Queen's Economics Department Working Paper No. 1044

Improving the Reliability of Bootstrap Tests with the Fast Double Bootstrap

Russell Davidson
McGill University

James MacKinnon
Queen's University

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

3-2006

Improving the Reliability of Bootstrap Tests with the Fast Double Bootstrap

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13236 Marseille cedex 02, France

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

Russell.Davidson@mcgill.ca

and

James G. MacKinnon

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

jgm@econ.queensu.ca

Abstract

We first propose two procedures for estimating the rejection probabilities of bootstrap tests in Monte Carlo experiments without actually computing a bootstrap test for each replication. These procedures are only about twice as expensive (per replication) as estimating rejection probabilities for asymptotic tests. We then propose a new procedure for computing bootstrap P values that will often be more accurate than ordinary ones. This "fast double bootstrap" is closely related to the double bootstrap, but it is far less computationally demanding. Simulation results for three different cases suggest that this procedure can be very useful in practice.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Silvia Gonçalves, two referees, and seminar participants at Carleton, Dalhousie, the University of Toronto, Cornell, Penn State, and Rochester for comments on the two unpublished papers — Davidson and MacKinnon (2000a) and MacKinnon (2004) — on which this paper is based.

March, 2006.

1. Introduction

The most appealing way to perform a bootstrap test is to calculate a bootstrap P value. This may be done by seeing where the test statistic falls in the empirical distribution of a number of bootstrap test statistics. The bootstrap P value is simply the proportion of the bootstrap statistics that are more extreme than the actual test statistic. When this P value is sufficiently small, we reject the null hypothesis.

Theory suggests that bootstrap tests will generally perform better in finite samples than asymptotic tests, in the sense that they will commit errors that are of lower order in the sample size n; see, among others, Hall (1992) and Davidson and MacKinnon (1999b). A growing body of evidence from simulation experiments indicates that bootstrap tests do indeed yield more reliable inferences than asymptotic tests in a great many cases; see Davidson and MacKinnon (1999a, 2002a), MacKinnon (2002), Park (2003), and Gonçalves and Kilian (2004), among many others.

Although bootstrap P values are often very reliable, this is certainly not true in every case. For an asymptotic test, one way to check whether it is reliable is simply to use the bootstrap. If the asymptotic and bootstrap P values associated with a given test statistic are similar, we can be fairly confident that the former is reasonably accurate. Of course, having gone to the trouble of computing the bootstrap P value, we may well want to use it instead of the asymptotic one.

In a great many cases, however, asymptotic and bootstrap P values are quite different. When this happens, it is almost certain that the asymptotic P value is inaccurate, but we cannot be sure that the bootstrap one is accurate. One of the contributions of this paper is a technique for computing modified bootstrap P values which will tend to be similar to the ordinary bootstrap P value when the latter is reliable but more accurate when it is unreliable. This technique is closely related to the double bootstrap proposed in Beran (1988), but it is far less expensive to compute. We therefore call it the **fast double bootstrap**, or **FDB**. The amount of computational effort needed to compute an FDB P value, beyond that needed to obtain an ordinary (single) bootstrap P value, is roughly equal to the amount needed to compute the latter in the first place.

In the next section, we discuss basic concepts of bootstrap testing. We emphasize the distinction between symmetric and equal-tail bootstrap tests, which can be important for tests based on statistics that can take either sign. Then, in Section 3, we review some theoretical results on the properties of bootstrap tests. In Section 4, we propose two computationally efficient ways to estimate the performance of bootstrap tests in simulation experiments. In Section 5, we introduce the fast double bootstrap and show how FDB P values may be computed with only twice as much effort as single bootstrap P values. In Section 6, we discuss the relationship between the fast double bootstrap and the double bootstrap itself. Then, in Sections 7, 8, and 9, we present results from simulation experiments on three different types of hypothesis test. These provide some evidence on how well the procedures proposed in this paper work in practice. Section 10 draws some conclusions.

2. Bootstrap Tests

Let τ denote a test statistic, and let $\hat{\tau}$ denote the realized value of τ for a particular sample of size n. The statistic τ is assumed to be asymptotically pivotal, so that the bootstrap yields asymptotic refinements, as we discuss in the next section. For a test that rejects when $\hat{\tau}$ is in the upper tail, such as most tests that asymptotically follow a χ^2 distribution, the true P value of $\hat{\tau}$ is $1 - F(\hat{\tau})$, where F is the cumulative distribution function, or CDF, of τ under the null hypothesis.

If we do not know F, we can often estimate it by using the bootstrap. We generate B bootstrap samples, each of which is used to calculate a bootstrap test statistic τ_j^* for $j = 1, \ldots, B$. We can then estimate $F(\hat{\tau})$ by $\hat{F}_B^*(\hat{\tau})$, where $\hat{F}_B^*(\tau)$ is the empirical distribution function, or EDF, of the τ_j^*. This EDF is often referred to as the **bootstrap distribution**. Then the bootstrap P value is

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

that is, the fraction of the bootstrap samples for which τ_j^* is larger than $\hat{\tau}$. For a test at significance level α, we reject the null hypothesis whenever $\hat{p}^*(\hat{\tau}) < \alpha$.

When τ can take on either positive or negative values, as is the case whenever it has the form of a t statistic, we often wish to perform a two-tailed test. In this case, there are two ways to proceed. The first is to assume that the distribution of τ is symmetric around zero in finite samples, just as it is asymptotically. This leads to the **symmetric bootstrap P value**

$$\hat{p}_{\mathrm{S}}^*(\hat{\tau}) = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}\big(|\tau_j^*| > |\hat{\tau}|\big). \tag{2}$$

As before, we reject the null hypothesis when $\hat{p}_{\mathrm{S}}^*(\hat{\tau}) < \alpha$. There is evidently a close relationship between (1) and (2). Suppose that τ is a t statistic, so that τ^2 is asymptotically $\chi^2(1)$. Then the P value for $\hat{\tau}$ based on (2) is identical to the P value for $\hat{\tau}^2$ based on (1), when both are calculated using the same set of τ_j^*.

The symmetry assumption may often be too strong, in which case we can instead base a test on the **equal-tail bootstrap P value**

$$\hat{p}_{\mathrm{ET}}^*(\hat{\tau}) = 2 \min\left(\frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_j^* < \hat{\tau}),\ \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_j^* > \hat{\tau}) \right). \tag{3}$$

Here we calculate P values for one-tailed tests in each tail and reject if either of these P values is less than $\alpha/2$. Thus we reject when $\hat{\tau}$ either falls below the $\alpha/2$ quantile or above the $1 - \alpha/2$ quantile of $\hat{F}_B^*(\tau)$. The leading factor of 2 is needed because it is twice as likely that $\hat{\tau}$ will be far out in one or other tail of the bootstrap distribution as that it will be far out in one specified tail.

The power of tests based on symmetric and equal-tail bootstrap P values against certain alternatives may be quite different, as is shown in Section 8. Moreover, these tests may have different finite-sample properties when the distribution of τ is not symmetric around zero. There is reason to believe that the bootstrap may perform better in finite samples for tests based on (2) than for tests based on (3), because the order of the bootstrap refinement is often higher for two-tailed than for one-tailed tests; see Hall (1992).

3. Inference from Bootstrap Tests

Beran (1988) shows that bootstrap inference is refined when the quantity bootstrapped is asymptotically pivotal. We formalize the idea of pivotalness by means of a few formal definitions. A **data-generating process**, or **DGP**, is any rule sufficiently specific to allow artificial samples of arbitrary size to be simulated on the computer. Thus all parameter values and all probability distributions must be provided in the specification of a DGP. A **model** is a set of DGPs. Models are usually generated by allowing parameters and probability distributions to vary over admissible sets. A test statistic is a random variable that is a deterministic function of the data generated by a DGP and, possibly, other exogenous variables.

A test statistic τ is a **pivot** for a model $\mathbb{M}$ if, for each sample size n, its distribution is independent of the DGP $\mu \in \mathbb{M}$ which generates the data from which τ is calculated. The **asymptotic distribution** of a test statistic τ for a DGP μ is the limit, if it exists, of the distribution of τ under μ as the sample size tends to infinity. The statistic τ is **asymptotically pivotal** for $\mathbb{M}$ if its asymptotic distribution exists for all $\mu \in \mathbb{M}$ and is independent of μ. Most test statistics commonly used in econometric practice are asymptotically pivotal under the null hypotheses they test, since asymptotically they have distributions, like standard normal or chi-squared, that do not depend on unknown parameters.

If τ is a pivot, then bootstrap inference is exact, even for finite B, provided that $\alpha(B+1)$, or $(\alpha/2)(B+1)$ in the case of an equal-tail test, is an integer. Such a test is often called a **Monte Carlo test**; see Dufour and Khalaf (2001). However, if τ is an asymptotic pivot but not an exact pivot, its distribution depends on which particular DGP μ generates the data used to compute it. In this case, bootstrap inference is no longer exact in general. The bootstrap samples used to estimate the finite-sample distribution of τ are generated by a **bootstrap DGP**, which is in general different from the DGP that generated the original data.

Suppose that data are generated by a DGP μ_0, which belongs to $\mathbb{M}$, and used to compute a realization $\hat{\tau}$ of the random variable τ. Then, for a test that rejects for large values of the statistic, the P value we would ideally like to compute is

$$p(\hat{\tau}) \equiv \Pr_{\mu_0}(\tau > \hat{\tau}). \tag{4}$$

This P value is by construction a drawing from the $\mathrm{U}(0,1)$ distribution.

In practice, (4) can be neither computed analytically nor estimated by simulation, because the DGP μ_0 that generated the observed data is unknown. If τ is an exact pivot, this does not matter, since (4) can be computed using any DGP in $\mathbb{M}$. However, if τ is only an asymptotic pivot, the **theoretical bootstrap *P* value** is defined by

$$p^*(\hat{\tau}, \hat{\mu}) \equiv \Pr_{\hat{\mu}}(\tau > \hat{\tau}), \tag{5}$$

where $\hat{\mu}$ is a (random) bootstrap DGP in $\mathbb{M}$, determined in some suitable way from the same data as those used to compute $\hat{\tau}$. We denote by μ^* the random DGP of which $\hat{\mu}$ is a realization. Observe that $p^*(\hat{\tau}, \hat{\mu})$ is what the bootstrap P value $\hat{p}^*(\hat{\tau})$ defined in (1) converges to as $B \to \infty$.

Let the asymptotic CDF of the asymptotic pivot τ be denoted by F. Throughout this paper, we assume that F is continuous and strictly increasing on its support. At nominal level α, an asymptotic test that rejects for large values of the statistic does so whenever the asymptotic P value $1 - F(\hat{\tau}) < \alpha$. In order to avoid having to deal with different asymptotic distributions, or tests which reject in the left-hand tail or in both tails of the distribution, it is convenient to replace a raw statistic τ by its asymptotic P value, of which the asymptotic distribution under the null is always $\mathrm{U}(0, 1)$ under our assumption. For the remainder of this section, τ denotes a test statistic that is asymptotically distributed as $\mathrm{U}(0, 1)$.

The **rejection probability function**, or **RPF**, provides a measure of the true rejection probability of an asymptotic test for a finite sample. This function, which gives the rejection probability under μ of a test at nominal level α, is defined as follows:

$$R(\alpha, \mu) \equiv \Pr_{\mu}(\tau < \alpha). \tag{6}$$

It is clear that $R(\cdot, \mu)$ is the CDF of τ under μ. For ease of the exposition, we assume that, for all $\mu \in \mathbb{M}$, $R(\cdot, \mu)$ is continuous and strictly increasing on $[0, 1]$, although this is not true of a bootstrap distribution based on resampling from a finite number of observations. In most such cases, the assumption is still a very good approximation.

The information contained in the function R is also provided by the **critical value function**, or **CVF**, defined implicitly by the equation

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

$Q(\alpha, \mu)$ is just the α quantile of τ under μ. It follows from (6) and (7) that

$$R\big(Q(\alpha, \mu), \mu\big) = \alpha, \quad \text{and} \quad Q\big(R(\alpha, \mu), \mu\big) = \alpha, \tag{8}$$

from which it is clear that, for given μ, R and Q are inverse functions.

