The Power of Bootstrap and Asymptotic Tests

by

Russell Davidson

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

Department of Economics
McGill University
Montreal, Quebec, Canada
H3A 2T7

email: **russell@ehess.cnrs-mrs.fr**

and

James G. MacKinnon

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

email: **jgm@qed.econ.queensu.ca**

Abstract

We show that the power of a bootstrap test will generally be very close to the level-adjusted power of the asymptotic test on which it is based, provided the latter is calculated properly. Our result, when combined with previous results on approximating the rejection frequency of bootstrap tests, provides a way to simulate the power of both asymptotic and bootstrap tests easily and inexpensively. Some Monte Carlo results for omitted variable tests in logit models illustrate the theoretical results of the paper, demonstrate that the level-adjusted power of asymptotic tests can vary greatly depending on the method used for level adjustment, and show how useful our approximate method can be.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Don Andrews, Joel Horowitz, two referees, and numerous seminar participants for comments on earlier versions.

November, 2002

1. Introduction

In recent years, it has become common to use the bootstrap to perform hypothesis tests in econometrics. Its use for this purpose has been advocated by Horowitz (1994, 1997), Davidson and MacKinnon (1999), and several others. If the bootstrap is to benefit from asymptotic refinements, the original test statistic must be asymptotically pivotal under the null hypothesis, that is, its asymptotic distribution under the null must not depend on any unknown features of the data generating process (DGP). With such a test statistic, the errors committed by using the bootstrap are generally of an order lower by a factor of either $n^{-1/2}$ or n^{-1}, where n is the sample size, than the errors committed by relying on asymptotic theory; see Beran (1988), Hall (1992), Davison and Hinkley (1997), or Davidson and MacKinnon (1999).

A convenient way to perform bootstrap inference is to compute bootstrap P values. After computing a test statistic, say τ, in the usual way, one uses a random bootstrap DGP, denoted by μ^* and constructed so as to satisfy the null hypothesis under test, to generate B bootstrap samples, each of which is used to compute a bootstrap test statistic τ_j^*, $j = 1, \ldots, B$. The bootstrap P value may then be estimated by the proportion of bootstrap statistics that are more extreme than τ. As $B \to \infty$, this estimated bootstrap P value will tend to the "ideal" bootstrap P value $p^*(\tau)$, which is defined as

$$p^*(\tau) \equiv \Pr_{\mu^*}\big(\mathrm{Rej}(\tau)\big),$$

where $\mathrm{Rej}(\tau)$ is the rejection region for a test for which the critical value is τ. For a one-tailed test that rejects in the upper tail, for instance, $\mathrm{Rej}(\tau)$ is just the set of real numbers greater than τ. In this paper, we ignore the fact that $p^*(\tau)$ has to be estimated. The effect of the estimation error can be made as small as desired by appropriate choice of B; see Davidson and MacKinnon (2000).

If the original data are generated by a DGP μ, the "true" P value $p(\tau) \equiv \Pr_{\mu}\big(\mathrm{Rej}(\tau)\big)$, which is just a deterministic function of τ, is by construction a drawing from the uniform distribution $U(0, 1)$. But, since the bootstrap DGP μ^* is a function of the data, the bootstrap P value $p^*(\tau)$ is in general drawn from a different distribution. Consequently, the rejection probability (RP) of a bootstrap test at nominal level α is in general different from α, even when μ satisfies the null hypothesis under test.

It is natural to ask whether bootstrapping a test has any effect on its power. Answering this question is complicated by the fact that asymptotic tests often suffer from substantial size distortion. In simulation studies, it is common to adjust for this distortion by using critical values for which the RP under some DGP μ_0 that satisfies the null hypothesis is exactly equal to the desired nominal level. With statistics that are not exactly pivotal, the adjustment depends on the specific choice of μ_0.

Conventional asymptotic power analysis relies on the notion of a *drifting DGP*, which, as the sample size tends to infinity, drifts to μ_0. In order to study the difference between the power of a bootstrap test and the adjusted power of the asymptotic test on which it is based, which we call the *bootstrap discrepancy*, a suitable drifting DGP must be chosen. We demonstrate in Section 2 that, for any choice of drifting DGP, the bootstrap

discrepancy may be of either sign and is, in general, of the same order in n as the size distortion of the bootstrap test.

In Section 3, we consider how best to choose the drifting DGP. We argue that the objective should be minimization of the bootstrap discrepancy, and we show that this is feasible only if τ and μ^* are asymptotically independent in a sense that we make precise. In Davidson and MacKinnon (1999), we showed that asymptotic independence of this sort leads to a reduction in the order of bootstrap size distortion. We characterize a class of drifting DGPs that serves to extend this result to the bootstrap discrepancy.

In Section 4, we propose an extension to power analysis of a procedure given in Davidson and MacKinnon (2001) for estimating the RP of a bootstrap test by simulation. This procedure, which is conceptually simple and computationally inexpensive, allows one to estimate the power of bootstrap and asymptotic tests inexpensively in Monte Carlo experiments. In Section 5, we present some Monte Carlo results for tests of omitted variables in a logit model. Section 6 concludes.

2. The Power of Bootstrap and Asymptotic Tests

Suppose that a test statistic t has a fixed known asymptotic distribution under the null hypothesis, represented by a probability measure P^∞ defined on the real line and absolutely continuous with respect to Lebesgue measure. It is convenient to replace t by another test statistic, which we denote by τ, of which the nominal asymptotic distribution is uniform on $[0, 1]$. This is most conveniently done by replacing t by its asymptotic P value, so that τ is given by $P^\infty(\text{Rej}(t))$, the probability mass in the part of the asymptotic distribution that is more extreme than t. The asymptotic test based on t rejects the null hypothesis at level α whenever $\tau < \alpha$. In the remainder of this section, without loss of generality, we consider statistics in this P value form.

To discuss power, we must consider DGPs that do not satisfy the null hypothesis. The asymptotic theory of power makes use of nonnull drifting DGPs, which are determined by a DGP belonging to the null, plus a perturbation that is usually $O(n^{-1/2})$; see Davidson and MacKinnon (1993, Chapter 12). The appropriate rate, usually $n^{-1/2}$, at which a nonnull DGP μ drifts towards the null is chosen so that the RP of the test associated with τ tends neither to 0 nor to 1 as $n \to \infty$ for levels α different from 0 or 1. Note that, even if the test is associated with a specific alternative hypothesis, an asymptotic power analysis does not require that a nonnull drifting DGP should belong to it; see Davidson and MacKinnon (1987).[1]

In what follows, we limit ourselves to parametric null hypotheses. By this, we mean that the set of DGPs $\mathbb{M}_0$ that satisfy the null are in one-one correspondence with the elements of a k–dimensional parameter space Θ. We assume that the drifting DGP of interest to us can be embedded in a $(k+1)$–dimensional model, comprised of the DGPs in the set $\mathbb{M}_1$, parametrized by $\Theta \times U$, where U is an interval in $\mathbb{R}$ with the origin as an interior point. The DGP that corresponds to the parameter vector $(\boldsymbol{\theta}, \delta) \in \Theta \times U$

[1] In that paper, we did not use the term "drifting DGP". Rather, we spoke of a "sequence of local DGPs". We much prefer the newer terminology, which has the advantage of making clear the link with Pitman drift.

belongs to the null model if and only if $\delta = 0$. The drifting DGP itself is such that, for sample size n, it is characterized by the parameters $(\boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{t}, n^{-1/2}\delta_0)$, for some $\boldsymbol{\theta}_0 \in \Theta$ and $\delta_0 \in U$, and some k–vector $\boldsymbol{t}$. This drifting DGP drifts towards the DGP in $\mathbb{M}_0$, denoted μ_0, that corresponds to the parameters $(\boldsymbol{\theta}_0, 0)$ for all n. Such a DGP, where the parameters are independent of n, will be called a *fixed DGP*.

The parametric bootstrap DGP μ^* is the DGP whose parameters are given by an estimator $\hat{\boldsymbol{\theta}}$ that is consistent under the null, and $\delta = 0$, so that $\mu^* \in \mathbb{M}_0$ by construction. The maximum likelihood estimator of the model $\mathbb{M}_0$ is asymptotically efficient, and so it is a sensible choice for $\hat{\boldsymbol{\theta}}$, but other consistent estimators, for instance the MLE for a model that represents the alternative hypothesis for the test, can be used without affecting the results to be developed in this section. Since $\hat{\boldsymbol{\theta}}$ depends on n, so does μ^*, which is thus a drifting DGP that drifts entirely within $\mathbb{M}_0$. Under a fixed DGP $\mu \in \mathbb{M}_0$ with parameter vector $\boldsymbol{\theta}_0$, $\operatorname{plim} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, and so μ^* drifts towards μ_0. In fact, under weak regularity conditions to be specified later, $\hat{\boldsymbol{\theta}}$ also has a plim of $\boldsymbol{\theta}_0$ under DGPs that drift to $\boldsymbol{\theta}_0$, from which it follows that μ^* drifts to μ_0 in this case as well.

