Queen's Economics Department Working Paper No. 1023

Applications of the Fast Double Bootstrap

James MacKinnon
Queen's University

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

2-2006

Applications of the Fast Double Bootstrap

by

James G. MacKinnon

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

jgm@econ.queensu.ca

Abstract

The fast double bootstrap, or FDB, is a procedure for calculating bootstrap P values that is much more computationally efficient than the double bootstrap itself. In many cases, it can provide more accurate results than ordinary bootstrap tests. For the fast double bootstrap to be valid, the test statistic must be asymptotically independent of the random parts of the bootstrap data generating process. This paper presents simulation evidence on the performance of FDB tests in three cases of interest to econometricians. One of the cases involves both symmetric and equal-tail bootstrap tests, which, interestingly, can have quite different power properties. Another highlights the importance of imposing the null hypothesis on the bootstrap DGP.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. I am grateful to Russell Davidson, Silvia Gonçalves, and seminar participants at Cornell, Penn State, and the University of Rochester for comments.

December, 2005

1. Introduction

The simplest, and usually the most informative, way to perform a bootstrap test is to compute a bootstrap P value. We first compute the test statistic itself in the usual way. Then we generate a number of bootstrap samples and use each of them to compute a bootstrap statistic. Finally, we calculate a bootstrap P value as 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.

Bootstrap tests based on asymptotically pivotal test statistics should generally perform better in finite samples than tests based on asymptotic theory, in the sense that they will commit errors of lower order in the sample size n; see, among others, Hall (1992) and Davidson and MacKinnon (1999). However, this does not mean that bootstrap tests always perform acceptably well in finite samples. In theory, double bootstrap P values, which are discussed in Section 3, should lead to more reliable tests than ordinary bootstrap P values. However, the double bootstrap tends to be very computationally demanding. Davidson and MacKinnon (2001) therefore suggested what they called the **fast double bootstrap**, or **FDB**. It is very much less expensive to compute than the double bootstrap itself, but its validity requires stronger conditions.

There is, at present, only limited evidence on how useful the fast double bootstrap is likely to be in practice. Ideally, it would yield a P value similar to the ordinary bootstrap P value when the latter is reliable, and a P value more accurate than the ordinary bootstrap P value when the latter is unreliable. One objective of this paper is to see whether this is likely to be the case for several commonly encountered tests.

The next section discusses bootstrap tests, especially in the context of two-tailed tests, where there is more than one way to proceed. The following two sections discuss double bootstrap tests and fast double bootstrap tests, respectively. Section 5 presents simulation results for some regression-based tests for serial correlation. Section 6 presents similar results for tests for ARCH errors. Section 7 presents results for t tests in a model estimated by two-stage least squares. The simulation results suggest that using the fast double bootstrap can often improve the performance of bootstrap tests. The improvement is modest in many cases but quite substantial in some.

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; see Beran (1988). 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 $F(\tau)$ is the cumulative distribution function, or CDF, of τ.

The problem is that we often do not know the function $F(\tau)$. At one time, the only practical way to estimate $F(\hat{\tau})$ was generally to use an asymptotic distribution function. With modern computing facilities, however, we can often employ the bootstrap

instead. A bootstrap DGP is used to 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 is often called the **bootstrap distribution**. Then the bootstrap P value is

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

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

In some cases, the τ_j^* follow precisely the same distribution as τ. That is, $F^*(\tau)$, the limit of $\hat{F}_B^*(\tau)$ as $B \to \infty$, is identical to $F(\tau)$. In such a case, a bootstrap test is called a Monte Carlo test and, if B is chosen so that $\alpha(B+1)$ is an integer, it is exact; see Dufour and Khalaf (2001) for a review of the literature on Monte Carlo tests. In this paper, however, we will be concerned with the much more common situation in which $F^*(\tau)$ differs from $F(\tau)$ in finite samples.

When τ can take on either positive or negative values, for example, when 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**

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

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 has the form of a t statistic, so that τ^2 is asymptotically $\chi^2(1)$. Then the P value for $\hat{\tau}$ based on (2) must be identical to the P value for $\hat{\tau}^2$ based on (1), when both are calculated using the same set of τ_j^*. Rejecting when $\hat{p}_{\mathrm{S}}^*(\hat{\tau}) < \alpha$ is equivalent to rejecting when $|\hat{\tau}|$ is greater than the $1-\alpha$ quantile of the EDF of the $|\tau_j^*|$.

