, as in a size-power plot, and then reading off the value of power that corresponds to true size .05.

In addition to the three power functions for the asymptotic tests, Figures 2 and 3 also show the results of a number of Monte Carlo experiments for bootstrap tests. Each experiment involved 50,000 replications. For each replication, we generated the data in exactly the same way as before, using the DGP with parameters $(\boldsymbol{\beta}_1, \delta\boldsymbol{\iota})$, computed both test statistics, and then used the parametric bootstrap based on 399 bootstrap samples to estimate a P value for each of them. The bootstrap test rejected the null hypothesis if the estimated P value was less than .05. The bullets in the figures show the proportion of replications for which this procedure led to rejection. The power of the bootstrap tests has not been size-adjusted, because the theory of Section 3 says that there is no need to do so, and it would have doubled the already very high computational costs of these experiments.

The results for case 1 for the ES test are shown in Figure 2. The test works so well in this experiment that there is evidently no need to bootstrap it. As the theory of Section 3 predicts, the power of the bootstrap test is always extremely close to the adjusted power of the asymptotic test based on the pseudo-true null, which we will henceforth refer to as PT-adjusted power. However, as there is very little difference between any of the power functions, the fact that the power of the bootstrap test is a little closer to the PT-adjusted power function than it is to the other two power functions does not provide much evidence in support of the theory.

In contrast, Figure 3 does provide strong support for the theory of Section 3. The three power curves are quite different, and the power of the bootstrap test is always very much closer to the PT-adjusted curve than it is to either of the other two curves. However, as the theory predicts, the power of the bootstrap test is not quite identical to the PT-adjusted power of the asymptotic test. This may partly be attributable to power loss from the use of only 399 bootstrap samples. The differences between the naive-adjusted and PT-adjusted curves may seem surprisingly large, in view of the fact that the OPG test, although it clearly overrejects, certainly does not perform terribly when $\delta = 0$. Evidently, the OPG test must overreject much more severely for some pseudo-true values $(\boldsymbol{\beta}_0^1, \mathbf{0})$, especially ones associated with positive values of δ, than it does for $(\boldsymbol{\beta}_1, \mathbf{0})$.

Case 2 differs from case 1 in four respects. The single nonconstant regressor in $\boldsymbol{X}_t$ is now distributed as $\chi^2(3)$, recentered and rescaled to have mean 0 and variance 1; the correlation between it and the regressors in $\boldsymbol{Z}_t$ is now negative, with squared correlation 0.6; the sample size is now 150; and $-1.1 \leq \delta \leq 1.1$. The larger sample size and smaller range of values for δ were needed to avoid too many replications for which the logit routine did not converge. Nonconvergence was actually less of a problem in case 2 than in case 1; it never occurred for $-0.3 \leq \delta \leq 0.9$, and it occurred less than .01% of the time for the worst case that we bootstrapped, $\delta = -1.0$.

For case 2, the asymptotic form of the ES test performed so well, rejecting the null 4.77% of the time at the 5% level, that there is little reason to use the bootstrap. With 100,000 replications, we can be very confident (at the .01 level) that the true size of the test is less than .0495 and greater than .0459. Thus there is clearly some size distortion, but it is very small. The three asymptotic power functions are about as close to each other as they are in Figure 2, and the power of the bootstrap test always lies very close to the PT-adjusted power function. Once again, the ES test is more powerful than the OPG test, substantially so, in this case, for large, negative values of δ. Since these results are so similar to those in Figure 2, we do not present them graphically.

Results for the OPG test are presented in Figure 4, which is comparable to Figure 3 except that the scale of the horizontal axis is different. Once again, the PT-adjusted power function is quite different from the naive-adjusted one, but now the differences are most striking for negative values of δ. The power of the bootstrap test is very much closer to the PT-adjusted curve than it is to either of the other two curves, but it is noticeably below that curve for large, negative values of δ. Although the lower power of the bootstrap test may be due in part to the use of only 399 bootstrap samples, the difference is too large to be explained by this factor alone.

Horowitz (1994) suggested that, when bootstrapping is expensive, it would be possible to estimate the power of bootstrap tests without actually doing any bootstrapping. His suggestion was, essentially, to compute PT-adjusted power functions for the asymptotic tests on which the bootstrap tests are based. It is evident from Figures 2 through 4 that this procedure would work well, but not perfectly, for the tests dealt with here.

Although this section has focused on the OPG form of the LM test, practitioners are strongly advised never to use it. In our experiments, the ES test was always at least as powerful, on a size-adjusted basis, as the OPG test, and it was often substantially more powerful. Moreover, the asymptotic form of the ES test performed so well that bootstrapping was really not necessary. Of course, bootstrapping might well be necessary for smaller sample sizes, models with different regressors and parameter values, and so on.