We can analyze the performance of a test based on τ for a given sample size n by use of two functions that depend on the nominal level α of the test and the DGP μ. The first of these is the *rejection probability function*, or RPF. This function, which gives the true rejection probability under μ of a test at nominal level α, is defined as

$$R(\alpha, \mu) \equiv \Pr_\mu(\tau \leq \alpha). \tag{1}$$

In this definition, everything except α should properly be indexed by n, but we prefer an uncluttered notation without this explicit indexing. Throughout, we assume that, under any DGP μ we consider, the distribution of τ has support $[0, 1]$ and is absolutely continuous with respect to the uniform distribution on that interval.

For given μ and n, $R(\alpha, \mu)$ is just the CDF of τ evaluated at α. The inverse of the RPF is the *critical value function*, or CVF, which is defined implicitly by the equation

$$\Pr_\mu\big(\tau \leq Q(\alpha, \mu)\big) = \alpha. \tag{2}$$

It is clear from (2) that $Q(\alpha, \mu)$ is the α quantile of the distribution of τ under μ. In addition, the definitions (1) and (2) imply that

$$R\big(Q(\alpha, \mu), \mu\big) = Q\big(R(\alpha, \mu), \mu\big) = \alpha \tag{3}$$

for all α and μ. As $n \to \infty$, $R(\alpha, \mu)$ and $Q(\alpha, \mu)$ both tend to α for all DGPs $\mu \in \mathbb{M}_0$. If an asymptotic test is exact in finite samples, then we have $R(\alpha, \mu) = \alpha$ and $Q(\alpha, \mu) = \alpha$ for all α, for all n, and for all μ in the null.

The bootstrap critical value for τ at nominal level α is $Q(\alpha, \mu^*)$. This is a random variable which is asymptotically nonrandom and equal to α, since, whether or not the true DGP belongs to the null hypothesis, the parametric bootstrap DGP μ^* does so. Any size distortion of the bootstrap test under a DGP μ in the null arises from the possibility that, in a finite sample, $Q(\alpha, \mu^*) \neq Q(\alpha, \mu)$.

A bootstrap test based on τ rejects at nominal level α if $\tau < Q(\alpha, \mu^*)$. Therefore, applying the increasing transformation $R(\cdot, \mu^*)$ to both sides and using (3), we see that

the bootstrap test rejects whenever

$$R(\tau, \mu^*) < R\big(Q(\alpha, \mu^*), \mu^*\big) = \alpha. \tag{4}$$

Thus the bootstrap P value is just $R(\tau, \mu^*)$. This can be interpreted as a bootstrap test statistic. The probability under μ that the bootstrap test rejects at nominal level α is

$$\Pr{}_\mu\big(\tau < Q(\alpha, \mu^*)\big) = \Pr{}_\mu\big(R(\tau, \mu^*) < \alpha\big). \tag{5}$$

For all sample sizes, and for all DGPs, fixed or drifting, in $\mathbb{M}_1$, let the random variable p be defined by

$$p = R(\tau, \mu). \tag{6}$$

Since $R(\cdot, \mu)$ is the CDF of τ under μ, p is distributed as $U(0,1)$ for all n and for all μ. Further, define the random variable q as

$$q = R\big(Q(\alpha, \mu^*), \mu\big) - R\big(Q(\alpha, \mu_0), \mu\big), \tag{7}$$

where, if μ is a fixed DGP in $\mathbb{M}_0$, $\mu_0 = \mu$. If instead μ is a drifting DGP, then μ_0 is the fixed DGP in $\mathbb{M}_0$ to which it drifts. Instead of applying the transformation $R(\cdot, \mu^*)$ to both sides of the inequality $\tau < Q(\alpha, \mu^*)$, as we did to obtain (4), we can apply the transformation $R(\cdot, \mu)$ to both sides of this inequality. When we do this and use (6), we see that rejection by the bootstrap test is equivalent to the inequality

$$p < R\big(Q(\alpha, \mu_0), \mu\big) + q. \tag{8}$$

The first term on the right-hand side of (8) is the RP under μ of the asymptotic test when the true α-level critical value of the DGP μ_0 is used. If $\mu = \mu_0$, it is equal to α. If not, it is what would usually be called the size-corrected, or level-adjusted,[2] power of the asymptotic test under μ at level α.

From equation (7), it is clear that q is just the difference in the RPs under μ according to whether the bootstrap critical value or the critical value correct for μ_0 is used. Since μ^* converges to μ_0 as $n \to \infty$, it follows that q tends to zero asymptotically. The rate at which $q \to 0$ depends on the extent of bootstrap refinements.

In the analysis that follows, we abbreviate $R(Q(\alpha, \mu_0), \mu)$ to just R. The marginal distribution of p under μ is, by construction, just $U(0,1)$. Let $F(q \,|\, p)$ denote the CDF of q conditional on p. The RP of the bootstrap test under μ is then

$$\begin{aligned} \Pr{}_\mu(p < R + q) &= E_\mu\big(\Pr{}_\mu(q > p - R|p)\big) \\ &= E_\mu\big(1 - F(p - R \,|\, p)\big) \\ &= 1 - \int_0^1 F(p - R \,|\, p)\, dp, \end{aligned} \tag{9}$$

where the last line follows because $p \sim U(0,1)$.

[2] The former term is probably more common in the econometrics literature, even if its use of the word "size" is incorrect in most contexts. Horowitz and Savin (2000) use the expression "Type I critical value" to refer to $Q(\alpha, \mu_0)$.

The values of $R(\cdot,\cdot)$ must belong to the interval $[0,1]$, and so the support of the random variable q is the interval $[-R, 1-R]$. This implies that, for any $p \in [0,1]$,

$$F(-R\,|\,p) = 0 \quad \text{and} \quad F(1-R\,|\,p) = 1. \tag{10}$$

On integrating by parts in (9) and changing variables, we find using (10) that the RP of the bootstrap test is

$$\begin{aligned} 1 - \Big[pF(p-R\,|\,p)\Big]_0^1 + \int_0^1 p\,dF(p-R\,|\,p) &= \int_{-R}^{1-R} (x+R)\,dF(x\,|\,R+x) \\ &= R + \int_{-\infty}^{\infty} x\,dF(x\,|\,R+x). \end{aligned} \tag{11}$$

We refer to the integral in (11) as the *bootstrap discrepancy*. When μ belongs to the null, the bootstrap discrepancy is the error in rejection probability (ERP) of the bootstrap test, and so it tends to zero as $n \to \infty$ at least as fast as q. When μ is a nonnull drifting DGP, the bootstrap discrepancy is the difference between the RP of the bootstrap test at nominal level α and that of the level-adjusted asymptotic test.

The following theorem shows that the bootstrap discrepancy tends to zero as $n \to \infty$ at the same rate under drifting DGPs as under the null.

Theorem 1

Let τ be a test statistic with asymptotic distribution $U(0,1)$ under all DGPs in a finite-dimensional null hypothesis model $\mathbb{M}_0$, with parameter space $\Theta \subseteq \mathbb{R}^k$. Let $\mathbb{M}_1$ be a $(k+1)$–dimensional model with parameter space $\Theta \times U$, where $U \subseteq \mathbb{R}$ contains the origin as an interior point, for which the set of DGPs characterized by the parameters $(\boldsymbol{\theta}, 0)$ are the DGPs of $\mathbb{M}_0$. Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta} \in \Theta$ that is root-n consistent under the null. Under regularity conditions specified in the Appendix, the bootstrap discrepancy, as defined above in (11), for a parametric bootstrap test based on $\hat{\boldsymbol{\theta}}$, has the same rate of convergence to zero as the sample size n tends to infinity for all levels α and for all drifting DGPs in $\mathbb{M}_1$ characterized by the sequence of parameters $(\boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{t}, n^{-1/2}\delta_0)$, for some $\boldsymbol{\theta}_0 \in \Theta$ and $\delta_0 \in U$, and some k–vector $\boldsymbol{t}$.

All proofs are found in the Appendix.