The bootstrap test rejects at nominal level α if $\tau < Q(\alpha, \mu^*)$, that is, if τ is smaller than the α quantile of τ under the bootstrap DGP. By acting on both sides with $R(\cdot, \mu^*)$, this condition can also be expressed as

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

This makes it clear that the bootstrap P value is just $R(\tau, \mu^*)$. It follows that, if R actually depends on μ^*, that is, if τ is not an exact pivot, the bootstrap test is not equivalent to the asymptotic test, because the former depends not only on τ, but also on the random DGP μ^*. If the true DGP is μ_0, the actual **rejection probability**, or **RP**, of the bootstrap test at nominal level α is

$$\Pr_{\mu_0}\big(\tau < Q(\alpha, \mu^*)\big) = \Pr_{\mu_0}\big(R(\tau, \mu^*) < \alpha\big). \tag{9}$$

In Davidson and MacKinnon (1999b), it is shown that some bootstrap tests enjoy a further refinement, over and above that due to the use of an asymptotic pivot, if τ and μ^* are asymptotically independent. In addition, such asymptotic independence makes it possible to obtain an approximate expression for the RP of a bootstrap test. If τ and μ^* are fully independent under the true DGP, the RP (9) becomes

$$\mathrm{E}_{\mu_0}\Big(\Pr_{\mu_0}\big(\tau < Q(\alpha, \mu^*) \;\big|\; \mu^*\big)\Big) = \mathrm{E}_{\mu_0}\Big(R\big(Q(\alpha, \mu^*), \mu_0\big)\Big). \tag{10}$$

Although this is an exact result only if τ and μ^* are independent, it is approximately true whenever τ and μ^* are only asymptotically independent. As we discuss in the next section, (10) can easily be estimated approximately by simulation.

The asymptotic independence assumption is not very restrictive. A great many test statistics are asymptotically independent of all parameter estimates under the null hypothesis. This is generally true for extremum estimators where the estimates under the null lie in the interior of the parameter space, and for many statistics including all of the classical test statistics for models estimated by nonlinear least squares and maximum likelihood; see Davidson and MacKinnon (1999b). However, it is usually not true for inefficient estimators.

4. Approximating Bootstrap Rejection Frequencies

The conventional way to estimate the bootstrap RP (9) for a given sample size n by simulation is to generate M samples of size n using the DGP μ_0, where, for reasonable accuracy, M must be large. For each replication, indexed by $m = 1, \ldots, M$, a realization τ_m of the statistic τ is computed from the simulated sample, along with a realization $\hat{\mu}_m$ of the bootstrap DGP μ^*. Then B bootstrap samples are generated using $\hat{\mu}_m$, and bootstrap statistics τ^*_{mj}, $j = 1, \ldots, B$ are computed. The realized bootstrap P value for replication m is then

$$\hat{p}^*_m(\tau_m) \equiv \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau^*_{mj} < \tau_m), \tag{11}$$

where we continue to assume that the τ_m and the τ^*_{mj} are in asymptotic P value form. The estimate of (9) is then the proportion of the $\hat{p}^*_m(\tau_m)$ that are less than α. The whole procedure requires the computation of $M(B + 1)$ statistics. If B is not very

large, what is estimated is not really (9), but rather the RP of a bootstrap test based on just B bootstrap repetitions. The bootstrap statistics τ^*_{mj} are realizations of a random variable that we denote as τ^*.

If one wishes to compare the RP of the bootstrap test with that of the underlying asymptotic test, a simulation estimate of the latter can be obtained directly as the proportion of the τ_m less than α. Of course, estimation of the RP of the asymptotic test by itself requires the computation of only M statistics.

The fundamental idea of this paper is that it is possible to obtain a much less expensive approximate estimate of the quantity (10), as follows. As before, for $m = 1, \ldots, M$, the DGP μ_0 is used to draw realizations τ_m and $\hat{\mu}_m$. In addition, $\hat{\mu}_m$ is used to draw a single bootstrap statistic τ^*_m. The τ^*_m are therefore IID realizations of the variable τ^*. We estimate (10) as the proportion of the τ_m that are less than $\hat{Q}^*(\alpha)$, the α quantile of the τ^*_m. This yields the following estimate of the RP of the bootstrap test:

$$\widehat{\mathrm{RP}}_{\mathrm{A}} \equiv \frac{1}{M} \sum_{m=1}^{M} \mathrm{I}\big(\tau_m < \hat{Q}^*(\alpha)\big), \tag{12}$$

where the "A" stands for "approximate" to remind us that (10) is generally valid only as an approximation. As a function of α, $\widehat{\mathrm{RP}}_{\mathrm{A}}$ is an estimate of the CDF of the bootstrap P value (5). Davidson and MacKinnon (1999b) propose a different procedure for estimating the rejection probability of a bootstrap test. Since its performance is almost always worse than that of the procedure proposed here, we do not discuss it.

The estimate (12) is only an approximate estimate of the RP of the bootstrap test not only because it rests on the assumption of the full independence of τ and μ^*, but also because its limit as $B \to \infty$ is not precisely (10). Instead, its limit is something that differs from (10) by an amount of a smaller order of magnitude than the difference between (10) and the nominal level α.

In order to make this statement precise, we begin by obtaining an explicit expression for the CDF of the random variable τ^*. Conditional on the bootstrap DGP μ^*, the CDF of τ^* evaluated at α is $R(\alpha, \mu^*)$. Therefore, if μ^* is generated by the DGP μ_0, the unconditional CDF of τ^* is

$$R^*(\alpha, \mu_0) \equiv \mathrm{E}_{\mu_0}\big(R(\alpha, \mu^*)\big). \tag{13}$$

Denote the α quantile of the distribution of τ^* under μ_0 as $Q^*(\alpha, \mu_0)$. It is defined implicitly by the equation $R^*(Q^*(\alpha, \mu_0), \mu_0) = \alpha$, that is,

$$\mathrm{E}_{\mu_0}\big(R(Q^*(\alpha, \mu_0), \mu^*)\big) = \alpha. \tag{14}$$

We now make some assumptions about the bootstrap procedure under study. First, we suppose that τ is an approximate pivot (in approximate P value form) for the the null-hypothesis model $\mathbb{M}_0$. Thus, for any DGP $\mu \in \mathbb{M}_0$, $R(\alpha, \mu) - \alpha$ is small in some appropriate sense. Otherwise, it would not be sensible to use the given bootstrap

procedure. Next, we assume that R is continuously differentiable with respect to its first argument α for all $\mu \in \mathbb{M}_0$. Thus the statistic τ has a continuous density for all $\mu \in \mathbb{M}_0$. Unlike the first assumption, this assumption *is* restrictive, but it is made in order to simplify the following discussion, the result of which holds true under much less restrictive conditions which are, however, harder to specify precisely. Finally, we assume that $R'(\alpha, \mu) - 1$, where R' denotes the derivative of R with respect to its first argument, is small in the same sense as that in which $R(\alpha, \mu) - \alpha$ is small.

The assumption about the derivative R' implies that $Q(\alpha, \mu) - \alpha$ is small for $\mu \in \mathbb{M}_0$. The definition (13) implies that $R^*(\alpha, \mu) - \alpha$ is small, and so also $Q^*(\alpha, \mu) - \alpha$. By Taylor's Theorem, we can see that

$$R\big(Q(\alpha, \mu^*), \mu_0\big) - R\big(Q(\alpha, \mu_0), \mu_0\big) = (1 + \eta_1)\big(Q(\alpha, \mu^*) - Q(\alpha, \mu_0)\big), \tag{15}$$

where η_1 is a small random quantity. Similarly,

$$R\big(Q^*(\alpha, \mu_0), \mu^*\big) - R\big(Q(\alpha, \mu^*), \mu^*\big) = (1 + \eta_2)\big(Q^*(\alpha, \mu_0) - Q(\alpha, \mu^*)\big), \tag{16}$$

and

$$R\big(Q(\alpha, \mu_0), \mu_0\big) - R\big(Q^*(\alpha, \mu_0), \mu_0\big) = (1 + \eta_3)\big(Q(\alpha, \mu_0) - Q^*(\alpha, \mu_0)\big), \tag{17}$$

where η_2 and η_3 are also small random quantities. If we add equations (15), (16), and (17) together, remembering the identity $R(Q(\alpha, \mu), \mu) = \alpha$, we see that

$$R\big(Q(\alpha, \mu^*), \mu_0\big) + \Big(R\big(Q^*(\alpha, \mu_0), \mu^*\big) - \alpha\Big) - R\big(Q^*(\alpha, \mu_0), \mu_0\big)$$

can be expressed as the sum of quantities that are the product of two small quantities. Taking expectations under μ_0 and using (14), we find that

$$\mathrm{E}_{\mu_0}\Big(R(Q(\alpha, \mu^*), \mu_0)\Big) = R\big(Q^*(\alpha, \mu_0), \mu_0\big) \tag{18}$$

plus the expectation of a sum of products of two small quantities. The left-hand side of (18) is just the second expression in (10), while the right-hand side is the limit of (12) as $M \to \infty$. Thus the error in using (12) to estimate (10) is of smaller order than the difference between the RP of the bootstrap test and the nominal level α. This suggests that the $\widehat{\mathrm{RP}}_{\mathrm{A}}$ procedure will tend to be relatively accurate when the bootstrap test works well and relatively inaccurate when it works poorly.

In practice, it is not necessary to convert test statistics to approximate P value form in order to estimate rejection probabilities. Drawings of the statistics may be obtained in whatever form is most convenient and then sorted in order from the most extreme values to the least extreme. For each value of α of interest, it is then straightforward to compute the proportion of realizations of the statistic more extreme than the realization of the bootstrap statistic in position αM in the sorted list.

For given M, the $\widehat{\text{RP}}_{\text{A}}$ procedure requires about twice as much computational effort as performing an experiment for the asymptotic test, since we need only $2M$ test statistics, the τ_m and the τ_m^*. However, more replications are required to achieve a given level of accuracy. The $\widehat{\text{RP}}_{\text{A}}$ procedure results in drawings of τ and τ^* that are asymptotically independent if τ and μ^* are asymptotically independent. Thus the variance of the estimated RP for a bootstrap test with a given actual RP will always be larger than the variance of the estimated RP for an asymptotic test with the same actual RP. For the asymptotic test, the only source of error is the randomness of the τ_m. For $\widehat{\text{RP}}_{\text{A}}$, there is also the randomness of the τ_m^*, which causes $\hat{Q}^*(\alpha)$ to be random. Thus more replications are needed to achieve a given level of accuracy.

A modified version of this procedure may be used to obtain positively correlated drawings of τ and τ^* and thus reduce the variance of the estimated RP of the bootstrap test. This modified procedure works as follows. Once $\hat{\mu}_m$ has been obtained for replication m, a new set of random numbers, independent of those used to obtain $\hat{\mu}_m$, is drawn. These are then used to compute both τ_m and τ_m^*, the former using μ_0, the latter using $\hat{\mu}_m$. The resulting substantial positive correlation between the τ_m and the τ_m^* reduces the variance of the estimated RP. An additional advantage of this method is that τ and μ^* are genuinely, and not just asymptotically, independent. We call this modified procedure $\widetilde{\text{RP}}_{\text{A}}$. It necessarily involves more computational cost per replication than $\widehat{\text{RP}}_{\text{A}}$, but it may require substantially fewer replications; see Section 7.

The $\widehat{\text{RP}}_{\text{A}}$ and $\widetilde{\text{RP}}_{\text{A}}$ procedures proposed here can be used to estimate the power of bootstrap tests as well as their size. The only thing that changes is that the τ_m are calculated using data generated by a DGP, say μ_1, which does not satisfy the null hypothesis. These data are used to obtain the realizations $\hat{\mu}_m$ of the bootstrap DGP, which in turn are used to generate the data from which the τ_m^* are calculated. See Davidson and MacKinnon (2006) for details.

5. Fast Double Bootstrap *P* Values

The procedures proposed in the previous section are useful only in the context of Monte Carlo experiments. But any procedure that gives an estimate of the RP of a bootstrap test, or, equivalently, of the CDF of the bootstrap P value, allows one to compute a corrected P value. This is just the estimated RP for a bootstrap test at nominal level equal to the uncorrected bootstrap P value. The idea behind the fast double bootstrap is to bootstrap the $\widehat{\text{RP}}_{\text{A}}$ procedure of the previous section, replacing the unknown true DGP μ_0 by the bootstrap DGP $\hat{\mu}$. At least potentially, this leads to more accurate inference than conventional procedures based on the bootstrap P values (1), (2), or (3).

The details are as follows. For each of B bootstrap replications, two different bootstrap statistics are generated. For bootstrap replication j, a bootstrap data set, denoted by $\boldsymbol{y}_j^*$, is first drawn from the bootstrap DGP $\hat{\mu}$. In the same way as the original data are used to obtain both the realized test statistic $\hat{\tau}$ and the realized bootstrap DGP $\hat{\mu}$, the simulated data $\boldsymbol{y}_j^*$ are used to compute two things: a bootstrap statistic, denoted

by τ_j^*, and a second-level bootstrap DGP, denoted by μ_j^{**}. Next, a further simulated data set, denoted by $\boldsymbol{y}_j^{**}$, is drawn using this second-level bootstrap DGP, and a second-level bootstrap test statistic, τ_j^{**}, is computed. This is completely analogous to the $\widehat{\mathrm{RP}}_{\mathrm{A}}$ procedure of the previous section: Here the M drawings τ_m and τ_m^* are replaced by the B drawings τ_j^* and τ_j^{**}, respectively.