The symmetry assumption may often be excessively strong. If we do not wish to assume symmetry, we can instead base a test on the **equal-tail bootstrap *P* value**

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

Here we calculate P values for one-tailed tests in each tail and reject if either of these P values is less than $\alpha/2$. In other words, 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 factor of 2 is needed because it is twice as likely that $\hat{\tau}$ will be far out in either tail of the bootstrap distribution as that it will be far out in one specified tail. When τ has the form of a t statistic, the procedure is analogous to forming an equal-tail percentile t confidence interval. For a

Monte Carlo test based on the equal-tail P value (3) to be exact, it is required that $(\alpha/2)(B+1)$ be an integer.

As we will see below, the power of tests based on symmetric and equal-tail bootstrap P values against certain alternatives may be quite different. 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 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. Double Bootstrap Tests

The double bootstrap was proposed by Beran (1988). In various forms, it can be used for a variety of purposes, including the calculation of either confidence intervals or P values. In particular, it can be used to calculate P values that should, at least in theory, be more accurate than ordinary bootstrap P values.

Let $G(x)$ denote the CDF of the distribution that the bootstrap P values follow. If $F(\tau) = F^*(\tau)$, and $B = \infty$, this distribution would just be $\mathrm{U}(0,1)$. In general, however, it will be unknown. The idea of the double bootstrap is simply to estimate $G(x)$ by using a second level of bootstrap.

The first step in the double bootstrap is to generate B_1 first-level bootstrap samples that are used to compute bootstrap statistics τ_j^* for $j = 1, \ldots, B_1$. Then the ordinary bootstrap P value $\hat{p}^*(\hat{\tau})$ is calculated using one of the formulae discussed in the previous section. For concreteness, let us assume that (1) is used. The second step is to obtain a second-level bootstrap DGP for each first-level bootstrap sample indexed by j, in essentially the same way as the first-level bootstrap DGP was obtained from the actual sample. Each second-level bootstrap DGP is then used to generate B_2 bootstrap samples that are used to compute test statistics τ_{jl}^{**} for $l = 1, \ldots, B_2$.

For each first-level bootstrap sample indexed by j, we can compute the second-level bootstrap P value

$$\hat{p}_j^{**} = \frac{1}{B_2} \sum_{l=1}^{B_2} I(\tau_{jl}^{**} > \tau_j^*). \tag{4}$$

This is the P value for the bootstrap statistic τ_j^* based on the EDF of the τ_{jl}^{**}. We then use the $\hat{p}_j^{**}$ to calculate the **double-bootstrap P value** as

$$\hat{p}^{**}(\hat{\tau}) = \hat{G}^*_{B_2}\big(\hat{p}^*(\hat{\tau})\big) = \frac{1}{B_1} \sum_{j=1}^{B_1} I\big(\hat{p}_j^{**} \leq \hat{p}^*(\hat{\tau})\big). \tag{5}$$

Thus $\hat{p}^{**}$ is equal to the proportion of the second-level bootstrap P values that are smaller (and hence more extreme) than the first-level bootstrap P value. The inequality in (5) is not strict, because, depending on the values of B_1 and B_2, there may well be cases for which $\hat{p}_j^{**} = \hat{p}^*(\hat{\tau})$.

Comparing (5) and (1), we can see that, just as $\hat{F}^*_B(\hat{\tau})$ is used to estimate $F(\hat{\tau})$, so $\hat{G}^*_{B_2}(\hat{p}^*)$ is used to estimate $G(\hat{p}^*)$. If $\hat{\tau}$, the τ^*_j, and the τ^{**}_{jl} all came from the same distribution, then the $\hat{p}^{**}_j$ should be uniformly distributed on the zero-one interval, and, as $B_1 \to \infty$ and $B_2 \to \infty$, we should find that $\hat{p}^{**}(\hat{\tau}) = \hat{p}^*(\hat{\tau})$. In this case, the double bootstrap yields exactly the same inferences as the ordinary 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 $\hat{\tau}$ must be 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.