Remarks:

1. The bootstrap discrepancy, for a nonnull DGP, is the difference between the RP of the bootstrap test at *nominal* level α and the RP of the level-adjusted asymptotic test. A theoretical comparison of the two tests might better be based on the RP of the level-adjusted bootstrap test. Let $D(\alpha, \mu)$ denote the bootstrap discrepancy at level α for a DGP μ that drifts to $\mu_0 \in \mathbb{M}_0$. The nominal level α' at which the RP of the bootstrap test is exactly α under μ_0 satisfies the equation $\alpha' + D(\alpha', \mu_0) = \alpha$. Thus

$$\alpha' - \alpha = -D(\alpha', \mu_0), \tag{12}$$

and so $\alpha' - \alpha$ is of the same order as the bootstrap discrepancy. The bootstrap RP at nominal level α' is, by (11), $R(Q(\alpha', \mu_0), \mu) + D(\alpha', \mu)$. Thus the difference between this RP and the level-adjusted RP of the asymptotic test, which is $R(Q(\alpha, \mu_0), \mu)$, is

$$D(\alpha', \mu) + \Big(R\big(Q(\alpha', \mu_0), \mu\big) - R\big(Q(\alpha, \mu_0), \mu\big)\Big). \tag{13}$$

In the regularity conditions for Theorem 1, we assume that both R and Q are continuously differentiable with respect to their first argument. It therefore follows from (12) that the two terms in expression (13) are of the same order, namely, that of the bootstrap discrepancy.

2. The preceding remark implies that three quantities all tend to zero at the same rate as $n \to \infty$. They are (i) the ERP of the bootstrap test under the null, (ii) the difference under a nonnull drifting DGP between the power of the bootstrap test at nominal level α and the level-adjusted power of the asymptotic test, and (iii) the difference between the level-adjusted power of the bootstrap test and that of the asymptotic test. Just what the common rate of convergence is, expressed as a negative power of n, depends on the extent of bootstrap refinement.

A result similar to part of this result was obtained by Horowitz (1994), who showed that, if $R(\cdot, \mu)$ converges to the asymptotic distribution of τ at rate $n^{-j/2}$ for DGPs μ in the null, then the difference $R(\alpha, \mu^*) - R(\alpha, \mu_0)$ is of order $n^{-(j+1)/2}$ in probability. This is normally, but not always, also the order of the ERP of the bootstrap test.

3. Explicit expressions for the bootstrap discrepancy to leading order can often be obtained with the aid of Edgeworth expansions. See Hall (1988) and Hall (1992) for background. An explicit example is found in Abramovitch and Singh (1985), where the statistic is the t statistic for the mean of an IID sample. These authors express the bootstrap discrepancy under a nonnull drifting DGP indirectly in terms of the Hodges-Lehmann deficiency.

4. Although the bootstrap discrepancy is usually difficult to compute, it has an intuitive interpretation. Because the density of q is very small except in a short interval around 0, the second term in (11) can be approximated by $\int_{-\infty}^{\infty} x \, dF(x \mid R)$, that is, the expectation of q conditional on p being equal to R. By (6), $p = R$ is equivalent to $\tau = Q(\alpha, \mu_0)$, that is, to the condition that the statistic is at the margin between rejection and nonrejection using the critical value correct for μ_0. Taylor expansion of (7) around μ_0 then shows that the bootstrap discrepancy is approximately the bias of the bootstrap critical value, $Q(\alpha, \mu^*)$, thought of as an estimator of the critical value $Q(\alpha, \mu_0)$, conditional on being at the margin of rejection, scaled by the sensitivity of the RP to the critical value.

5. In general, neither the bootstrap discrepancy nor the difference (13) in level-adjusted powers can be signed. Thus, in any particular finite-sample case, either the asymptotic test or the bootstrap test could turn out to be more powerful.

6. If the statistic τ is exactly pivotal under the null hypothesis for any sample size, then $Q(\alpha, \mu)$ does not depend on μ if μ is in the null. Since the bootstrap DGP μ^* is by construction in the null, it follows that the random variable q of (7) is identically zero in this case, and so also, therefore, the three quantities of remark 2.

3. The Choice of a Drifting DGP

We are usually interested in practice in a particular sample size, say N, and we conduct asymptotic analysis as an approximation to what happens for a DGP μ^N defined just for that sample size. The drifting DGP used in asymptotic power analysis is a theoretical construct, but, as we will see in the simulations presented in Section 5, the bootstrap discrepancy can vary greatly with the specific choice of drifting DGP.

The parametrization $(\boldsymbol{\theta}, \delta)$ of the extended model $\mathbb{M}_1$ is not necessarily well adapted to the estimator $\hat{\boldsymbol{\theta}}$, since this estimator is not in general consistent for $\boldsymbol{\theta}$ except for DGPs in $\mathbb{M}_0$. We therefore introduce the following reparametrization. For each fixed DGP $\mu \in \mathbb{M}_1$, let $\boldsymbol{\phi} = \text{plim}_\mu \hat{\boldsymbol{\theta}}$. The model $\mathbb{M}_1$ is now to be parametrized by $\boldsymbol{\phi}$ and δ. By construction, $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\phi}$ over the full extended model $\mathbb{M}_1$. For the null model $\mathbb{M}_0$, $\delta = 0$, and the $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ parametrizations coincide.

Consider a drifting DGP μ constructed as follows. We start with a DGP $\mu^N \in \mathbb{M}_1$ defined for sample size N. We require first that the parameters of μ for sample size N should be those, say $(\boldsymbol{\phi}, \delta)$, that characterize μ^N in the new parametrization. Then, for any sample size n, the parameters are specified as

$$(\boldsymbol{\phi}, (n/N)^{-1/2}\delta). \tag{14}$$

An important property of such a drifting DGP is given in the following theorem.

Theorem 2

> Assume the regularity conditions of Theorem 1. Under a drifting DGP with parameters as specified in (14), the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\phi})$ is normal, and is the same as under the fixed DGP with parameters $(\boldsymbol{\phi}, 0)$. There is a positive integer j such that, for the random variable q of (7), $n^{(j+1)/2}q$ is asymptotically normal with asymptotic expectation of zero.

Davidson and MacKinnon (1999) show that, if, under a DGP $\mu_0 \in \mathbb{M}_0$ associated with the parameter vector $\boldsymbol{\theta}_0$, the estimator $\hat{\boldsymbol{\theta}}$ which determines the bootstrap DGP μ^* is such that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and τ are independent under their joint asymptotic distribution, then the ERP of the bootstrap test, which is the bootstrap discrepancy under the null, converges to zero at a rate faster by at least $n^{-1/2}$ than when this asymptotic independence does not hold. It is natural to enquire whether this more rapid convergence extends to drifting DGPs. The next theorem shows that it does if the drifting DGP is of the type given by (14).

Theorem 3

> Under the regularity conditions of Theorem 1, if $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and τ are independent under their joint asymptotic distribution for all DGPs in $\mathbb{M}_0$, then the rate of convergence to zero of the bootstrap discrepancy as the sample size $n \to \infty$ is faster than that of the random variable q defined in (7) by a factor of $n^{-1/2}$ or better for all DGPs in $\mathbb{M}_0$ and for all drifting DGPs in $\mathbb{M}_1$ of type (14).

A special case of particular interest arises if the estimator $\hat{\boldsymbol{\theta}}$ is the MLE for model $\mathbb{M}_0$. In this case, it is well known that, under any DGP in $\mathbb{M}_0$, $\hat{\boldsymbol{\theta}}$ is asymptotically independent of any classical test statistic that tests whether $\mathbb{M}_0$ is correctly specified.

It is possible in this case to give a meaningful characterization of the $(\boldsymbol{\phi}, \delta)$ parametrization of the extended model $\mathbb{M}_1$. Consider a DGP $\mu_1 \in \mathbb{M}_1$ associated with parameters $(\boldsymbol{\theta}, \delta)$. Let $(\boldsymbol{\phi}, \delta)$ be the corresponding parameters in the reparametrization induced by the MLE for $\mathbb{M}_0$. Thus $\boldsymbol{\phi}$ is the probability limit under μ_1 of the quasi-maximum likelihood estimator (QMLE) for $\mathbb{M}_0$. The Kullback-Leibler information criterion (KLIC) is an asymmetric measure of the distance from one DGP, defined for a given sample size, to another. Let the first DGP be μ_1 for some sample size n, and consider the problem of minimizing the KLIC from μ_1 to a DGP in $\mathbb{M}_0$ for the same n. By definition of the KLIC, the parameters of the minimizing DGP maximize the expectation under μ_1 of the loglikelihood function of the null model for sample size n. Let these parameters be $\boldsymbol{\phi}_n$. The reason for this notation is that White (1982) showed that $\operatorname{plim} \boldsymbol{\phi}_n = \boldsymbol{\phi}$.