Precisely how the fast double bootstrap, or FDB, P value is calculated depends on how the single bootstrap P value is calculated. For the moment, we maintain the convention that τ is asymptotically $\mathrm{U}(0,1)$. In this case, the bootstrap P value is

$$\hat{p}^* \equiv \hat{p}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau_j^* < \hat{\tau}), \tag{19}$$

which is analogous to (1). Except for simulation randomness, a completely correct P value would be the RP under the true DGP μ_0 of the bootstrap test at nominal level $\hat{p}^*$. We can estimate this correct P value by use of the $\widehat{\mathrm{RP}}_{\mathrm{A}}$ procedure based on the bootstrap DGP $\hat{\mu}$ instead of the unknown μ_0. In order to do so, we calculate the $\hat{p}^*$ quantile of the τ_j^{**}, denoted by $\hat{Q}_B^{**}(\hat{p}^*)$ and defined implicitly by the equation

$$\frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\Big(\tau_j^{**} < \hat{Q}_B^{**}\big(\hat{p}^*(\hat{\tau})\big)\Big) = \hat{p}^*(\hat{\tau}). \tag{20}$$

Of course, for finite B, there will be a range of values of Q_B^{**} that satisfy (20), and we must choose one of them somewhat arbitrarily. The FDB P value is now the bootstrap version of (12), namely,

$$\hat{p}_F^{**}(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\Big(\tau_j^* < \hat{Q}_B^{**}(\hat{p}^*(\hat{\tau}))\Big). \tag{21}$$

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

If τ is not asymptotically $\mathrm{U}(0,1)$, and we wish to reject when it is large, then the single bootstrap P value is calculated by (1), and we need the $1-\hat{p}^*$ quantile of the τ_j^{**}, which is defined implicitly by the equation

$$\frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\big(\tau_j^{**} > \hat{Q}_B^{**}(1-\hat{p}^*)\big) = \hat{p}^*. \tag{22}$$

The FDB P value is then given by

$$\hat{p}_F^{**}(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\big(\tau_j^* > \hat{Q}_B^{**}(1-\hat{p}^*)\big). \tag{23}$$

If $\hat{p}^*(\hat{\tau}) = 0$, as may happen quite often when the null hypothesis is false, then it seems natural to define $\hat{Q}_B^{**}(1-\hat{p}^*) = \hat{Q}_B^{**}(1)$ as the largest observed value of the τ_j^{**}, although there are certainly other possibilities. Similarly, when $\hat{p}^*(\hat{\tau}) = 1$, it seems natural to define $\hat{Q}_B^{**}(1-\hat{p}^*) = \hat{Q}_B^{**}(0)$ as the smallest observed value of the τ_j^{**}.

In order to perform an equal-tail FDB test, we need to compute two FDB P values. One is given by (23), and the other by

$$\hat{p}_F^{**}(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\big(\tau_j^* < \hat{Q}_B^{**}(1-\hat{p}^*)\big). \tag{24}$$

The equal-tail FDB P value is then equal to twice the minimum of (23) and (24), by analogy with (3).

A modified version of our FDB procedure would bootstrap the $\widetilde{\mathrm{RP}}_\mathrm{A}$ rather than the $\widehat{\mathrm{RP}}_\mathrm{A}$ procedure. The only difference relative to the above formulas is that both the τ_j^* and the τ_j^{**} would be generated using the bootstrap DGP according to the $\widetilde{\mathrm{RP}}_\mathrm{A}$ recipe instead of the $\widehat{\mathrm{RP}}_\mathrm{A}$ one. However, because this would significantly increase the cost per bootstrap repetition, we have not investigated this modified FDB procedure.

It is of interest to see how well FDB tests work under ideal conditions, when τ, the τ_j^*, and the τ_j^{**} are all independent drawings from the same distribution. To investigate this question, we generate all three statistics from a standard normal distribution for various values of B between 99 and 3999 and then calculate single and FDB bootstrap P values. Single bootstrap tests always reject just about 5% of the time at the .05 level and just about 1% of the time at the .01 level. So do the FDB tests, but only when B is sufficiently large. There is a very noticeable tendency for the FDB tests to reject too often when B is not large, especially in the equal-tail case.

To quantify this tendency, the difference between the FDB rejection frequency and the single bootstrap rejection frequency is regressed on $1/(B+1)$ and $1/(B+1)^2$, with no constant term. These regressions fit extremely well. Results for tests at the .05 and .01 levels are presented in Table 1. There are 54 experiments for the one-tailed and symmetric tests and 47 experiments for the equal-tail tests. There are fewer experiments for the latter because values of B for which $\alpha(B+1)$ is an integer but $(\alpha/2)(B+1)$ is not (namely, 99 and 299) cannot be used. Since each experiment uses 1 million replications, the experimental error should be very small.

In addition to coefficients and standard errors, Table 1 shows the fitted values from each of the regressions for $B = 199$, $B = 999$, and $B = 1999$. It can be seen that all the FDB tests, especially the equal-tail ones, tend to overreject when B is small. Precisely why this is happening is not clear, although it is probably related to the way in which quantiles are estimated. This is a problem for studies of the finite-sample properties of FDB tests, which must avoid using small values of B. In practice, however, overrejection should not be a problem, because any sensible investigator will use a large value of B whenever the bootstrap P value is not well above, or well below, the level of the test; see Davidson and MacKinnon (2000b) for a discussion of how to choose B sequentially when it is expensive to calculate bootstrap test statistics.

6. Relations with the Double Bootstrap

The genuine double bootstrap, as originally laid out in Beran (1988), bootstraps, not the $\widehat{\mathrm{RP}}_{\mathrm{A}}$ procedure, but rather the much more expensive conventional procedure described at the beginning of Section 4. Again the idea is to estimate the RP of the single bootstrap test for a nominal level equal to the single bootstrap P value. Briefly, one proceeds as follows. After having computed the realization $\hat{\tau}$ of the statistic τ and the realization $\hat{\mu}$ of the bootstrap DGP from the real data, one uses B_1 first-level bootstrap samples to compute bootstrap statistics τ_j^* for $j = 1, \ldots, B_1$. If we assume as usual that τ is approximately $\mathrm{U}(0,1)$, the next step is to calculate the first-level bootstrap P value $\hat{p}^*(\hat{\tau})$ according to (19).

Each first-level bootstrap sample is also used to generate a second-level bootstrap DGP μ_j^{**}, which is then used to generate B_2 bootstrap samples from which we compute second-level bootstrap test statistics τ_{jl}^{**} for $l = 1, \ldots, B_2$. For the j^{th} first-level bootstrap sample, the second-level bootstrap P value is

$$\hat{p}_j^{**} = \frac{1}{B_2} \sum_{l=1}^{B_2} \mathrm{I}(\tau_{jl}^{**} < \tau_j^*); \tag{25}$$

compare (11). The **double-bootstrap P value** is the proportion of the $\hat{p}_j^{**}$ that are less than $\hat{p}^*(\hat{\tau})$:

$$\hat{p}^{**}(\hat{\tau}) = \frac{1}{B_1} \sum_{j=1}^{B_1} \mathrm{I}\big(\hat{p}_j^{**} \leq \hat{p}^*(\hat{\tau})\big). \tag{26}$$

The inequality in (26) is not strict, because there may well be cases for which $\hat{p}_j^{**} = \hat{p}^*(\hat{\tau})$. For this reason, it is desirable that $B_2 \neq B_1$.

If $\hat{\tau}$, the τ_j^*, and the τ_{jl}^{**} all come from the same distribution, then, for $B_1 = B_2 = \infty$, the double bootstrap yields exactly the same inferences as the single bootstrap. Suppose, instead, that the bootstrapping process causes the distribution of the τ_j^* to contain fewer extreme values than the distribution of τ itself. Therefore, the P values associated with moderately extreme values of $\hat{\tau}$ are too small. But it is reasonable to expect that the distributions of the τ_{jl}^{**} contain even fewer extreme values than the distribution of the τ_j^*. Therefore, the $\hat{p}_j^{**}$ should tend to be too small, at least for small values of $\hat{p}^*(\hat{\tau})$. This implies that the double-bootstrap P value $\hat{p}^{**}(\hat{\tau})$ will be larger than $\hat{p}^*(\hat{\tau})$, which is exactly what we want. By a similar argument, $\hat{p}^{**}(\hat{\tau})$ will tend to be smaller than $\hat{p}^*(\hat{\tau})$ when the distribution of the τ_j^* contains more extreme values than the distribution of τ itself. The same intuition applies as well to the FDB.

Of course, even the double bootstrap cannot be expected to work perfectly. Just as the first-level bootstrap distribution may provide an inadequate approximation to the distribution of τ under μ_0, so may the distribution of the second-level bootstrap P values provide an inadequate approximation to that of the first-level P values. In principle, any bootstrap procedure, including the double bootstrap, may or may not provide an acceptable approximation to the true P value associated with $\hat{\tau}$.

The advantage of the double bootstrap, relative to the new FDB procedure, is that it does not require any sort of independence between the bootstrap DGP and the test statistic. But this comes at an enormous computational cost. For each of B_1 bootstrap samples, we need to compute $B_2 + 1$ test statistics. Thus the total number of test statistics that must be computed is $1 + B_1 + B_1 B_2$. Even if B_2 is somewhat smaller than B_1, as is often recommended, this will be vastly more expensive than computing $1 + 2B$ test statistics, for reasonable values of $B \approx B_1$ and $B_2 \leq B_1$. For example, if $B = B_1 = 999$ and $B_2 = 399$, the FDB procedure involves computing 1999 test statistics, while the genuine double bootstrap involves computing no fewer than 399,601 of them.

7. Tests for Omitted Variables in Probit Models

In this section and the next two, we present results from a number of simulation experiments designed to see whether the procedures proposed in this paper can work well enough to be useful in practice. In this section, we focus on whether $\widehat{\mathrm{RP}}_{\mathrm{A}}$ and $\widetilde{\mathrm{RP}}_{\mathrm{A}}$ provide good approximations to the actual rejection probabilities for single bootstrap tests based on $\hat{p}^*$, and on whether the FDB procedure can yield bootstrap tests with smaller errors in rejection probability than single bootstrap tests.

We begin by studying the OPG version of the LM test for omitted variables in the probit model. This test has noticeably worse finite-sample properties than other tests of the same hypothesis, such as the LR test and the efficient score version of the LM test; see Davidson and MacKinnon (1984). Therefore, it should rarely be used in practice. However, its poor finite-sample performance makes it a good example for the study of alternative bootstrap procedures.

The probit model we study can be written as

$$\mathrm{E}(y_t \,|\, \boldsymbol{X}_t) = \Phi(\boldsymbol{X}_{1t}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}\boldsymbol{\beta}_2), \tag{27}$$

where y_t is a binary dependent variable that can equal 0 or 1, $\Phi(\cdot)$ is the cumulative standard normal distribution function, $\boldsymbol{X}_t = [\boldsymbol{X}_{1t} \;\; \boldsymbol{X}_{2t}]$ is a $1 \times k$ vector of exogenous variables, with $k = k_1 + k_2$, $\boldsymbol{\beta}_1$ is a k_1–vector, and $\boldsymbol{\beta}_2$ is a k_2–vector. The null hypothesis is that $\boldsymbol{\beta}_2 = \mathbf{0}$. The OPG test statistic is the explained sum of squares from a regression of an n–vector of 1s on the derivatives of the contributions to the loglikelihood with respect to each of the parameters, evaluated at the ML estimates under the null hypothesis. See Davidson and MacKinnon (1984) for more details.

We report experimental results for two different cases. In Case 1, $k_1 = 2$, $k_2 = 6$, and $\boldsymbol{\beta}_1^\top = [0 \;\; 1]$. The number of restrictions, k_2, is relatively large because the finite-sample performance of the test becomes worse as k_2 increases, and preliminary experiments revealed that the finite-sample performance of the single bootstrap test does likewise. In Case 2, $k_1 = 2$, $k_2 = 8$, and $\boldsymbol{\beta}_1^\top = [1 \;\; 2]$. With these parameter values, the bootstrap test performs worse because there are more restrictions, the probit model fits better, and the proportion of 0s in the sample is much less than one-half.