Of course, we cannot expect even the double bootstrap to work perfectly. Just as $F^*(\tau)$ may provide an inadequate approximation to $F(\tau)$, so may $G^*(x)$, the limit of $\hat{G}^*_{B_2}(x)$, provide an inadequate approximation to $G(x)$. In principle, any bootstrap procedure, including the double bootstrap, may or may not provide an acceptable approximation to the true P value associated with $\hat{\tau}$.

In practice, the most serious problem with the double bootstrap is that it is very costly in terms of computation. 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 $1 + B_1 + B_1 B_2$. For example, if $B_1 = 999$ and $B_2 = 399$, the double bootstrap will require the calculation of no fewer than 399,601 test statistics.

4. Fast Double Bootstrap Tests

The double bootstrap is costly because we need to generate B_2 second-level bootstrap samples for every first-level bootstrap sample. This is necessary because the distribution of the τ^{**}_{jl} may not be independent of τ^*_j. If we make the assumption that this distribution is independent of τ^*_j, we can dramatically reduce the cost of the procedure. This is the key assumption of the FDB procedure.

For the fast double bootstrap test, only one second-level bootstrap statistic, τ^{**}_j, is calculated along with each τ^*_j. Let $\hat{Q}^{**}_B(1 - \hat{p}^*)$ denote the $1 - \hat{p}^*$ quantile of the τ^{**}_j. This quantile is defined implicitly by the equation

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

Of course, for finite B, there will be a range of values of Q^{**}_B that satisfy (6), and we will need to choose one of them in a somewhat arbitrary manner. 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^{**}. However, it is quite possible that different methods of estimating quantiles may affect the performance of FDB tests when B is not large.

The **FDB *P* value** is easily calculated from $\hat{Q}_B^{**}(1-\hat{p}^*)$ and the τ_j^*. It is just

$$\hat{p}_F^{**}(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I\big(\tau_j^* > \hat{Q}_B^{**}(1-\hat{p}^*)\big). \tag{7}$$

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

When the distribution of τ_{jl}^{**} does not depend on τ_j^*, there is a very close relationship between $\hat{p}_F^{**}(\hat{\tau})$ defined in equation (7) and $\hat{p}^{**}(\hat{\tau})$ defined in equation (5). Let $Q^{**}(x)$ denote the limit of $\hat{Q}_B^{**}(x)$ as $B \to \infty$, and let $p^* = \Pr(\tau_j^* > \hat{\tau})$. Then, in the limit, the FDB P value is simply

$$p_F^{**}(\hat{\tau}) = \Pr\big(\tau_j^* > Q^{**}(1-p^*)\big). \tag{8}$$

The probability in (8), which is conditional on the original sample, is being taken with respect to the distribution of the τ_j^*. In contrast, from (4) and (5), the limit of the double bootstrap P value as $B_1 \to \infty$ and $B_2 \to \infty$ is

$$\begin{aligned} p^{**}(\hat{\tau}) &= \Pr\big(\Pr(\tau_{jl}^{**} > \tau_j^*) < p^*(\hat{\tau})\big) \\ &= \Pr\Big(\Pr(\tau_{jl}^{**} > \tau_j^*) < \Pr\big(\tau_{jl}^{**} > Q^{**}(1-p^*)\big)\Big), \end{aligned} \tag{9}$$

where the equality in the second line here uses the the definition of $Q^{**}(1-p^*)$ that is analogous to (6). The outer probability in the expression in the second line of (9) is being taken with respect to the distribution of the τ_j^*, and the inner ones with respect to the distribution of the τ_{jl}^{**}. Since these two distributions are assumed to be independent, it is clear that

$$\Pr(\tau_{jl}^{**} > \tau_j^*) < \Pr\big(\tau_{jl}^{**} > Q^{**}(1-p^*)\big)$$

if and only if

$$\tau_j^* > Q^{**}(1-p^*).$$

But, as we can see from (8), the probability of this event is precisely what $p_F^{**}(\hat{\tau})$ is equal to. Thus we see that, in the limit when B, B_1, and B_2 are infinite, the FDB P value (7) must be equivalent to the double bootstrap P value (5) whenever the distribution of the τ_{jl}^{**} does not depend on τ_j^*.