The parameters $\boldsymbol{\phi}$ are usually called the pseudo-true parameters for μ_1. We refer to the fixed DGP $\mu_0 \in \mathbb{M}_0$ with parameters $\boldsymbol{\phi}$ as the *pseudo-true* DGP. Theorem 3 tells us that the more rapid convergence to zero of the bootstrap discrepancy for classical test statistics, resulting from their asymptotic independence of the MLE for the null model, extends to DGPs that drift from a given nonnull DGP for sample size N to the corresponding pseudo-true DGP according to the drift scheme (14).

A slight modification of the scheme (14) leads to a drifting DGP starting from μ_1 for sample size N with the property that the endpoint μ_0 is the pseudo-true DGP not only for sample size N but for all sample sizes. Quite generally, let the pseudo-true parameters associated with the DGP with parameters $(\boldsymbol{\theta}, \delta)$ be $(\Pi_n(\boldsymbol{\theta}, \delta), 0)$ for sample size n. Clearly, $\Pi_n(\boldsymbol{\theta}, 0) = \boldsymbol{\theta}$ for all n. Let $\Phi_n(\boldsymbol{\theta}, \delta)$ be the inverse of Π_n for given δ, so that $\Pi_n(\Phi_n(\boldsymbol{\theta}, \delta), \delta) = \boldsymbol{\theta}$ and $\Phi_n(\Pi_n(\boldsymbol{\theta}, \delta), \delta) = \boldsymbol{\theta}$. Then, for sample size n, the drifting DGP has parameters

$$(\Phi_n(\boldsymbol{\phi}, (n/N)^{-1/2}\delta), (n/N)^{-1/2}\delta), \tag{15}$$

where $(\boldsymbol{\phi}, \delta)$ are the parameters in the $\boldsymbol{\phi}$ parametrization for μ_1 at the reference sample size N. It is clear that, as $n \to \infty$, (15) drifts towards $(\boldsymbol{\phi}, 0)$.

Although the drifting DGP (15) does not follow scheme (14), the following corollary shows that the result of Theorem 3 continue to hold for (15). Moreover, the bootstrap discrepancy is the same to leading order for (15) and the DGP that drifts from μ_1 for sample size N to μ_0 according to (14). Drifting DGPs for which the bootstrap discrepancy is the same to leading order will be called *asymptotically equivalent*.

Corollary

> The bootstrap discrepancy is the same to leading order for the drifting DGP with $\boldsymbol{\phi}$ parameters $(\boldsymbol{\phi}, n^{-1/2}\delta)$ for sample size n and for drifting DGPs for which the parameters are $(\boldsymbol{\phi} + n^{-1/2}\boldsymbol{p}_n, n^{-1/2}\delta)$, if $\boldsymbol{p}_n$ tends to zero as $n \to \infty$.

The question of what null DGP μ_0 is most appropriate for the level adjustment of either a bootstrap or an asymptotic test does not seem to have an unambiguous answer in general. If, for a given nominal level α, there exists a $\mu_0 \in \mathbb{M}_0$ which maximizes the RP of the test based on τ, then there are good arguments for basing level adjustment on this μ_0, in which case one can legitimately speak of "size adjustment." However, as pointed out by Horowitz and Savin (2000), such a μ_0 may not exist, or, if it does, it may be intractable to compute its parameters, or it may lead to a size-adjusted power

no greater than the size. In addition, μ_0 will in general depend on α, μ_1, and n, thereby making such size adjustment essentially impossible in practice.

Study of the power of asymptotic tests is usually based on Monte Carlo experiments. As Horowitz and Savin (2000) point out, it is common for such studies to perform some sort of level adjustment, but most do so on the basis of an essentially arbitrary choice of the null DGP μ_0 used to generate critical values. Horowitz and Savin are critical of level adjustment in Monte Carlo experiments based on anything other than the $\mu_0 \in \mathbb{M}_0$ with parameters given by the plim of $\hat{\boldsymbol{\theta}}$ under the DGP μ_1 for which power is to be studied. The thrust of their argument is that, since only the bootstrap offers any hope of performing level adjustment with any reasonable accuracy in practice, level adjustment in Monte Carlo experiments, to be meaningful, should in the large-sample limit coincide with the bootstrap level adjustment. This is the case for a parametric bootstrap based on $\hat{\boldsymbol{\theta}}$ if μ_1 is thought of as a *fixed* DGP, since then the parameters of the bootstrap DGP converge as $n \to \infty$ to those of μ_0.

It is illuminating to examine this argument in the light of the results of this paper. Asymptotic analysis of power is not feasible with fixed nonnull DGPs, which is why we have considered drifting DGPs. But if all that is required of these is that they start at μ_1 for a given sample size, and drift to *some* DGP in the null, then the bootstrap DGP will also drift to that null DGP, which might therefore seem to be indicated for level adjustment. Such a conclusion would clearly be unsatisfactory.

The bootstrap is usually the best way to do level adjustment in practice. Therefore, if Monte Carlo experiments on level-adjusted power are to be informative, we should try to do level adjustment in experiments using a null DGP μ_0 that in some sense minimizes the bootstrap discrepancy for DGPs that drift to it from μ_1. In this way, one would minimize the difference between the rejection probability $R(Q(\alpha, \mu_0), \mu_1)$, which can be estimated with arbitrary accuracy by simulation for any given μ_1 and μ_0, and the RP of the bootstrap test. Finite-sample simulation results would then be as close as possible to the actual behavior of the bootstrap.

It is through the random variable q of (7) that the bootstrap discrepancy depends on μ_0. To leading order, the discrepancy is the expectation of q conditional on the statistic τ being at the margin of rejection. Although Theorem 2 shows that $n^{(j+1)/2}q$ has asymptotic expectation of 0 under drifting DGPs of type (14), the conditional expectation is different from 0 unless $n^{(j+1)/2}q$ and τ are asymptotically uncorrelated, and is not in general smaller for DGPs of type (14) than for other drifting DGPs.

If $n^{(j+1)/2}q$ and τ are asymptotically uncorrelated, then, by Theorem 3, the bootstrap discrepancy is an order of magnitude smaller under (14) than under other drifting DGPs. However, the DGP μ_0 used for level adjustment is still dependent on the specific estimator $\hat{\boldsymbol{\theta}}$ used to define the bootstrap DGP. If $\hat{\boldsymbol{\theta}}$ is not asymptotically equivalent to the MLE for $\mathbb{M}_0$, then μ_0 is not the null DGP that minimizes the KLIC from μ_1. It is still possible that τ is asymptotically independent of an asymptotically inefficient $\hat{\boldsymbol{\theta}}$, in which case μ_0 minimizes the bootstrap discrepancy, and so should certainly be used in Monte Carlo experiments. There is, however, no unique choice of μ_0 that minimizes the discrepancy for all root-n consistent estimators $\hat{\boldsymbol{\theta}}$. The story is clean only when $\hat{\boldsymbol{\theta}}$ is the MLE for $\mathbb{M}_0$, or is asymptotically equivalent to it. Then μ_0 is uniquely defined

in a way that is independent of the parametrization of $\mathbb{M}_0$, since the KLIC and the KLIC-minimizing DGP are parametrization independent.

Many applications of the bootstrap use, not the parametric bootstrap DGP we have considered, but a DGP that is at least partially nonparametric, based on some sort of resampling. Although we conjecture that much of the analysis of this section applies as well to the nonparametric bootstrap, there are technical difficulties in the way of proving this. For Theorem 1, these arise in connection with the LAN property (see the Appendix) when DGPs in the null hypothesis are not uniquely characterized by a finite-dimensional parameter, and for Theorem 2, there are analogous difficulties with the LAE property. Beran (1997) makes use of an ingenious construction to sidestep the problem, and it seems likely that a similar technique would work here. For Theorem 3, the main difficulty is that asymptotic independence of τ and μ^* cannot, in general, be achieved simply by using a classical test statistic and an asymptotically efficient estimator under the null.

4. Approximate Bootstrap Rejection Probabilities

The quantity $R(Q(\alpha, \mu_0), \mu)$, which is the power of an asymptotic test based on τ against the DGP μ at level α when level adjustment is based on the null DGP μ_0, can be straightforwardly estimated by simulation. For each of M replications, compute two test statistics, one of them generated by μ and the other by μ_0. Find the critical value τ_c such that the rejection frequency in the M replications under μ_0 is α; τ_c is our estimate of $Q(\alpha, \mu_0)$. $R(Q(\alpha, \mu_0), \mu)$ is then estimated by the rejection frequency under μ with critical value τ_c. If desired, we can study how power depends on level using a "size-power curve," as suggested by Davidson and MacKinnon (1998).