For each of the two cases, we perform 161 experiments, with 100,000 replications each, for all sample sizes between 40 and 200. The exogenous variables, other than the constant, are redrawn from the standard normal distribution for each replication, so as to avoid undue dependence on the design matrix. In these experiments, rejection frequencies of the asymptotic test and approximate rejection probabilities of the bootstrap test (both $\widehat{\text{RP}}_\text{A}$ and $\widetilde{\text{RP}}_\text{A}$) are estimated. In addition, we perform 17 much more expensive experiments, for $n = 40, 50, 60, \ldots, 200$, also with 100,000 replications, in which we estimate the actual performance of the bootstrap test and the FDB test using $B = 999$. The results of these experiments are presented graphically in Figures 1 through 3.

Figure 1 shows the rejection frequencies for the asymptotic tests, which always overreject severely, much more so for Case 2 than for Case 1. As expected, the overrejection gradually diminishes as the sample size increases, after an initial increase for Case 2, but it remains quite substantial even at $n = 200$.

Figures 2 and 3 pertain to Cases 1 and 2, respectively. Each of these figures shows the rejection frequencies for the (single) bootstrap and FDB tests, along with the approximate rejection probabilities given by $\widehat{\text{RP}}_\text{A}$ and $\widetilde{\text{RP}}_\text{A}$. Compared with the asymptotic tests, the bootstrap tests always perform remarkably well. However, they may either underreject or overreject for small sample sizes and then underreject for a range of somewhat larger sample sizes. The reason for the initial overrejection in Case 2 is explained below. Moreover, as can be seen from the figures, the approximations $\widehat{\text{RP}}_\text{A}$ and $\widetilde{\text{RP}}_\text{A}$ are almost always very good indeed, except for the very smallest sample size.

In Section 4, we discussed the relationship between the variances of $\widehat{\text{RP}}_\text{A}$, $\widetilde{\text{RP}}_\text{A}$, and the estimated rejection probability for the asymptotic test. In order to investigate this matter, we regress both estimates of bootstrap rejection probability errors on a number of powers of $n^{-1/2}$ (with no constant, since asymptotically there is no error) for each of the two cases. The standard errors of the preferred regressions are estimates of the magnitude of experimental error. For $\widehat{\text{RP}}_\text{A}$, these standard errors are 0.000869 and 0.001004 for Cases 1 and 2, respectively. For $\widetilde{\text{RP}}_\text{A}$, the corresponding standard errors are 0.000650 and 0.000765. Thus it appears that, as expected, $\widetilde{\text{RP}}_\text{A}$ can produce results with noticeably less experimental error than $\widehat{\text{RP}}_\text{A}$.

In almost all cases, the FDB test outperforms the single bootstrap test. This is most noticeable for the smallest sample sizes and, in Case 2, for the larger sample sizes where the single bootstrap test systematically underrejects. The few cases in which FDB does not perform better seem to occur when both methods perform very well, and some of them can probably be attributed to experimental error.

The tendency of the single bootstrap test to overreject in very small samples for Case 2 has a simple explanation. In these cases, ML estimation of the null model not infrequently achieves a perfect fit. When this happens, the test statistic is equal to zero. As is well known, probit models tend to fit too well in small samples. Therefore, the slope coefficients used to generate the bootstrap samples tend to be larger than the ones used to generate the original samples; see MacKinnon and Smith (1998). This means that perfect fits are achieved more often for the bootstrap samples than they

are for the original samples. In consequence, there are fewer large values of the τ_j^* than there are of the τ_j, and the bootstrap P values are therefore biased downwards. This problem tends to go away rapidly as n increases.

We calculate genuine double bootstrap P values in a few experiments. Despite having only 10,000 replications, with $B_1 = 399$ and $B_2 = 199$, these experiments are far more expensive than any of the others. Table 2, in which these results are reported, provides no evidence to suggest that double bootstrap P values are any more accurate than FDB P values. In Case 2 with 40 observations, where perfect fits occur with some frequency, the double bootstrap actually performs substantially less well. In the other five cases, bearing in mind that the standard errors of the estimated rejection frequencies are roughly 0.0022, there is little to choose between them.

8. Tests for Serial Correlation

The simulation experiments of the second set concern tests for serial correlation. They are designed in part to shed light on the choice between symmetric and equal-tail tests, which can have quite different power properties.

Commonly-used tests for serial correlation are not exact in models with lagged dependent variables or nonnormal disturbances. Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \gamma y_{t-1} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2), \tag{28}$$

where there are n observations, and $\boldsymbol{X}_t$ is a $1 \times k$ vector of observations on exogenous variables. The null hypothesis is that $\rho = 0$. A simple and widely-used test statistic for serial correlation in this model is the t statistic on $\hat{u}_{t-1}$ in a regression of y_t on $\boldsymbol{X}_t$, y_{t-1}, and $\hat{u}_{t-1}$. This procedure is proposed in Durbin (1970) and Godfrey (1978). The test statistic is asymptotically distributed as $\text{N}(0, 1)$ under the null hypothesis. Since this test can either overreject or underreject in finite samples, it is natural to use the bootstrap in an effort to improve its finite-sample properties.

In order to bootstrap the Durbin-Godfrey test under weak assumptions about the ε_t, we first estimate the regression in (28) by ordinary least squares. This yields $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}$, and a vector of residuals with typical element $\hat{u}_t$. It is then natural to generate the bootstrap data using the semiparametric bootstrap DGP

$$y_t^* = \boldsymbol{X}_t\hat{\boldsymbol{\beta}} + \hat{\gamma} y_{t-1}^* + u_t^*, \tag{29}$$

where the u_t^* are obtained by resampling the vector of rescaled residuals with typical element $(n/(n-k-1))^{1/2}\hat{u}_t$. The initial value y_0^* is set equal to the actual pre-sample value y_0. The bootstrap DGP (29) imposes the IID assumption on the disturbances without imposing any additional distributional assumptions.

In all the reported experiments, the disturbances are normally distributed, the first column of the $\boldsymbol{X}$ matrix is a constant, and the remaining columns are generated by independent, stationary AR(1) processes with normal innovations and parameter ρ_x. A

new $\boldsymbol{X}$ matrix is drawn for each replication. Both asymptotic and bootstrap rejection frequencies are found to depend strongly on k, ρ_x, σ_ε, and γ, as well as on the sample size n. Since the performance of the asymptotic test improves rapidly as n increases, $n = 20$ is used for most of the experiments.

Asymptotic results are based on 200,000 replications for values of γ between -0.99 and 0.99 at intervals of 0.01. Bootstrap results are based on 100,000 replications for values of γ between -0.9 and 0.9 at intervals of 0.1 using 1999 bootstrap samples. This is an unusually large number to use in a Monte Carlo experiment. It is used because the results in Table 1 suggest that the equal-tail FDB tests will tend to overreject noticeably if B is not quite large.

Results under the null

Figure 4 shows three sets of rejection frequencies for the performance of asymptotic and bootstrap tests under the null hypothesis when $n = 20$. These are representative of the results for a much larger number of similar experiments. Rejection frequencies for tests at the .05 level are shown on the vertical axis, and γ is shown on the horizontal axis. Each row concerns the same set of experiments. Results for the asymptotic test are shown in both panels. The left-hand panel shows rejection frequencies for symmetric bootstrap and FDB tests, and the right-hand panel shows rejection frequencies for equal-tail bootstrap and FDB tests.

The first row of the figure contains results for a case in which all the bootstrap tests work very well. In the left-hand panel, we see that there is very little difference between the rejection frequencies for the symmetric bootstrap test, based on (2), and for its FDB variant. This is not merely true on average, but also for every replication: The correlation between the two P values is 0.999 for every value of γ. Thus an investigator who performs both tests would obtain extremely similar results and would probably conclude, correctly, that the bootstrap P value is very reliable.

In the right-hand panel of the first row of the figure, we see that the equal-tail bootstrap test is generally not quite as reliable as the symmetric bootstrap test. Moreover, the FDB procedure yields noticeably different rejection frequencies which are, in most cases, closer to the nominal level of .05. However, the correlation between the two P values is still very high at approximately 0.996 for all values of γ.

The second and third rows of the figure show results for cases in which, on average, the bootstrap tests do not work as well. In both cases, $\sigma = 10$, which is ten times larger than for the case in the first row, and $k = 6$, which is twice as large. Thus the bootstrap DGP depends on more parameters, and they are estimated less precisely. The only difference between the two cases is that $\rho_x = 0.8$ in the second row, and $\rho_x = -0.8$ in the third row.

Several interesting results are evident in the second and third rows of the figure. All four bootstrap tests generally work much better than the asymptotic test on which they are based. It is apparent that a symmetric bootstrap test can overreject when an equal-tail test underrejects, and *vice versa*. However, the equal-tail tests seem to be a bit more prone to overreject than the symmetric tests. The FDB tests generally work

better than the single bootstrap tests, especially when the latter are least reliable. Nevertheless, the correlations between the single bootstrap and FDB tests remain quite high. They are never less than 0.976 for the equal-tail tests and 0.998 for the symmetric ones.

It is of interest to see how fast the performances of the single bootstrap and FDB tests improve as the sample size increases. Figure 5 contains six panels, comparable to those in Figure 4. In each of these experiments, γ is fixed at a value associated with relatively poor performance of at least one of the tests for $n = 20$, and n takes on the values 10, 14, 20, 28, 40, 56, 80, 113, 160, 226, and 320. Each of these sample sizes is larger than the previous one by a factor of approximately $\sqrt{2}$. As before, there are 100,000 replications, and $B = 1999$.

The left-hand panel of the first row shows that the symmetric bootstrap and FDB tests work extremely well for all sample sizes when $k = 3$ and $\sigma_\varepsilon = 1$. There is essentially nothing to choose between them. However, as can be seen from the right-hand panel, the equal-tail tests tend to underreject for very small values of n in this case, with the FDB tests underrejecting less severely than the single bootstrap tests.

The next two rows of the figure, in which $k = 6$ and $\sigma_\varepsilon = 10$, are more interesting. We see both noticeable overrejection and noticeable underrejection by the single bootstrap tests. With a few exceptions, the FDB tests perform substantially better than the single bootstrap tests when the latter perform badly. The results in the right-hand panel of the second row and the left-hand panel of the third row are particularly dramatic. In these cases, the gain from using the FDB procedure is quite substantial.

It appears that the equal-tail FDB tests overreject slightly for large values of n. This appears to be a manifestation of the phenomenon seen in Table 1. Since the magnitude of the overrejection is just about what we would expect from the results in Table 1, allowing for a certain amount of experimental error, it would surely be even smaller if B were larger than 1999.

Results under the alternative

Figure 6 shows power functions for six sets of experiments. The value of ρ is on the horizontal axis, and the rejection frequency is on the vertical axis. Asymptotic results are based on 200,000 replications for 199 values of ρ between -0.99 and 0.99, and bootstrap results are based on 100,000 replications for 19 values of ρ between -0.9 and 0.9. Every panel shows results for both symmetric and equal-tail tests. Because the single bootstrap and FDB tests always have essentially the same power, their symbols always overlap. Thus it may not be immediately apparent that the same symbols are used as in Figures 4 and 5.

In the first two rows of the figure, $n = 20$. In the four panels in these rows, the shapes of the asymptotic power functions differ dramatically from the inverted bell shape that they must have asymptotically. The power functions for the symmetric bootstrap tests always have essentially the same shape as those for the asymptotic tests, although with a vertical displacement that is quite large in the case of the left-hand panel in the second row. This vertical displacement arises because the asymptotic test overrejects

quite severely under the null hypothesis. The symmetric bootstrap test, which does not overreject, inevitably has noticeably less power against all alternatives.

In contrast, the shapes of the power functions for the equal-tail bootstrap tests are dramatically different from the ones for the symmetric bootstrap tests. The former have somewhat less power in whichever direction the asymptotic tests have high power, but they have much more power in the other direction. Specifically, when ρ_x and γ are both positive, the equal-tail tests always have more power against positive values of ρ than the symmetric tests, and the differences are often dramatic. Since this is a case that we might expect to encounter quite frequently, this is an important result.

In the third row of Figure 6, $n = 40$. Increasing the value of n brings the shape of the asymptotic power functions much closer to the inverted bell shape that they should have, as can be seen by comparing the left-hand panel in the top row with the left-hand panel in the bottom row and the left-hand panel in the middle row with the right-hand panel in the bottom row. However, it does not change the results about the power of the symmetric and equal-tail bootstrap tests. The equal-tail tests have somewhat less power against negative values of ρ and a great deal more power against positive values than do the symmetric tests, because the power functions of the former are much closer to being symmetric about $\rho = 0$.