If we wished to reject when the test statistic was in the lower tail of the distribution, we would replace equation (7) by

$$\hat{p}_F^{**}(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I\big(\tau_j^* < \hat{Q}_B^{**}(1-\hat{p}^*)\big). \tag{10}$$

Note that $\hat{p}^*$ is still defined as in (1) here, so that, in contrast to the usual case, large values of $\hat{p}^*$ would lead to rejection by the ordinary bootstrap test. To obtain an equal-tail FDB P value, we could compute both (7) and (10) and then take twice the minimum of these two FDB P values, as in equation (3).

In most cases of interest to econometricians, the distribution of a test statistic is asymptotically independent of the distribution of the parameter estimates under the null hypothesis, and this also applies to many nonparametric and semiparametric bootstrap DGPs. Whether or not the fast double bootstrap will perform well depends, in part, on how much this asymptotic independence carries over to finite samples. Of course, the independence assumption merely guarantees that the fast double bootstrap is equivalent to the double bootstrap when B, B_1, and B_2 are large. It does not guarantee that the latter will yield accurate inferences in finite samples. In fact, as we will see in Section 6, double bootstrap P values can be somewhat inaccurate.

Davidson and MacKinnon (2001, 2002) studied several cases in which the FDB works well. In the case of the "J test" for nonnested linear regression models (Davidson and MacKinnon, 1981), which is treated as a one-tailed test, the FDB yields substantially more accurate results than ordinary bootstrap tests in cases where the latter are somewhat inaccurate. There have also been simulations by Lamarche (2004) in the context of testing for unknown structural breaks, by Omtzigt and Fachin (2002) in the context of cointegrated VARs, and by Davidson (2006) in the context of testing cointegration for fractionally integrated processes.

It is of interest to see how well FDB tests work under ideal conditions, when τ, the τ_j^*, and the τ_j^{**} all come from the same distribution. To investigate this question, I generated all three statistics from the standard normal distribution for various values of B between 99 and 3999 and then calculated ordinary and FDB bootstrap P values. Ordinary bootstrap tests always rejected just about 5% of the time at the .05 level and just about 1% of the time at the .01 level. So did the FDB tests, but only when B was sufficiently large. There was a very noticeable tendency for the FDB tests to reject too often when B was not large. The overrejection was much more severe for equal-tail FDB tests than for symmetric or one-tailed tests.

To quantify this tendency, I regressed the difference between the FDB rejection frequency and the ordinary bootstrap rejection frequency 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 were 54 experiments for the one-tailed and symmetric tests and 47 experiments for the equal-tail tests. There were fewer experiments for the latter because values of B for which $\alpha(B+1)$ was an integer but

Table 1. Regressions for overrejection when B is small

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

$(\alpha/2)(B+1)$ was not (namely, 99 and 299) had to be removed. Since each experiment used 1 million replications, 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 that quantiles are estimated. 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 (2000) for a discussion of how to choose B sequentially when it is expensive to calculate bootstrap test statistics.

5. Tests for Serial Correlation

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

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \gamma y_{t-1} + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2), \tag{11}$$

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 was proposed by 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 error terms, we first estimate the regression in (11) 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

$$y_t^* = \boldsymbol{X}_t\hat{\boldsymbol{\beta}} + \hat{\gamma} y_{t-1}^* + u_t^*, \tag{12}$$

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 (12) imposes the IID assumption on the error terms without imposing any additional distributional assumptions.

A large number of experiments on the finite-sample properties of bootstrap versions of the Durbin-Godfrey test were performed. In all the reported experiments, the error terms were normally distributed, the first column of the $\boldsymbol{X}$ matrix was a constant, and the remaining columns were generated by independent, stationary AR(1) processes with parameter ρ_x. A new $\boldsymbol{X}$ matrix was drawn for each replication. Both asymptotic and bootstrap rejection frequencies were 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 improved rapidly as n increased, I used $n = 20$ 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 was 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 1 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 was 0.999 for every value of γ. Thus an investigator who performed both tests would obtain extremely similar results and would probably conclude, quite justifiably, that the bootstrap P value was 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. 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 ordinary bootstrap tests, especially when the latter are least reliable. Nevertheless, the correlations between the ordinary 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 performance of the ordinary bootstrap and FDB tests improves as the sample size increases. Figure 2 contains six panels, which are comparable to the six panels in Figure 1. 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 were 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 ordinary bootstrap tests.