Exactly this sort of simulation experiment was suggested by Horowitz (1994) to estimate expression (5), the power of a bootstrap test. However, this ignores the bootstrap discrepancy. The obvious, but computationally expensive, way to estimate (5) taking account of the bootstrap discrepancy is to generate M sets of data, indexed by m, from μ, and for each to compute a test statistic τ_m and a bootstrap DGP μ^*_m. For each m, generate B statistics from μ^*_m and find the critical value $\hat{Q}(\alpha, \mu^*_m)$ such that a fraction α of the bootstrap statistics are more extreme than it. Then the estimate of (5) is

$$\frac{1}{M} \sum_{m=1}^{M} I\big(\tau_m \in \text{Rej}(\hat{Q}(\alpha, \mu^*_m))\big), \tag{16}$$

the fraction of the M replications for which the bootstrap statistic leads to rejection. Horowitz also performed some simulations of this sort.

The procedure just described, which, of course, does not require the statistic τ to have an approximate $U(0, 1)$ distribution, involves the calculation of $M(B + 1)$ test statistics. Since the power of bootstrap tests is increasing in B (see Davidson and MacKinnon, 2000), we probably do not want to use B less than about 399, in which case this procedure is roughly 200 times as expensive as the one described above for an asymptotic test.

We now propose a method for estimating the power of bootstrap tests that takes (approximate) account of the bootstrap discrepancy at computational cost similar to that required for the level-adjusted power of an asymptotic test. The conditions of Theorem 3 are assumed, namely, that the parameters of the null DGP μ_0 are given by the plim of $\hat{\boldsymbol{\theta}}$ under μ, and that τ and $n^{(j+1)/2}q$ are asymptotically independent. The method has the further very considerable advantage that it does not require calculation of the parameters of μ_0. It is a slight modification of a method proposed in Davidson and MacKinnon (2001) for approximating the RP of a bootstrap test under the null.

From (11) it can be seen that the RP of a bootstrap test under μ is $R(Q(\alpha, \mu_0), \mu)$ plus the bootstrap discrepancy, which to leading order is just the expectation of q under the asymptotic independence assumption. Thus, using (7), the definition of q, the RP of the bootstrap is $\mathrm{E}_\mu\big(R(Q(\alpha, \mu^*), \mu)\big)$ to leading order. Davidson and MacKinnon (2001) proposed estimating this quantity as follows. For each replication, compute τ_m and μ_m^* as before, but now generate just one bootstrap sample and use it to compute a single test statistic, τ_m^*. Then calculate $\hat{Q}^*(\alpha)$, the α quantile of the τ_m^*. The approximate rejection probability is then

$$\widehat{\mathrm{RP}}_{\mathrm{A}} \equiv \frac{1}{M} \sum_{m=1}^{M} I\big(\tau_m \in \mathrm{Rej}(\hat{Q}^*(\alpha))\big). \tag{17}$$

The only difference between (16) and (17) is that the τ_m are compared to different estimated critical values inside the indicator functions. If the τ_m are independent of the μ_m^*, it should not make any difference whether we estimate the α quantile $Q(\alpha, \mu_m^*)$ separately for each m, or use the α quantile $Q^*(\alpha)$ of all the τ_m^* taken together. How well (17) approximates (16) in finite samples depends on how close τ is to being independent of μ^*. We do not claim that $\widehat{\mathrm{RP}}_{\mathrm{A}}$ works well in all circumstances. However, numerous simulation experiments, some of which are discussed in the next section, suggest that it often works very well in practice.

The amount of computation required to compute $\widehat{\mathrm{RP}}_{\mathrm{A}}$ is very similar to that required for an asymptotic test: Once again, we have to calculate $2M$ test statistics. But the $\widehat{\mathrm{RP}}_{\mathrm{A}}$ procedure is often a good deal simpler, because there is no need to calculate the parameters of μ_0 explicitly. When that is difficult, we can use Horowitz's (1994) argument in reverse, and claim that, with error no greater than the order of the bootstrap discrepancy, $\widehat{\mathrm{RP}}_{\mathrm{A}}$ estimates the level-adjusted power of the asymptotic test.

5. Testing for Omitted Variables in a Logit Model

In this section, we present the results of several Monte Carlo experiments, which deal with Lagrange multiplier tests for omitted variables in the logit model. We chose to examine the logit model for several reasons: It is not a regression model, the results of Horowitz (1994) and Davidson and MacKinnon (1998) 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, and it is easy to calculate the pseudo-true DGP for this model.

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

$$\mathrm{E}(y_t \,|\, \boldsymbol{X}_t, \boldsymbol{Z}_t) = F(\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Z}_t\boldsymbol{\gamma}) \equiv \big(1 + \exp(-\boldsymbol{X}_t\boldsymbol{\beta} - \boldsymbol{Z}_t\boldsymbol{\gamma})\big)^{-1}, \tag{18}$$

where y_t is an observation on a 0-1 dependent variable, $\boldsymbol{X}_t$ 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 corresponding vectors of unknown parameters. The null hypothesis that $\boldsymbol{\gamma} = \mathbf{0}$ may be tested in several ways. Two of the easiest are to use tests based on artificial regressions. The first of these 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 probit 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 (18) 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

$$1 = \frac{\tilde{f}_t(y_t - \tilde{F}_t)}{\tilde{F}_t(1 - \tilde{F}_t)} \left(\sum_{i=1}^{k} X_{ti} b_i + \sum_{i=1}^{r} Z_{ti} g_i \right) + \text{residual}, \tag{19}$$

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

$$\frac{y_t - \tilde{F}_t}{\big(\tilde{F}_t(1 - \tilde{F}_t)\big)^{1/2}} = \frac{\tilde{f}_t}{\big(\tilde{F}_t(1 - \tilde{F}_t)\big)^{1/2}} \left(\sum_{i=1}^{k} X_{ti} b_i + \sum_{i=1}^{r} Z_{ti} g_i \right) + \text{residual}. \tag{20}$$

In regressions (19) and (20), 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 order to level-adjust the tests, it is necessary to compute the parameters of the pseudo-true DGP that corresponds to whatever nonnull DGP actually generated the data. If $\log h(\boldsymbol{y}, \boldsymbol{\beta})$ denotes the loglikelihood evaluated at the parameters of a DGP in the null, and $\log g(\boldsymbol{y}, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_1)$ the loglikelihood at the parameters of the nonnull DGP, then the KLIC is

$$\mathrm{E}\big(\log g(\boldsymbol{y}, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_1) - \log h(\boldsymbol{y}, \boldsymbol{\beta})\big), \tag{21}$$

where the expectation is computed under the nonnull DGP. The parameter vector for the pseudo-true DGP is the $\boldsymbol{\beta}_0$ that minimizes (21) with respect to $\boldsymbol{\beta}$. For the model (18), this is just the vector that maximizes the expectation of $\log h(\boldsymbol{y}, \boldsymbol{\beta})$, namely,

$$\sum_{t=1}^{n} \Big(F(\boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{Z}_t\boldsymbol{\gamma}_1) \log F(\boldsymbol{X}_t\boldsymbol{\beta}) + \big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{Z}_t\boldsymbol{\gamma}_1)\big) \log\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big) \Big). \tag{22}$$

Here we have used the fact that $\mathrm{E}(y_t \mid \boldsymbol{X}_t, \boldsymbol{Z}_t) = F(\boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{Z}_t\boldsymbol{\gamma}_1)$.

The first-order conditions for maximizing expression (22) are

$$\sum_{t=1}^{n} \frac{f(\boldsymbol{X}_t\boldsymbol{\beta}_0)\boldsymbol{X}_t}{F(\boldsymbol{X}_t\boldsymbol{\beta}_0)\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta}_0)\big)}\big(F(\boldsymbol{X}_t\boldsymbol{\beta}_0) - F(\boldsymbol{X}_t\boldsymbol{\beta}_1 + \boldsymbol{Z}_t\boldsymbol{\gamma}_1)\big) = \mathbf{0}. \tag{23}$$

These equations give us the relationship between the nonnull DGP, which is characterized by $\boldsymbol{\beta} = \boldsymbol{\beta}_1$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1$, the pseudo-true DGP, which is characterized by $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma} = \mathbf{0}$, and what we may call the *naive null* DGP, which is characterized by $\boldsymbol{\beta} = \boldsymbol{\beta}_1$ and $\boldsymbol{\gamma} = \mathbf{0}$. Given $\boldsymbol{\gamma}_1$ and either $\boldsymbol{\beta}_0$ from the pseudo-true null or $\boldsymbol{\beta}_1$ from the naive null, we can solve equations (23) for the other $\boldsymbol{\beta}$ vector.