In several panels of Figure 6, the asymptotic tests are reasonably reliable under the null. Nevertheless, there are substantial gains in power to be had from using equal-tail bootstrap tests instead of asymptotic tests. This suggests that equal-tail bootstrap tests for serial correlation should be used routinely, even when (indeed, perhaps especially when) there is no reason to believe that asymptotic tests are unreliable.

9. Tests for ARCH

Since the seminal work of Engle (1982), it has been recognized that serial dependence in the variance of the disturbances of regression models using time-series data is a very common phenomenon. In the case of financial data at high or moderate frequencies, there is not much point simply testing for ARCH disturbances, because we know that we will find strong evidence of them, whether or not ARCH is actually the best way to model the properties of the disturbances. However, in the case of low-frequency financial data, or non-financial macroeconomic data, the hypothesis of serial independence is not unreasonable, and it may therefore make sense to test for ARCH.

Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \delta_1\sigma_{t-1}^2, \quad \varepsilon_t \sim \text{IID}(0,1). \tag{30}$$

The disturbances of this model follow the GARCH(1,1) process introduced by Bollerslev (1986). It is easy to generalize this process to have more lags of u_t^2, more lags of σ_t^2, or both. In this paper, however, attention is restricted to the GARCH(1,1) process, partly for simplicity, and partly because this process generally works extraordinarily well in practice.

The easiest way to test the null hypothesis that the u_t are IID in the model (30) is to run the regression

$$\hat{u}_t^2 = b_0 + b_1 \hat{u}_{t-1}^2 + \text{residual}, \tag{31}$$

where $\hat{u}_t$ is the t^{th} residual from an OLS regression of y_t on $\boldsymbol{X}_t$. The null hypothesis that $\alpha_1 = \delta_1 = 0$ can be tested by testing the hypothesis that $b_1 = 0$. For a simple derivation of the test regression (31), and an explanation of why it has just two coefficients even though the GARCH(1,1) model has three, see Davidson and MacKinnon (2004, Section 13.6).

There are several valid test statistics based on regression (31). These include the ordinary t statistic for $b_1 = 0$, which is asymptotically distributed under the null as $\mathrm{N}(0,1)$, and n times the centered R^2, which is asymptotically distributed as $\chi^2(1)$. Results are reported only for the second of these statistics, partly because it seems to be the most widely used test for ARCH, and partly because it generalizes easily to tests for higher-order ARCH and GARCH processes, in which there are more lags of $\hat{u}_t^2$ in the test regression. It would be interesting to compare the finite-sample performance of alternative tests, but that would require another paper.

Figures 7 through 10 report results from a number of simulation experiments which focus on the effects of the sample size and the distribution of the ε_t. In all the reported experiments, $\boldsymbol{X}_t$ consists of a constant and two independent, standard normal random variates, since changing the number of regressors has only a modest effect on the finite-sample behavior of the tests. The sample size takes on the values 10, 14, 20, 28, 40, 56, 80, 113, 160, 226, and 320. The ε_t are either standard normal, Student's t with 4 degrees of freedom, or $\chi^2(2)$ rescaled and recentered to have mean 0 and variance 1. The first of these distributions is in some sense the base case, the second involves severe leptokurtosis, and the third involves severe skewness. As is easily shown, the test statistics are invariant to the variance of the disturbances under the null hypothesis.

The top left-hand panel of Figure 7 shows rejection frequencies for the asymptotic test as a function of the sample size for the three different distributions of the ε_t. These results are based on 100,000 replications for each value of n. The test underrejects severely in all cases, especially when the ε_t are nonnormal. As we would expect, performance improves with the sample size, but the rate of improvement is fairly slow, especially when the ε_t are $t(4)$.

There are several ways to bootstrap this test. One possibility is to use a parametric bootstrap, drawing the simulated disturbances from the normal distribution. It is easy to see that this leads to an exact test when the disturbances actually are normally distributed. The test statistic depends solely on the $\boldsymbol{X}$ matrix and the vector of innovations $\boldsymbol{\varepsilon}$. The former is known. If the distribution of the latter is also known, then that of the test statistic does not depend on any unknown features of the DGP. It then follows by standard arguments for Monte Carlo tests that, if B is chosen so that $\alpha(B+1)$ is an integer, the parametric bootstrap test is exact; see Dufour *et al.* (2004).

The top right-hand panel of Figure 7 shows rejection frequencies for parametric bootstrap tests with $B = 1999$. As expected, these tests work perfectly when the ε_t are actually normally distributed. The very small deviations from a frequency of 0.05 are well within the margins of experimental error. However, the tests are evidently not exact when the disturbances are not normally distributed. For the largest sample sizes, they are no better than the corresponding asymptotic tests. Since the FDB tests performed almost identically to the parametric bootstrap tests on which they are based, results for parametric FDB tests are not shown.

Of course, we would not expect parametric bootstrap tests to perform well when they are based on an incorrect distributional assumption, and we would not expect the FDB procedure to help. It therefore seems natural to use a semiparametric bootstrap DGP, like the one in equation (29). Results for this procedure are shown in the bottom left-hand panel of Figure 7 and in the two left-hand panels of Figure 8. For the normal distribution, the semiparametric bootstrap test underrejects, quite noticeably so for the smaller sample sizes. Interestingly, except for the very smallest sample sizes, the FDB version performs considerably better. It appears to be essentially exact for $n \geq 80$, whereas the single semiparametric bootstrap test always underrejects.

It is more interesting to see what happens when the disturbances are nonnormal. When they are $t(4)$, the semiparametric bootstrap test underrejects quite severely for small sample sizes. However, its performance gradually improves as n increases, and there is a noticeable gain from using the FDB procedure, except when n is very small. When the disturbances are recentered $\chi^2(2)$, the underrejection is even more severe for small sample sizes, but the rate of improvement as n increases is much more rapid. Once again, there is generally a noticeable gain from using the FDB procedure.

The errors committed by the semiparametric bootstrap test must arise from the fact that the empirical distribution of the residuals provides an inadequate approximation to the distribution of the disturbances. One way to improve this approximation is to smooth the bootstrap disturbances. This can be done by using a kernel estimator. The kernel estimator of the CDF of u at the point u', using a sample of n residuals $\hat{u}_t$, is given by

$$\hat{F}_h(u') = \frac{1}{n} \sum_{t=1}^{n} K(\hat{u}_t, u', h), \tag{32}$$

where $K(\hat{u}_t, u', h)$ is a cumulative kernel, such as the standard normal CDF, called the Gaussian kernel, and h is the bandwidth; see Azzalini (1981) and Reiss (1981). A reasonable choice for h is $1.587\hat{\sigma}n^{-1/3}$, where $\hat{\sigma}$ is the standard deviation of the (possibly rescaled) residuals.

To draw bootstrap disturbances from (32), we simply resample from the residuals $\hat{u}_t$ and then add independent normal random variables with variance h^2. The resulting bootstrap disturbances have expectation zero because both the residuals and the normal random variates do. They also have too much variance, but they can easily be rescaled. However, in the context of tests for ARCH errors, this rescaling is not needed, because the test statistics are invariant to the variance of the disturbances.

The bottom right-hand panel of Figure 7 and the two right-hand panels of Figure 8 show the effects of using bootstrap disturbances that are smoothed in this way, where the Gaussian kernel with the bandwidth given above is used. When the disturbances are actually normal, resampling smoothed residuals works substantially better than resampling ordinary residuals for small sample sizes. This presumably occurs because smoothing brings the distribution of the disturbances closer to normality. However, there appears to be no appreciable gain from smoothing when the disturbances are $t(4)$ or recentered $\chi^2(2)$. As in the case where smoothing is not used, the FDB rejection frequencies are always noticeably closer to 0.05 than those of the single bootstrap, except when the sample size is very small.

Because Figures 7 and 8 deal only with tests at the 0.05 level, they do not tell the whole story. To show the effect of the level of the test, Figure 9 plots the difference between the rejection frequency and the level of the test for all levels between 0.005 and 0.25, at intervals of 0.005, for two sample sizes, 40 and 160. The nominal level is on the horizontal axis, and the "rejection frequency discrepancy" is on the vertical axis. Several interesting facts emerge from this figure. First, the asymptotic test can actually overreject for small levels. Second, for nonnormal disturbances, the distortion of the asymptotic test becomes steadily worse as the level increases. Finally, and of most interest for this paper, the improvement from using the FDB rather than the single bootstrap becomes larger as the level of the test increases. Moreover, the extent of the improvement is greater for $n = 160$ than for $n = 40$, especially in relative terms. To see this, compare the left-hand and right-hand panels in the second and third rows of the figure.

It is natural to ask whether the FDB procedure works as well as the full double bootstrap. Figure 10 provides some evidence on this point. The experiments are similar to those in the left-hand panels of Figure 8, except that values of n greater than 160 are omitted. They involve semiparametric bootstrap DGPs that use resampled residuals, with disturbances that are either $t(4)$ or recentered $\chi^2(2)$. There are 100,000 replications. However, because computational cost is an issue, B_1 and B_2 are just 399 and 199, respectively. Even with such a small value of B_2, the double bootstrap is about 100 times more expensive to compute than the FDB in this case.

In both panels of Figure 10, it is evident that all the bootstrap procedures work much better than the asymptotic test. Moreover, there is a clear ordering, with the FDB performing noticeably better than the single bootstrap, and the double bootstrap performing a little better than the FDB. The advantage of the double bootstrap over the FDB is a bit greater for the experiments with $\chi^2(2)$ disturbances than for the ones with $t(4)$ disturbances, but it is never striking. Thus, at least in this case, the failure of the FDB to work perfectly appears to be attributable mainly to the limitations of the double bootstrap itself rather than to a failure of the independence assumption.

10. Conclusions

In this paper, we have proposed two closely related techniques to solve two different problems. The first problem is the high cost of Monte Carlo experiments that involve bootstrap tests. Our $\widehat{\text{RP}}_{\text{A}}$ and $\widetilde{\text{RP}}_{\text{A}}$ procedures make it possible to study the finite-sample performance of bootstrap tests for only about three or four times the computational cost of studying asymptotic tests. In contrast, with the standard approach, the cost is generally hundreds of times as great. Of course, because our procedures are valid only under an independence assumption that may not hold in finite samples, it will always be essential to simulate actual bootstrap tests for at least a few cases to verify that they are yielding accurate results.

The second problem is the errors in rejection probability that sometimes occur for bootstrap tests. Our FDB procedure seems to reduce these errors quite substantially in some cases. In fact, for the probit model example we looked at, it does so just as effectively as a genuine double bootstrap procedure, but at very much less computational expense. For the ARCH errors example, it performs almost as well as a genuine double bootstrap.

We do not claim that the FDB procedure will always be useful. In some cases, if an investigator chooses the right test statistic to bootstrap and the right bootstrap DGP, single bootstrap tests may work so well that there is nothing to be gained by using the FDB. In such a case, the single bootstrap and FDB P values are likely to be very similar, and this fact may be taken as evidence that both procedures are reliable. In other cases, neither the double bootstrap nor the FDB may help very much.

Since the FDB was originally proposed in Davidson and MacKinnon (2000a), we and others have performed a number of additional simulations. In particular, Davidson and MacKinnon (2002b) show that the FDB works very well when applied to the J test for nonnested linear regression models. In most cases, the single bootstrap based on resampling residuals works adequately for the J test. However, there are extreme cases in which single bootstrap tests overreject noticeably, and, in these cases, FDB tests work very much better. For a number of other tests, however, using the FDB apparently leads to more modest improvements; see Omtzigt and Fachin (2002), Lamarche (2004), and Davidson (2006).

One interesting result of this paper is not specifically related to the FDB procedure. For the experiments discussed in Section 7, equal-tail bootstrap tests can be much more powerful than either asymptotic tests or symmetric bootstrap tests, even when the asymptotic tests are well-behaved under the null. This suggests that equal-tail bootstrap tests deserve closer investigation for a variety of problems where two-tailed tests are commonly used.

Based on all the experimental results, it is tempting to draw the following conclusions, which are admittedly somewhat speculative. When the single bootstrap works almost perfectly, the FDB procedure will produce extremely similar results, provided B is reasonably large. When the single bootstrap improves matters substantially, the FDB procedure will lead to a significant further improvement if there is still much room for

improvement. But when the single bootstrap leads to only modest improvement, the FDB procedure will not help matters very much.

References

Azzalini, A. (1981). "A note on the estimation of a distribution function and quantiles by a kernel method," *Biometrika*, **68**, 326–328.

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

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, **31**, 307–27.