7. Conclusions

In this paper, we have shown that the power of a bootstrap test differs from the size-adjusted power of the asymptotic test on which it is based by an amount that is of the same order, in the sample size, as the size distortion of the bootstrap test itself. Since this size distortion is generally either $O(n^{-3/2})$ or $O(n^{-2})$, there is good reason to believe that, in practice, bootstrap tests will have power very similar to the power of the underlying asymptotic tests. This theoretical result was confirmed and illustrated, for the case of tests for omitted variables in logit models, by the Monte Carlo results in Section 6.

The theoretical results of the paper are for the ideal bootstrap test, that is, a test based on an infinite number of bootstrap samples. As we discussed in Section 4, there will inevitably be some loss of power when the number of bootstrap samples is finite. This loss of power should be minimal if the number of bootstrap samples is reasonably large; 399 is the smallest number we would recommend. The number 399 is chosen, in part, because it is desirable that the number of bootstrap samples B should be such that $\alpha(B+1)$ is an integer, where α is the size of the test.

The theoretical and simulation results of the paper all depend on the size-adjustment of the asymptotic test being carried out using test size under what we call the pseudo-true null hypothesis. As we discussed in Section 5, it is very important to use this method of size adjustment, even in Monte Carlo experiments that do not involve bootstrap testing, if we wish accurately to compare the size-adjusted power of competing tests.

References

Barnard, G. A. (1963). "Contribution to discussion," *Journal of the Royal Statistical Society*, Series B, 25, 294.

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

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

Davidson, R. and J. G. MacKinnon (1995). "Graphical methods for investigating the size and power of hypothesis tests," revised version of Queen's Institute for Economic Research Discussion Paper No. 903, 1994.

Davidson, R. and J. G. MacKinnon (1996). "The size distortion of bootstrap tests," GREQAM Document de Travail No. 96A15.

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

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

Hope, A. C. A. (1968). "A simplified Monte Carlo significance test procedure," *Journal of the Royal Statistical Society*, Series B, 30, 582–598.

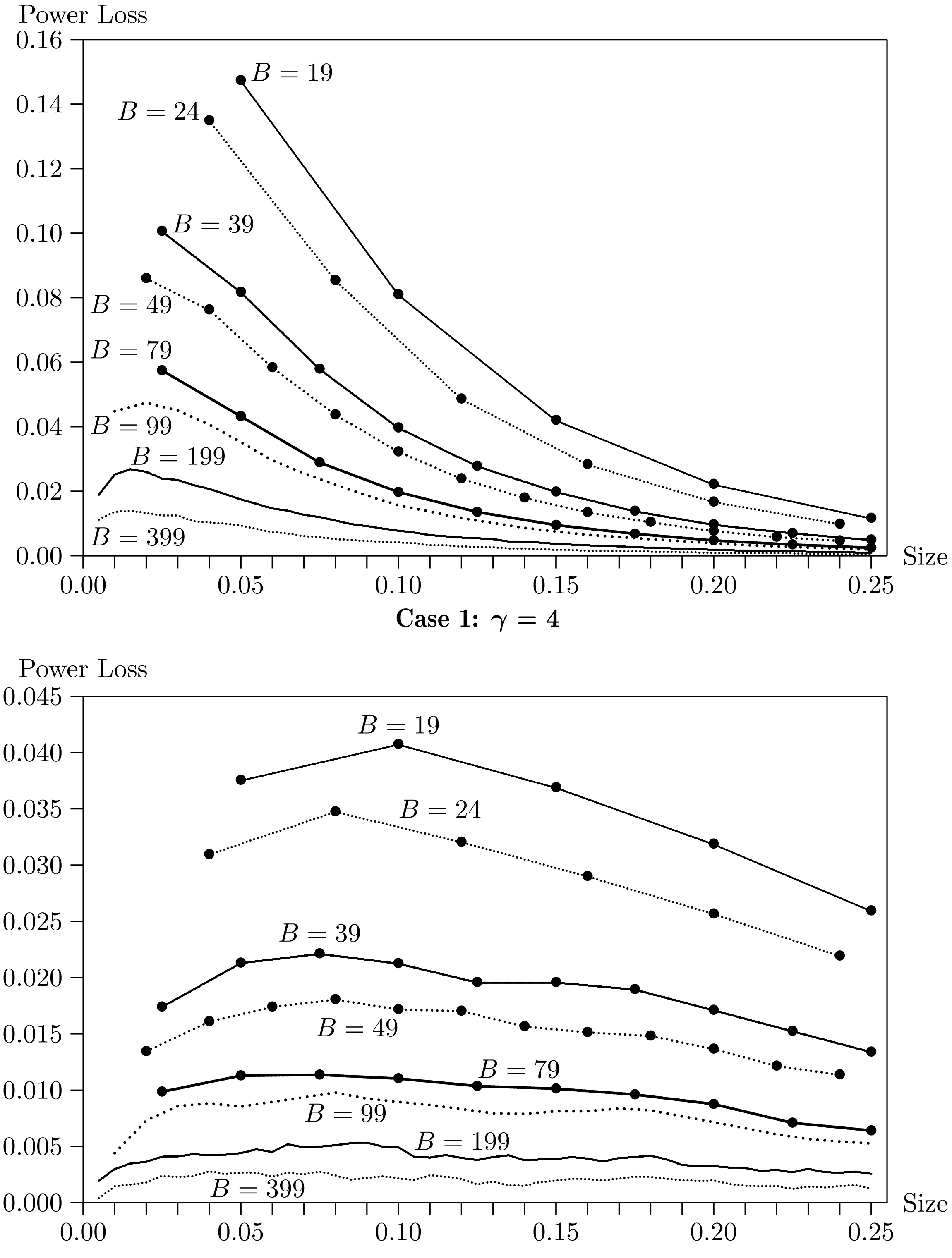
Horowitz, J. L. (1994). "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, 61, 395–411.