In our experiments, the vector $\boldsymbol{X}_t$ consisted of a constant term and two regressors that were distributed as $N(0, 1)$, and the vector $\boldsymbol{Z}_t$ consisted of eight regressors that were also $N(0, 1)$. The two non-constant regressors in $\boldsymbol{X}_t$ were uncorrelated, and the regressors in $\boldsymbol{Z}_t$ were correlated with one or the other of these, with correlation coefficient ρ. The number of regressors in $\boldsymbol{Z}_t$ was chosen to be quite large in order to make the tests, especially the OPG test, work relatively poorly. 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 DGP depends on the regressors, we used a single set of regressors in all the experiments, and all our experimental results are conditional on them.

In the experiments we report, we held the pseudo-true DGP constant and used equation (23) to vary $\boldsymbol{\beta}_1$ as δ, and hence $\boldsymbol{\gamma}_1$, varied. This procedure makes it relatively inexpensive to level-adjust the bootstrap test, since there is only one pseudo-true DGP to be concerned with. In a second set of experiments, of more conventional design, we held the naive null DGP constant and varied $\boldsymbol{\beta}_0$ as δ varied. Because both sets of experiments yielded similar results, we do not report those from the second set to conserve space.

In both sets of experiments, we set $\rho = -0.8$. By making ρ fairly large, we ensure that the pseudo-true and naive null DGPs will be quite different, except when the nonnull DGP is very close to the null hypothesis. Changing the sign of ρ would change the results in a predictable way: The figures would be roughly mirror images of the ones presented here. In our experiments, the constant term in either $\boldsymbol{\beta}_0$ or $\boldsymbol{\beta}_1$ is set to 0, and the two slope coefficients are set to 1. Thus, under the null hypothesis, approximately half of the y_t would be 0 and half would be 1. The sample size was 125. This relatively large sample size was used in order to avoid having more than a few replications for which the logit routine failed to converge. Nonconvergence, which is caused by perfect classifiers, is more of a problem for smaller sample sizes and for larger values of δ.

Figures 1 and 2 show estimated power functions for asymptotic and bootstrap tests at the .05 level, the former for the ES tests, and the latter for the OPG tests. These power functions are based on 200,000 replications for a large number of values of δ between -0.8 and 0.8 at intervals of 0.025. The unadjusted power function (the solid line) shows the power of the asymptotic test at the nominal .05 level. The two adjusted power functions (the dotted lines) show adjusted test power, calculated in two different ways.

For all values of δ, the naive adjustment method uses test performance for $(\boldsymbol{\beta}_1, \mathbf{0})$ as a benchmark. In contrast, the pseudo-true adjustment method uses test performance for $(\boldsymbol{\beta}_0, \mathbf{0})$ as a benchmark. The power function estimated by the $\widehat{\text{RP}}_\text{A}$ procedure is shown as a dashed line.

Figures 1 and 2 also show the results of nine Monte Carlo experiments for bootstrap tests. Each experiment involved 50,000 replications. 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 DGP was simply the logit model with parameters $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma} = \mathbf{0}$, and the bootstrap test rejected the null hypothesis whenever 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. Figure 2 also shows the level-adjusted power for the bootstrap test, based on the pseudo-true DGP.

From Figure 1, we see that the ES test works so well as an asymptotic test that there is no need to bootstrap it. There is essentially no difference between any of the power functions, which suggests that the ES test statistic is nearly pivotal in this case.

In contrast, from Figure 2, we see that the OPG test statistic is far from pivotal. As the theory predicts, the $\widehat{\text{RP}}_\text{A}$ estimated power function is very similar to the one adjusted using the pseudo-true null. However, both of these are quite different from the power function adjusted using the naive null. This confirms that the issues raised in Section 3 are empirically relevant: The null DGP used for level adjustment can have a substantial effect, and it is the pseudo-true null that yields a power function close to that of the bootstrap test. Use of the naive null in fact leads to the misleading appearance of greater power for the OPG test for some negative values of δ. It is misleading because the ES and OPG statistics are based on the same empirical scores, but the latter uses a noisier estimate of the information matrix, which should reduce its power.

The theory is also confirmed by the fact that, when the power of the bootstrap test is level-adjusted (because the test underrejects slightly), the correspondence of the power function with the $\widehat{\text{RP}}_\text{A}$ function is not as good as with no adjustment. This is as expected, since $\widehat{\text{RP}}_\text{A}$ estimates the *nominal* power of the bootstrap test.

6. Conclusions

Level adjustment of the power of tests based on nonpivotal statistics yields results that depend on the DGP in the null hypothesis used to provide a critical value. For a given choice of this null DGP, we show that the power of a bootstrap test differs from the level-adjusted power of the asymptotic test on which it is based by an amount that we call the bootstrap discrepancy. This discrepancy is of the same order, in the sample size n, as the size distortion of the bootstrap test itself.

Since the bootstrap constitutes the best way to do level adjustment in practice, it makes sense to use critical values from a null DGP that minimizes the bootstrap discrepancy to do level adjustment in simulation experiments. In this way, power as measured by simulation in finite samples is a good approximation to the power of a bootstrap test. The rate of convergence to zero of the bootstrap discrepancy when the sample size

tends to infinity is analyzed in connection with different drifting DGPs, and we show that convergence is fastest when the test statistic is asymptotically independent of the bootstrap DGP and when a particular sort of drift towards a particular null DGP is used. This result serves to extend to the analysis of power a previous result whereby the ERP of a bootstrap test is of lower order with asymptotic independence.

Level-adjusted power can be estimated efficiently by simulation if the appropriate null DGP for providing critical values can readily be characterized. We propose a new approximate method that requires no such calculation and yields better estimates of the power of bootstrap tests by taking account of the bootstrap discrepancy.

Our theoretical results are confirmed and illustrated, for the case of tests for omitted variables in logit models, by simulation results which show that level adjustment of our preferred type leads to power estimates close to the power of bootstrap tests, while a cruder form of level adjustment may give quite different results.

Appendix

We state the regularity conditions that we make for the proof of Theorem 1. The first is an assumption about the $(k+1)$–dimensional model $\mathbb{M}_1$ that contains the null hypothesis model $\mathbb{M}_0$.

ASSUMPTION 1: The model $\mathbb{M}_1$, parametrized by $\Theta \times U$, is locally asymptotically normal (LAN) at all fixed DGPs $\mu \in \mathbb{M}_1$.

Local asymptotic normality was introduced by Le Cam (1960). See also Beran (1997) for a more modern version of the definition. What is required is that, for any sample size n and for all $\boldsymbol{\eta} = (\boldsymbol{\theta}, \delta) \in \Theta \times U$, the difference $\ell_n(\boldsymbol{a}, \boldsymbol{\eta})$ between the log of the joint density of the dependent variables under the DGP corresponding to parameters $\boldsymbol{\eta} + n^{-1/2}\boldsymbol{a}_n$, where the sequence $\{\boldsymbol{a}_n\}$ converges to $\boldsymbol{a}$, and that under the DGP $\boldsymbol{\eta}$ (the loglikelihood ratio) should take the form

$$\ell_n(\boldsymbol{a}, \boldsymbol{\eta}) = \boldsymbol{a}^\top \boldsymbol{g}_n(\boldsymbol{\eta}) - 2\boldsymbol{a}^\top \boldsymbol{I}(\boldsymbol{\eta})\boldsymbol{a} + o_p(1)$$

for all $\boldsymbol{a} \in \mathbb{R}^{k+1}$, where the $(k+1) \times (k+1)$ matrix $\boldsymbol{I}(\boldsymbol{\eta})$ is the information matrix at $\boldsymbol{\eta}$, and the $(k+1)$–dimensional random vectors $\boldsymbol{g}_n(\boldsymbol{\eta})$ are such that

$$\boldsymbol{g}_n(\boldsymbol{\eta} + n^{-1/2}\boldsymbol{a}_n) = \boldsymbol{g}_n(\boldsymbol{\eta}) - \boldsymbol{I}(\boldsymbol{\eta})\boldsymbol{a} + o_p(1).$$

In addition, the expectation of $\boldsymbol{g}_n(\boldsymbol{\eta})$ is zero for the DGP $\boldsymbol{\eta}$, and, as $n \to \infty$, it tends in distribution to $N(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\eta}))$. As the name suggests, LAN models have the regularity needed for the usual properties of the MLE, including asymptotic normality.

ASSUMPTION 2: The estimator $\hat{\boldsymbol{\theta}}$ is locally asymptotically equivariant (LAE) at all fixed DGPs $\mu \in \mathbb{M}_0$.