Davidson, J. (2006). "Alternative bootstrap procedures for testing cointegration in fractionally integrated processes," *Journal of Econometrics*, forthcoming.

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 (1999a). "Bootstrap testing in nonlinear models," *International Economic Review*, **40**, 487–508.

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

Davidson, R., and J. G. MacKinnon (2000a). "Improving the reliability of bootstrap tests," Queen's Institute for Economic Research Discussion Paper No. 995.

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

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

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

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

Davidson, R., and J. G. MacKinnon (2006). "The power of bootstrap and asymptotic tests," *Journal of Econometrics*, forthcoming.

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

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

Durbin, J. (1970). "Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables," *Econometrica*, **38**, 410–421.

Engle, R. F. (1982). "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, **50**, 987–1007.

Godfrey, L. G. (1978). "Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables," *Econometrica*, **46**, 1293–1301.

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

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

Lamarche, J.-F. (2004). "The numerical performance of fast bootstrap procedures," *Computational Economics*, **23**, 379–389.

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

MacKinnon, J. G. (2004). "Applications of the fast double bootstrap," paper presented at Joint Statistical Meeting, Toronto.

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

Omtzigt, P., and S. Fachin (2002). "Bootstrapping and Bartlett corrections in the cointegrated VAR model," University of Amsterdam Discussion Paper No. 2002/15.

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

Reiss, R. D. (1981). "Nonparametric estimation of smooth distribution functions," *Scandinavian Journal of Statistics*, **9**, 65–78.

Table 1. Response surface regressions for FDB overrejection as a function of B

Test	$1/(B+1)$	$1/(B+1)^2$	$B = 199$	$B = 999$	$B = 1999$
Symmetric .05	0.3466 (.0102)	−6.02 (1.21)	0.001583	0.000341	0.000172
One-tailed .05	0.3489 (.0111)	−6.01 (1.32)	0.001595	0.000343	0.000173
Equal-tail .05	1.7654 (.0208)	−51.10 (4.77)	0.007550	0.001714	0.000870
Symmetric .01	0.3886 (.0073)	−15.84 (0.86)	0.001547	0.000373	0.000190
One-tailed .01	0.3695 (.0073)	−13.64 (0.86)	0.001507	0.000356	0.000181
Equal-tail .01	1.6865 (.0174)	−140.89 (3.96)	0.004910	0.001546	0.000808

Table 2. Rejection frequencies at .05 level, 10,000 replications

	n	B_1	B_2	Bootstrap	FDB	Double Bootstrap
Case 1	40	399	199	0.0402	0.0447	0.0462
Case 1	80	399	199	0.0447	0.0502	0.0511
Case 1	160	399	199	0.0514	0.0543	0.0531
Case 2	40	399	199	0.0685	0.0564	0.0330
Case 2	80	399	199	0.0503	0.0496	0.0500
Case 2	160	399	199	0.0488	0.0519	0.0527

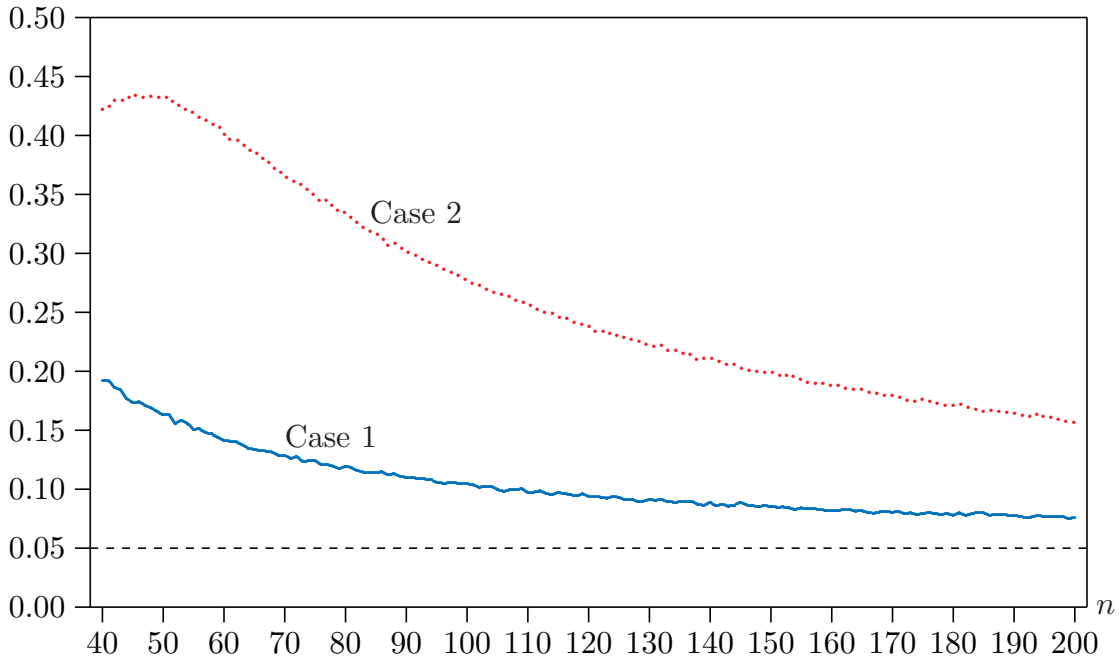

Figure 1. Asymptotic rejection frequencies at .05 level for probit OPG test

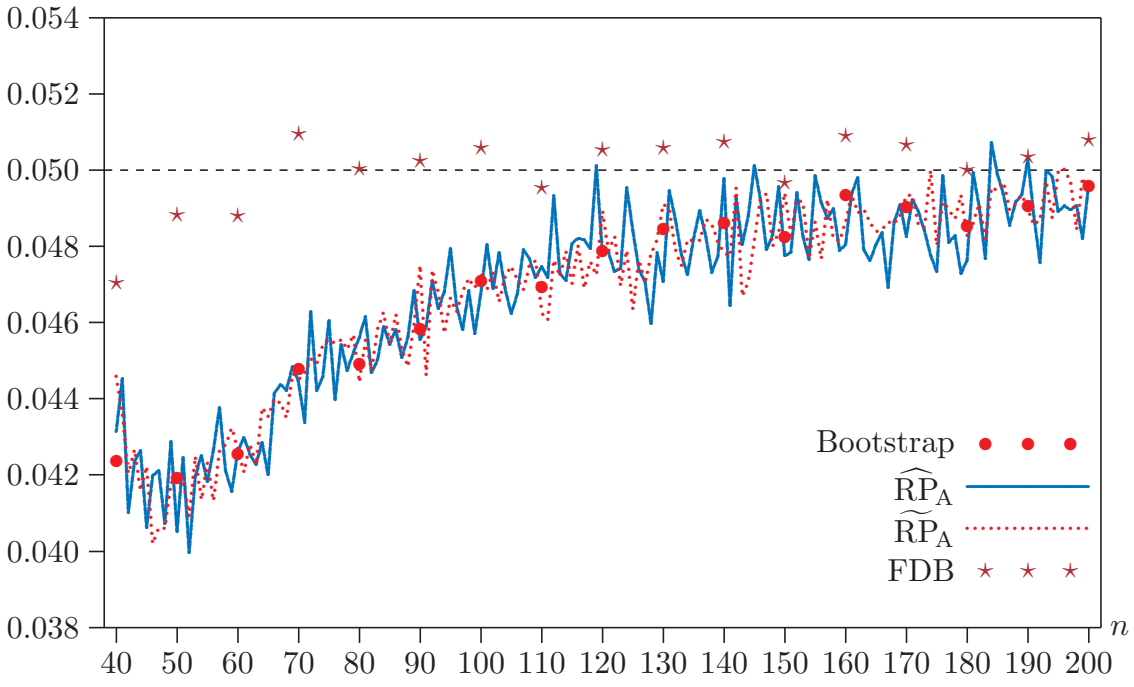

Figure 2. Bootstrap rejection frequencies at .05 level for probit OPG test, Case 1

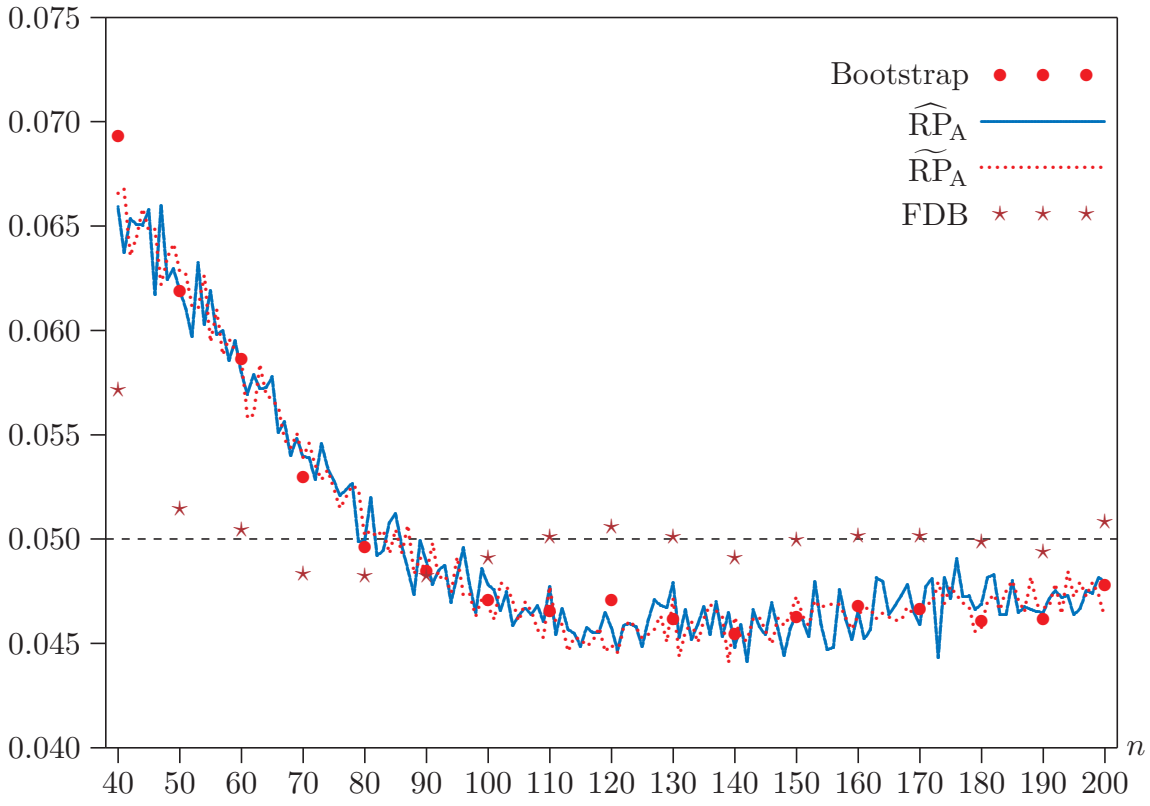

Figure 3. Bootstrap rejection frequencies at .05 level for probit OPG test, Case 2