Horowitz, J. L. (1995). “Bootstrap methods in econometrics: Theory and numerical performance,” paper presented at the 7th World Congress of the Econometric Society, Tokyo.

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

Marriott, F. H. C. (1979). “Barnard’s Monte Carlo tests: How many simulations?,” *Applied Statistics*, 28, 75–77.

White, H. (1982). “Maximum likelihood estimation of misspecified models,” *Econometrica*, 50, 1–26.



Case 1: $\gamma = 4$

Case 2: $\gamma = 2$

Figure 1. Power Loss from Bootstrapping

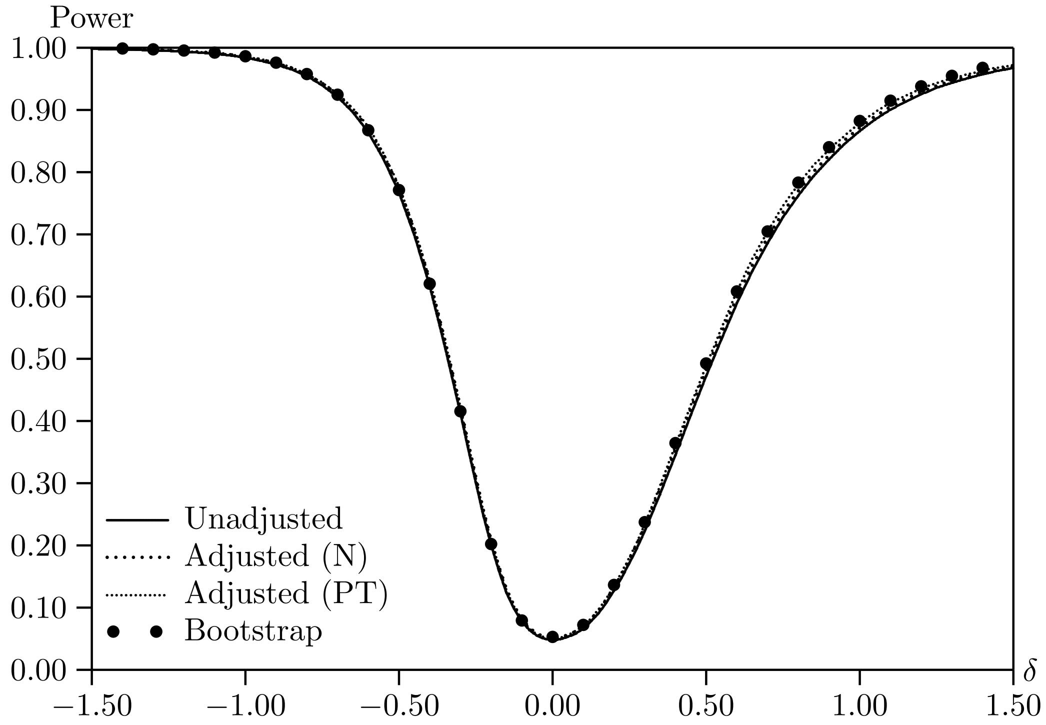


Figure 2. Power functions for logit ES tests, Case 1

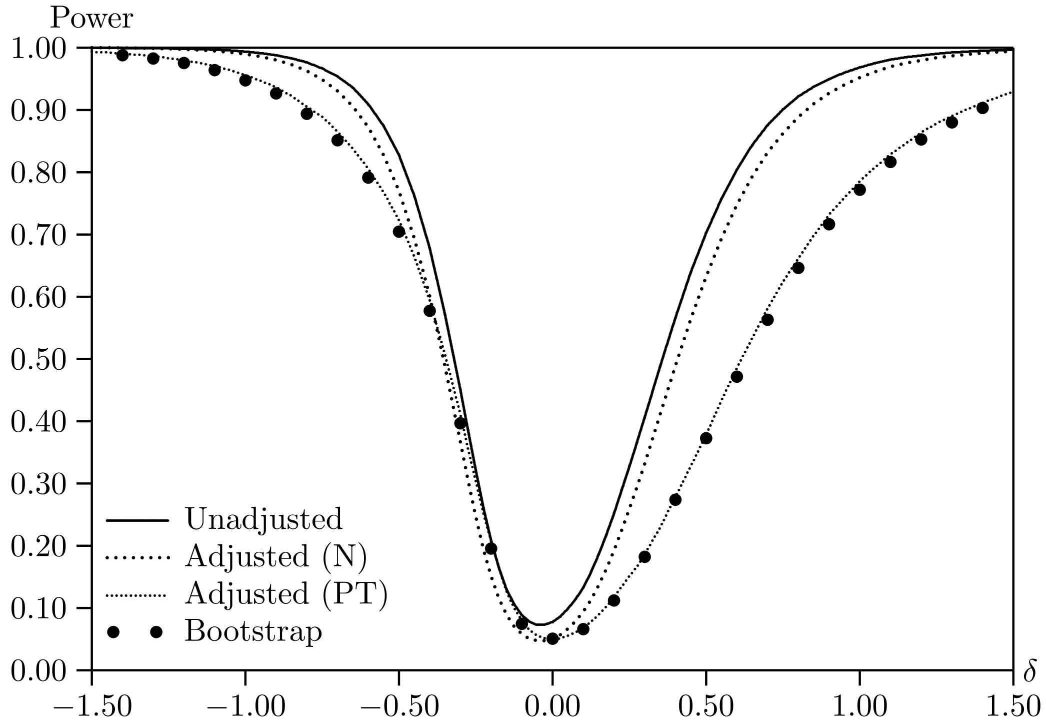


Figure 3. Power functions for logit OPG tests, Case 1

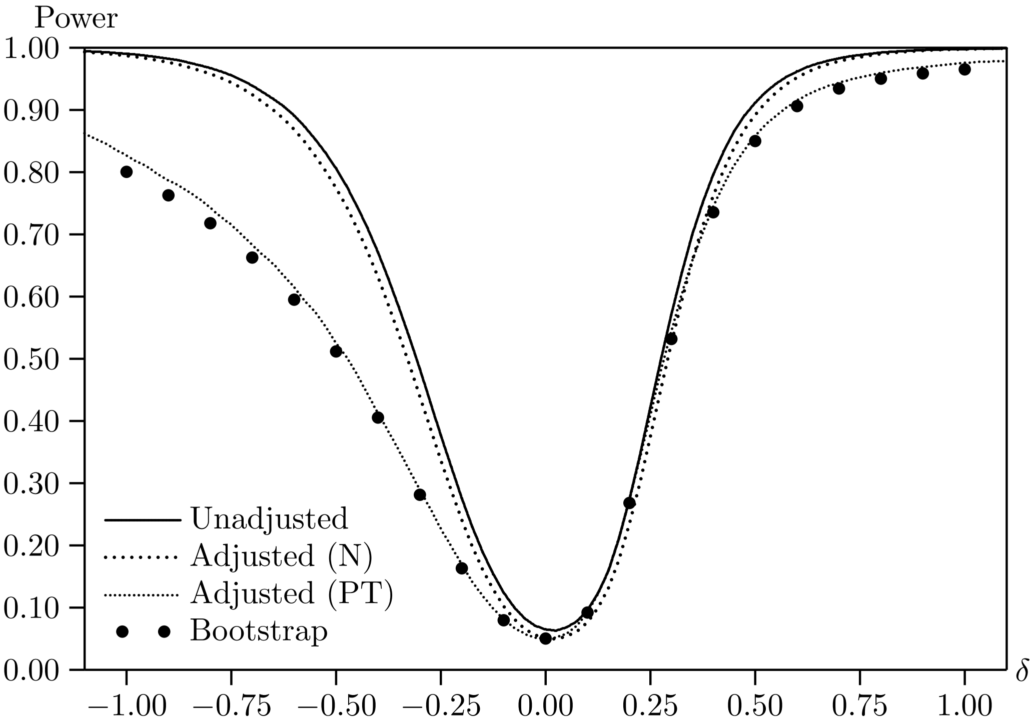


Figure 4. Power functions for logit OPG tests, Case 2