The definition of the LAE property is taken from Beran (1997). For a DGP $\mu_0 \in \mathbb{M}_0$ with parameter vector $\boldsymbol{\theta}_0$, consider a drifting DGP μ with parameters in the $(\boldsymbol{\phi}, \delta)$ reparametrization, which was introduced in Section 3, given by the sequence

$\{(\boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{t}_n, n^{-1/2}d_n)\}$, where $\boldsymbol{t}_n$ converges to a fixed k–vector $\boldsymbol{t}$, and d_n converges to $d \in \mathbb{R}$. The $(\boldsymbol{\phi}, \delta)$ parametrization is used because $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\phi}$ for the extended model $\mathbb{M}_1$. The LAE property requires that the random vectors $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - n^{-1/2}\boldsymbol{t}_n)$ converge in distribution under this drifting DGP to the asymptotic distribution of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ under μ_0, namely, $N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0, 0))$.

The LAE property is a condition which guarantees the usual desirable properties of the parametric bootstrap distribution of the estimator $\hat{\boldsymbol{\theta}}$ and excludes the possibility of bootstrap failure, as explained by Beran. It is likely that a weaker condition would suffice for our needs, where $\hat{\boldsymbol{\theta}}$ itself is not bootstrapped but simply serves to define the bootstrap distribution.

In the next assumption, we extend the LAN property to cover the alternative hypothesis against which the test statistic τ has maximal power.

ASSUMPTION 3: The test statistic t of which τ is the asymptotic P value is a statistic in either standard normal or χ^2 form asymptotically equivalent to a classical test statistic (LR, LM, or Wald) of the hypothesis represented by the model $\mathbb{M}_0$ against an alternative hypothesis represented by a $(k+r)$–dimensional LAN model $\mathbb{M}_2$ that includes the model $\mathbb{M}_0$ as a subset. Here r is the number of degrees of freedom of the test. Under any DGP $\mu \in \mathbb{M}_1$, the CDF $R(\alpha, \mu)$ of τ is a continuously differentiable function of α.

This assumption allows us to make use of results in Davidson and MacKinnon (1987), where it is shown that, under weak regularity conditions ensured by the LAN property, test statistics in standard normal or χ^2 form can always be associated with a model like $\mathbb{M}_2$, for which they are asymptotically equivalent to a classical test of $\mathbb{M}_0$ against $\mathbb{M}_2$.

Our final assumption is needed in order to be able to speak concretely about rates of convergence.

ASSUMPTION 4: For sample size n, the critical value function $Q(\alpha, \mu)$, defined in equation (2), can be expressed for all DGPs $\mu \in \mathbb{M}_0$ as

$$Q(\alpha, \mu) = \alpha + n^{-j/2}\gamma(\alpha, \mu), \tag{24}$$

where j is a positive integer, and the function γ is $O(1)$ as $n \to \infty$ and continuously differentiable with respect to the parameters of the DGP μ.

Since we assume that the statistic τ is expressed in asymptotic P value form, its asymptotic distribution is $U(0,1)$ for all DGPs in $\mathbb{M}_0$. It follows that, for $\mu \in \mathbb{M}_0$, $Q(\alpha, \mu) = \alpha + o(1)$. The relation (24) specifies the actual rate of convergence to zero of the remainder term.

Assumption 4 is precisely the assumption made in Beran (1988) in the analysis of the RP of bootstrap tests under the null.[3] It would have been possible to devise some more primitive conditions that, along with the other assumptions, would imply Assumption 4, but the clarity of the latter seems preferable.

[3] Beran makes the assumption about the function we denote as $R(\alpha, \mu)$, but, since R and Q are inverse functions, the assumption can equivalently be made about one or the other.

Proof of Theorem 1: By Assumption 3, the test statistic t of which τ is the asymptotic P value has a noncentral χ^2 asymptotic distribution under DGPs in $\mathbb{M}_1$ that drift to $\mathbb{M}_0$; this is the conclusion of the Theorem on page 1317 of Davidson and MacKinnon (1987). This distribution is completely characterized by the number r of degrees of freedom of the test and a scalar noncentrality parameter (NCP) λ that depends on the drifting DGP. Thus, for such a DGP μ, $R(\alpha, \mu)$ tends as $n \to \infty$ to $P(\alpha, \lambda)$, the probability mass in the tail of the $\chi^2_r(\lambda)$ distribution beyond the critical value for a test at level α as defined by the central χ^2_r distribution.

If μ is a fixed DGP in $\mathbb{M}_0$, then, by Assumption 4 (see also the footnote to it),

$$R(\alpha, \mu) = \alpha + n^{-j/2}\rho(\alpha, \mu),$$

for some $O(1)$ function $\rho(\alpha, \mu)$ that is continuously differentiable with respect to α by Assumption 3. Thus the sequence $\{\tau^n\}$ of test statistics for finite sample sizes n converges to a random variable with distribution $U(0, 1)$ under DGPs $\mu \in \mathbb{M}_0$.

The model $\mathbb{M}_1$ that contains the drifting DGPs of interest to us is LAN, by Assumption 1. A consequence of this is that the probability measures defined on $[0, 1]$ by the sequence $\{\tau^n\}$ under a DGP $\mu \in \mathbb{M}_0$ are *contiguous* to those defined by $\{\tau^n\}$ under a DGP that drifts to μ_0; see Roussas (1972) for a discussion of contiguity. Consequently, the sequence $\{\tau^n\}$ converges to the same limiting random variable under μ_0 and DGPs that drift to μ_0. By the argument in the first paragraph of the proof, under drifting DGPs with NCP λ, this variable has CDF $P(\alpha, \lambda)$.

By a slight abuse of notation, we write $Q(\alpha, \boldsymbol{\theta})$ for $Q(\alpha, \mu)$ when $\mu \in \mathbb{M}_0$ and $\boldsymbol{\theta}$ is the parameter vector associated with μ, and similarly for $\gamma(\alpha, \boldsymbol{\theta})$. Recall that the old ($\boldsymbol{\theta}$) and new ($\boldsymbol{\phi}$) parametrizations coincide on $\mathbb{M}_0$. Since the bootstrap DGP μ^* is in $\mathbb{M}_0$ and is characterized by the parameter vector $\hat{\boldsymbol{\theta}}$, we have that $Q(\alpha, \mu^*) = \alpha + n^{-j/2}\gamma(\alpha, \hat{\boldsymbol{\theta}})$. Then, from the definition (7), since the fixed DGP μ_0 is also in $\mathbb{M}_0$, the random variable q can be expressed as

$$q = R(\alpha + n^{-j/2}\gamma(\alpha, \hat{\boldsymbol{\theta}}), \mu) - R(\alpha + n^{-j/2}\gamma(\alpha, \boldsymbol{\theta}_0), \mu). \tag{25}$$

Since the function $R(\cdot, \mu)$ is continuously differentiable with respect to α, we may perform a Taylor expansion of (25) to obtain

$$q = n^{-j/2}\Big(P'(\alpha, \lambda)\big(\gamma(\alpha, \hat{\boldsymbol{\theta}}) - \gamma(\alpha, \boldsymbol{\theta}_0)\big) + o_p(1)\Big),$$

where λ is the NCP for μ, and $P'(\alpha, \lambda)$ is the derivative of $P(\alpha, \lambda)$ with respect to α.

Since $\gamma(\alpha, \boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta}$ by Assumption 4, Taylor's Theorem gives

$$\gamma(\alpha, \hat{\boldsymbol{\theta}}) - \gamma(\alpha, \boldsymbol{\theta}_0) = D_{\boldsymbol{\theta}}\gamma(\alpha, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O_p(n^{-1/2}).$$

It follows that $n^{(j+1)/2}q$ is a linear combination of the components of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ plus a variable that tends to zero in μ_0–probability, and, by contiguity, also in μ–probability.

By Assumption 2, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ has an asymptotically normal distribution, with finite variance and with mean zero under μ_0 and finite mean under μ.

The statistic t, if in χ^2 form, is a quadratic form in r asymptotically normal variables, with finite mean and variance, that have an asymptotically normal distribution jointly with $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$; again see Davidson and MacKinnon (1987) for details. If $r = 1$, t is itself an asymptotically normal variable. Thus, to leading order asymptotically under the drifting DGP μ, the joint distribution of the r variables used to construct t and $n^{(j+1)/2}q$ is multivariate normal. It follows that the distribution of $n^{(j+1)/2}q$ conditional on t, and so also on τ and on p, which are deterministic functions of t, is asymptotically normal with finite mean and variance.