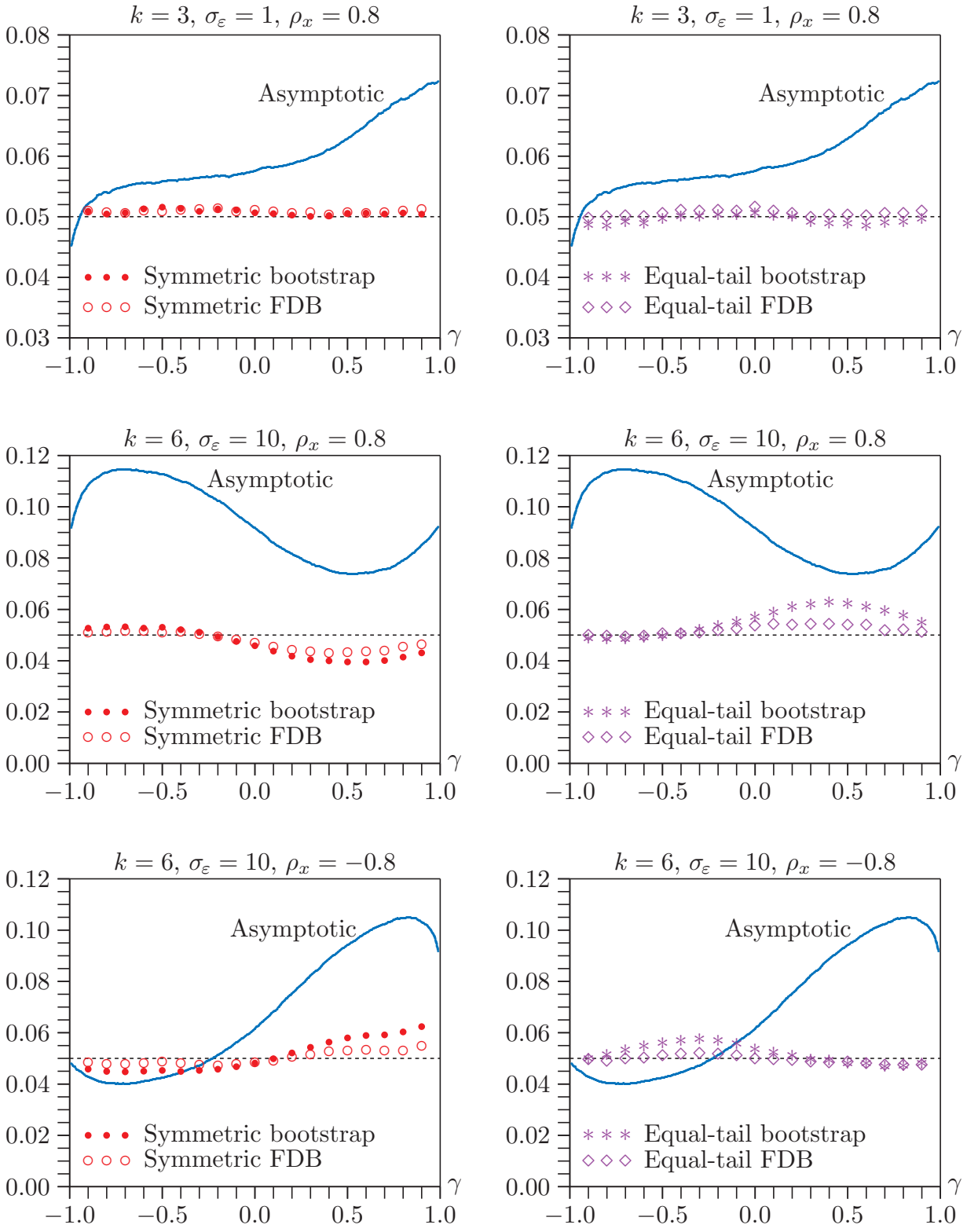

Figure 4. Durbin-Godfrey test rejection frequencies at .05 level under the null, $n = 20$

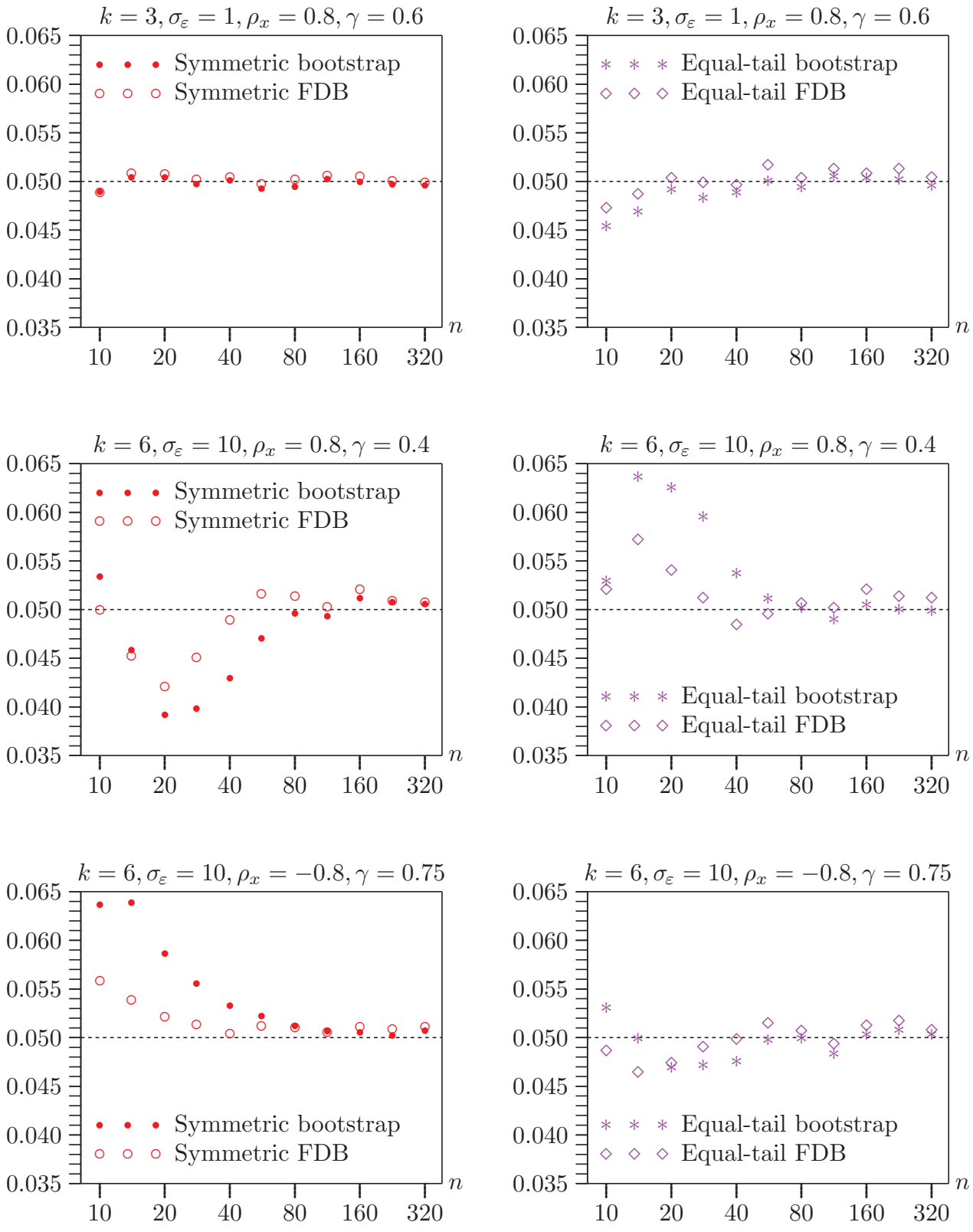

Figure 5. Durbin-Godfrey test rejection frequencies at .05 level under the null

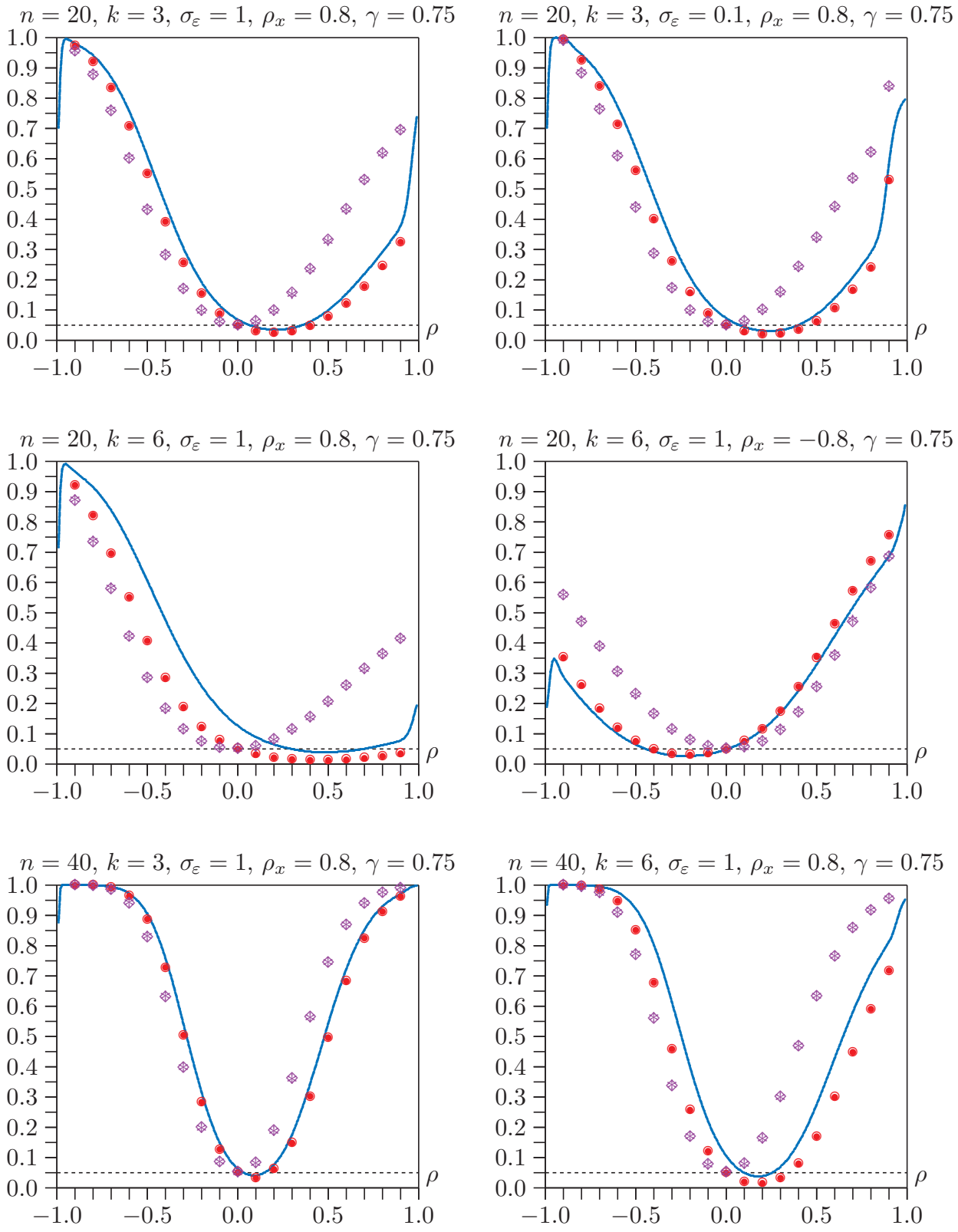

Figure 6. Power of Durbin-Godfrey tests at .05 level

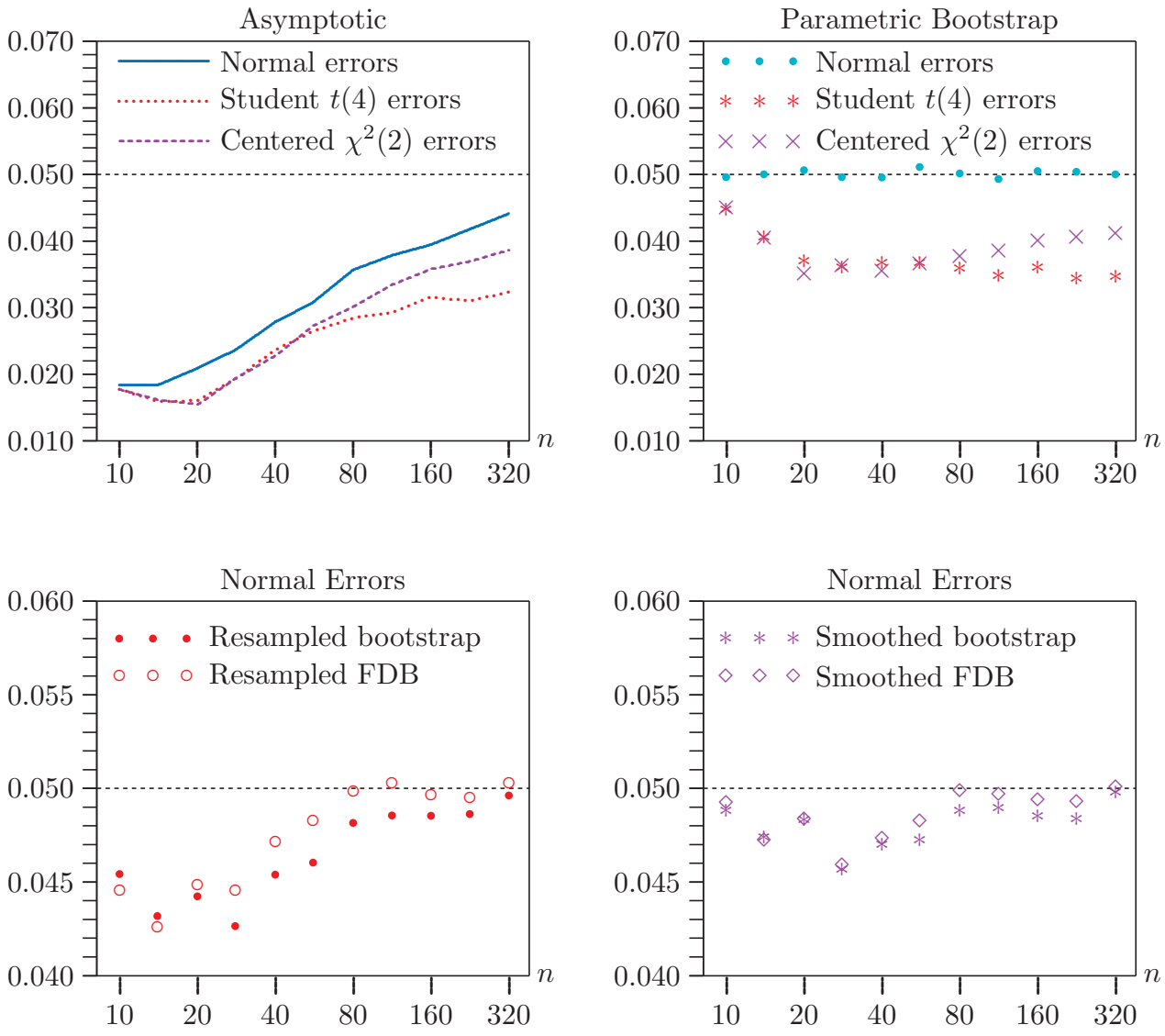

Figure 7. ARCH test rejection frequencies at .05 level under the null

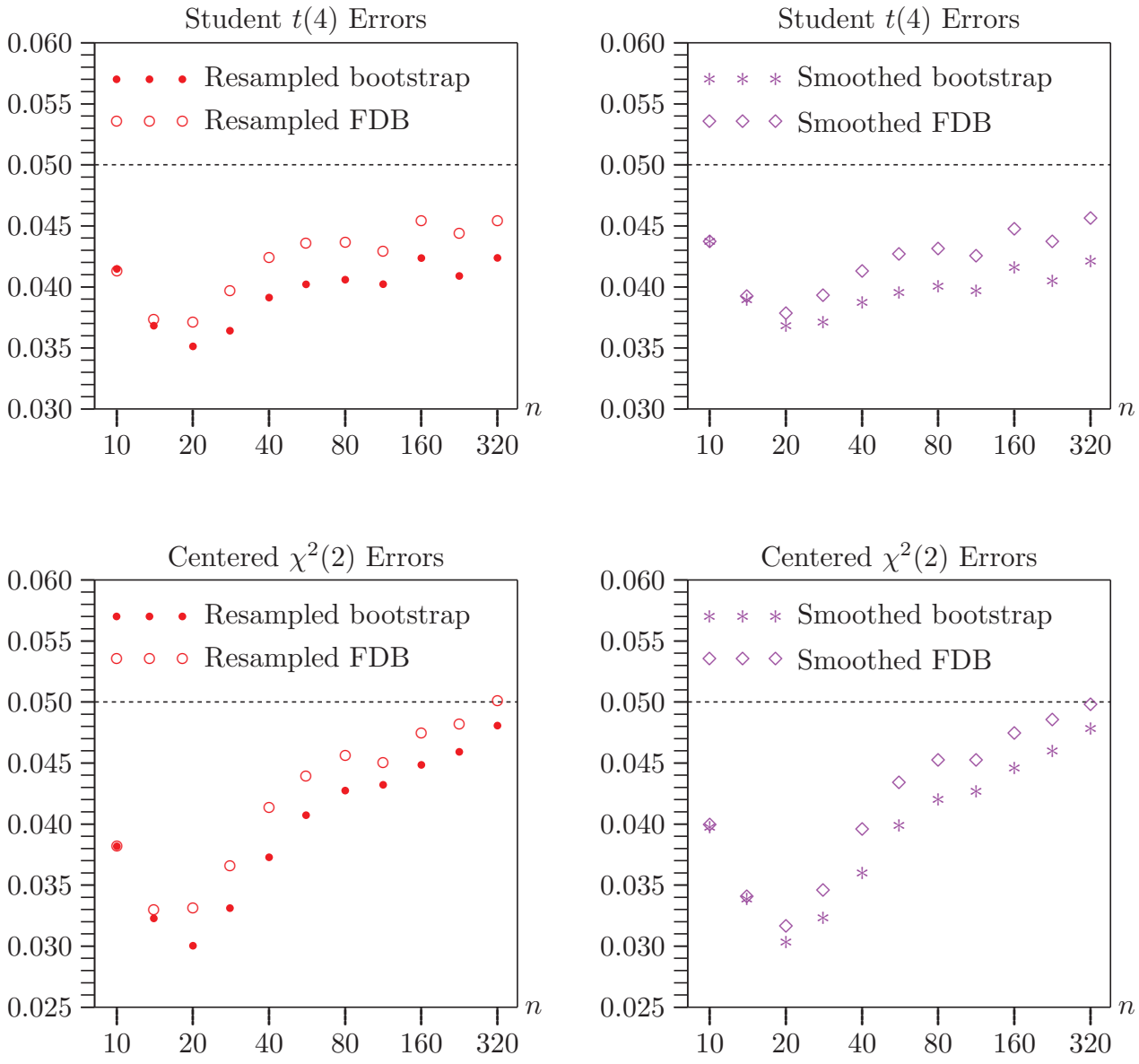

Figure 8. ARCH test rejection frequencies at .05 level under the null

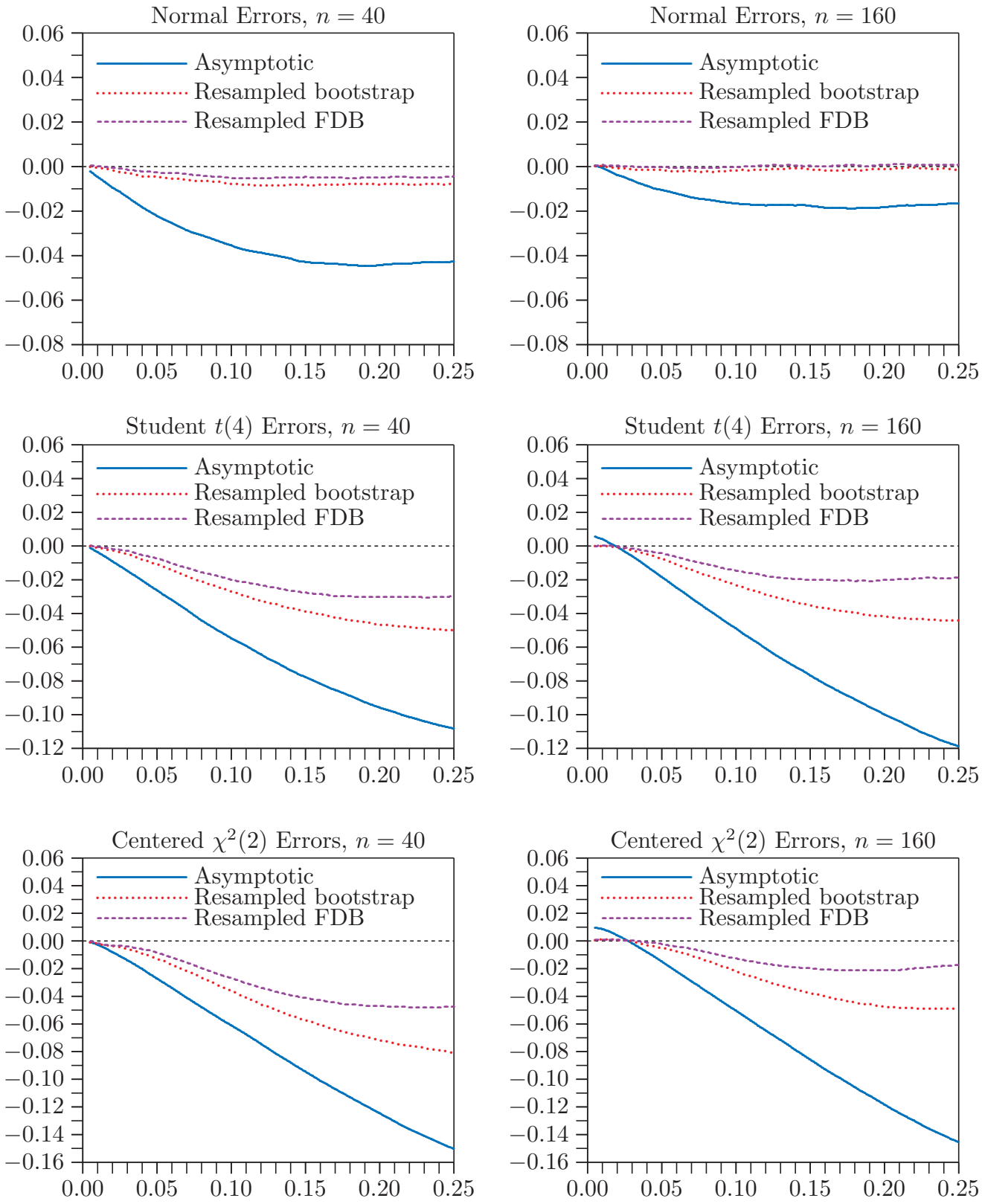

Figure 9. Rejection frequency discrepancy plots for ARCH tests

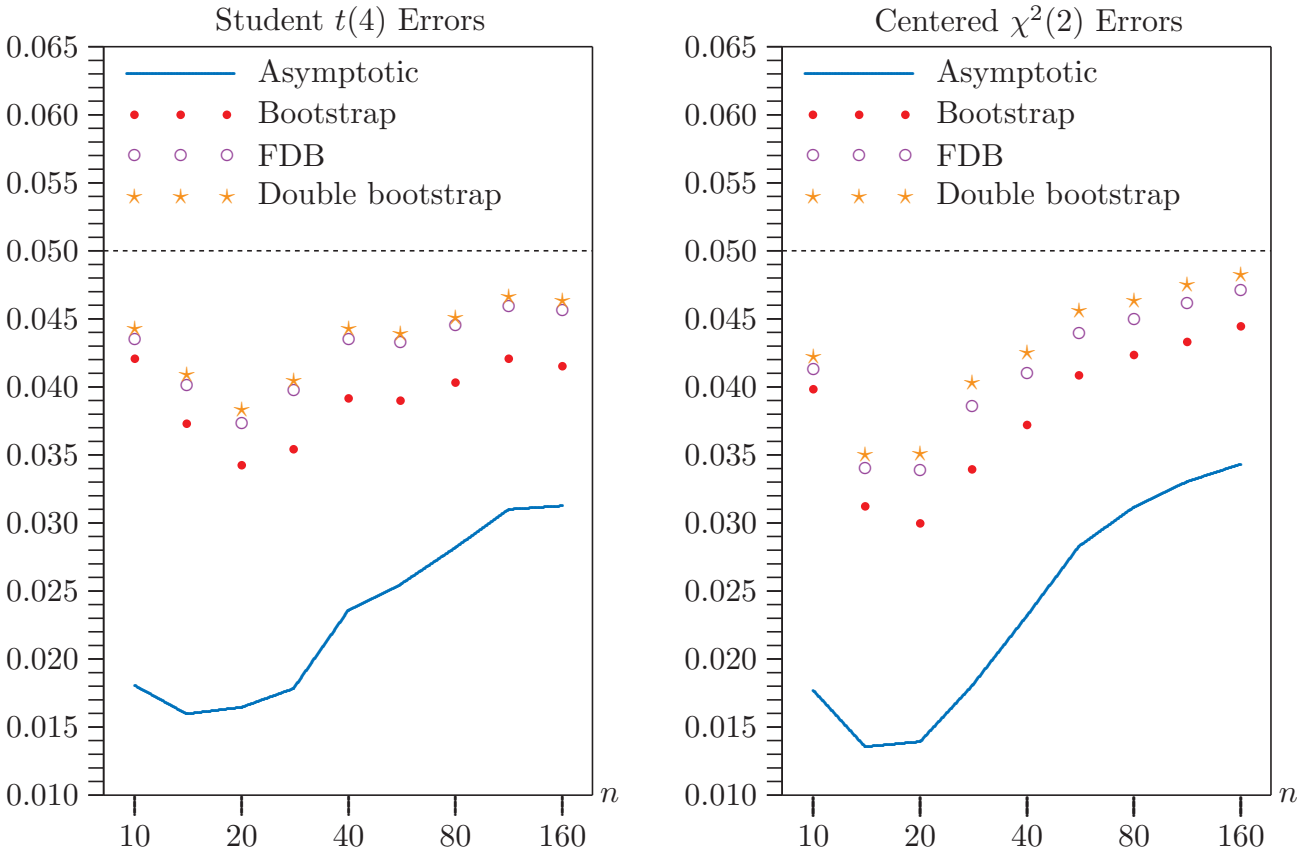

Figure 10. ARCH test rejection frequencies at .05 level under the null (resampled residuals)