Let the CDF of $n^{(j+1)/2}q$ conditional on p be denoted as $G(z \,|\, p)$. As $n \to \infty$, this tends to a normal CDF with finite mean and variance under the drifting DGP μ. By performing the change of variable $x = n^{-(j+1)/2}z$ in the expression for the bootstrap discrepancy given by (11), it can be seen that the discrepancy is

$$n^{-(j+1)/2} \int_{-\infty}^{\infty} z \, dG(z \,|\, R + n^{-(j+1)/2}z),$$

which is of order $n^{-(j+1)/2}$ under both the drifting DGP μ and the fixed null DGP μ_0. This completes the proof. ■

Proof of Theorem 2: Since the $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ parametrizations coincide on $\mathbb{M}_0$, the $\boldsymbol{\phi}$ in (14) can be identified with the parameters $\boldsymbol{\theta}_0$ of the null DGP to which (14) drifts. Asymptotic normality of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ was already shown in the proof of Theorem 1. That its asymptotic distribution is the same under the drifting DGP (14) as under the fixed DGP to which it drifts is then an immediate consequence of Assumption 2.

It was shown in the proof of Theorem 1 that, to leading order, $n^{(j+1)/2}q$ is a linear combination of the components of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ It follows that $n^{(j+1)/2}q$ is asymptotically normal with expectation zero under (14). ■

Proof of Theorem 3: The theorem supposes that, for any $\mu_0 \in \mathbb{M}_0$ with parameters $\boldsymbol{\theta}_0$, the statistic τ and $n^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ are independent under their joint asymptotic distribution. This independence holds also under DGPs that drift to $\mathbb{M}_0$, since, by contiguity, the joint asymptotic distribution of $n^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and the r asymptotically normal variables on which τ depends differs under drifting DGPs from what it is under DGPs in $\mathbb{M}_0$ only in its expectation, not its covariance matrix.

If the conditional CDF $F(q \,|\, p)$ is independent of p to leading order, then, from (11), the bootstrap discrepancy is to that order just the asymptotic expectation of q. The conclusion of this theorem now follows immediately from Theorem 2. ■

Proof of Corollary: The bootstrap discrepancy is determined by the joint distribution of τ and q, and to leading order by the joint asymptotic distribution of τ and $n^{(j+1)/2}q$. The latter is determined by the joint asymptotic distribution of τ and $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, which, by the LAE property, is the same for all drifting DGPs with parameters $(\boldsymbol{\phi} + n^{-1/2}\boldsymbol{p}_n, n^{-1/2}\delta)$ provided that $\boldsymbol{p}_n$ converges to zero. ■

References

Abramovitch, L. and K. Singh (1985). "Edgeworth corrected pivotal statistics and the bootstrap," *Annals of statistics*, 13, 116-132.

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

Beran, R. (1997). "Diagnosing bootstrap success," *Annals of the Institute of Statistical Mathematics*, 49, 1–24.

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 (1987). "Implicit alternatives and the local power of test statistics," *Econometrica* 55, 1305-1329.

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

Davidson, R. and J. G. MacKinnon (1998). "Graphical methods for investigating the size and power of hypothesis tests," *The Manchester School*, 66, 1998, 1–26.

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

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

Davidson, R. and J. G. MacKinnon (2001). "Improving the reliability of bootstrap tests," Queen's Institute for Economic Research, Discussion Paper No. 995, revised.

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

Hall, P. (1988). "Theoretical Comparison of Bootstrap Confidence Intervals," *Annals of Statistics*, 16, 927–953.

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

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

Horowitz, J. L. (1997). "Bootstrap methods in econometrics: Theory and numerical performance," in D. M. Kreps and K. F. Wallis (ed.), Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress, Cambridge, Cambridge University Press.

Horowitz, J. L., and N. E. Savin (2000). "Empirically relevant critical values for hypothesis tests," *Journal of Econometrics*, 95, 375–389.

Le Cam, L. (1960). "Locally Asymptotically Normal Families of Distributions," *University of California Publications in Statistics*, 3, 27–98.

Roussas, G. G. (1972). *Contiguity of Probability Measures*, Cambridge, Cambridge University Press.

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

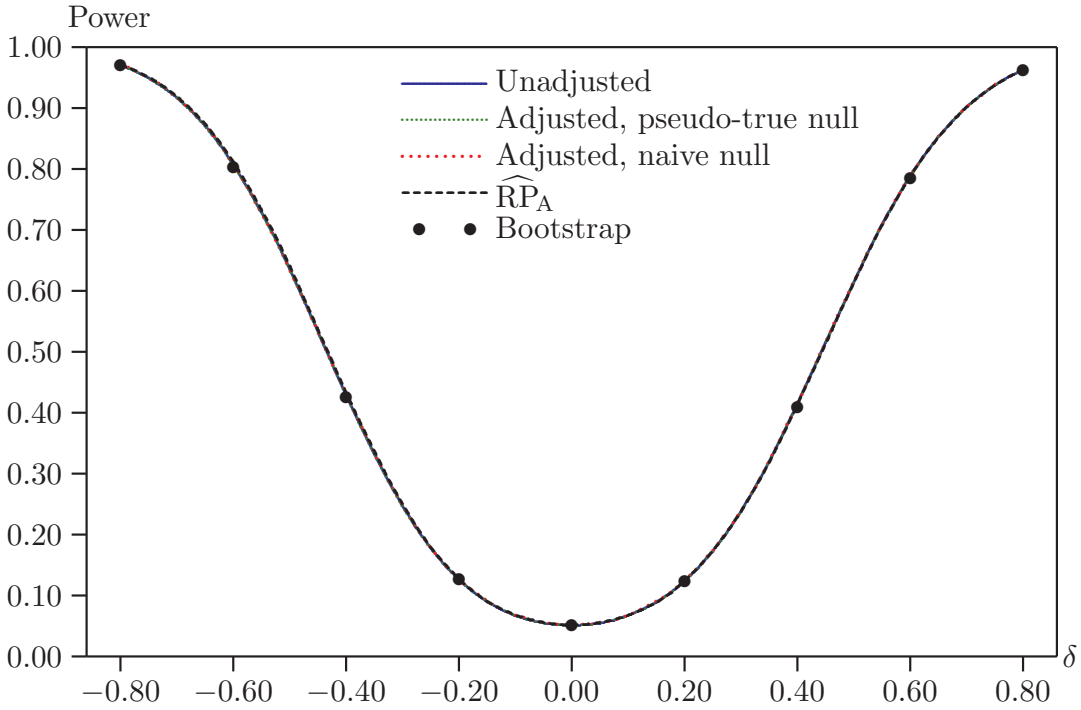

Figure 1. Power functions for logit ES tests

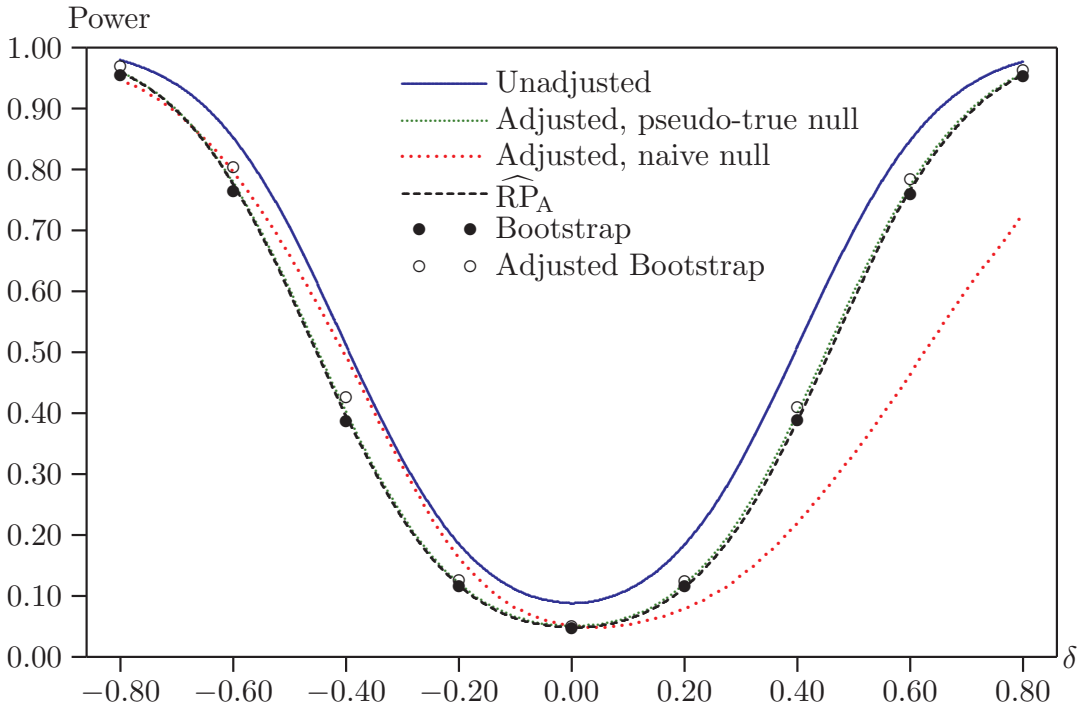

Figure 2. Power functions for logit OPG tests