The next two rows of the figure are more interesting. We see both noticeable overrejection and noticeable underrejection by the ordinary bootstrap tests. With a few exceptions, the FDB tests perform substantially better than the ordinary 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 that we saw in Table 1. Since the magnitude of the overrejection is just about what we would expect from the results in Figure 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 3 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 ordinary 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 1 and 2.

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 shapes of the power functions 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 the figure, $n = 40$. 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 $\rho = 0$.

Notice that, in several panels of Figure 3, the asymptotic tests are reasonably reliable under the null. Nevertheless, there are very substantial gains in power to be had by 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.

6. Tests for ARCH Errors

Since the seminal work of Engle (1982), it has been recognized that serial dependence in the variance of the error terms of regression models using times-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 errors, 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 error terms. 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 errors.

Consider the linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t = \sigma_t\varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \delta_1\sigma_{t-1}^2, \quad \varepsilon_t \sim \text{IID}(0,1). \tag{13}$$

The error terms 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 error terms are IID in the model (13) is to run the regression

$$\hat{u}_t^2 = b_0 + b_1\hat{u}_{t-1}^2 + \text{residual}, \tag{14}$$

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 (14), 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 (14). These include the ordinary t statistic for $b_1 = 0$, which is asymptotically distributed as $\text{N}(0,1)$, 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 errors, 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 4 through 7 report results from a number of simulation experiments which focused on the effects of the sample size and the distribution of the ε_t. In all the reported experiments, $\boldsymbol{X}_t$ consisted of a constant and two independent, standard normal random variates, since changing the number of regressors had only a modest effect on the finite-sample behavior of the tests. The sample size took on the values 10, 14, 20, 28, 40, 56, 80, 113, 160, 226, and 320. The error terms were either standard normal, Student's t with 4 degrees of freedom, or rescaled and recentered $\chi^2(2)$. The first of these distributions is in some sense the base case, the second involves severe leptokurtosis, and the third involves severe skewness. Since the test statistics can easily be shown to be invariant to the variance of the error terms under the null hypothesis, that aspect of the experimental design was not varied.

The top left-hand panel of Figure 4 shows rejection frequencies for the asymptotic test as a function of the sample size for the three different error distributions. These results are based on 100,000 replications for each value of n. The test underrejects

severely in all cases, especially when the error terms are nonnormal. As we would expect, performance improves with the sample size, but the rate of improvement is fairly slow, especially when the errors are $t(4)$.

There are several ways to bootstrap this test. One possibility is to use a parametric bootstrap, drawing the error terms from the normal distribution. It is easy to see that this will lead to an exact test when the errors 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 known, then the test statistic does not depend on any unknown features of the DGP. It then follows by standard arguments for Monte Carlo tests that, when 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 4 shows rejection frequencies for parametric bootstrap tests, with $B = 1999$. As expected, these tests work perfectly when the errors 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 error terms 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 (12). Results for this procedure are shown in the bottom left-hand panel of Figure 4 and in the two left-hand panels of Figure 5. 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 ordinary semiparametric bootstrap test always underrejects to some extent.

It is more interesting to see what happens when the error terms 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 error terms are centered $\chi^2(2)$, 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 error terms. One way to improve this approximation is to

smooth the bootstrap errors. 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

$$\hat{F}_h(u') = \frac{1}{n} \sum_{t=1}^{n} K(\hat{u}_t, u', h), \tag{15}$$

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 errors from (15), we simply resample from the residuals $\hat{u}_t$ and then add independent normal random variables with variance h^2. The resulting bootstrap errors have mean 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 error terms.

The bottom right-hand panel of Figure 4 and the two right-hand panels of Figure 5 show the effects of using bootstrap errors that were smoothed in this way, where the Gaussian kernel with the bandwidth given above was used. When the error terms 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 errors closer to normality. Interestingly, there appears to be no appreciable gain from smoothing when the errors are $t(4)$ or centered $\chi^2(2)$. As in the case where smoothing is not used, the FDB rejection frequencies are always noticeably closer to 0.05 than those of the ordinary bootstrap, except when the sample size is very small.

Because Figures 4 and 5 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 6 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 errors, 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 ordinary 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 7 provides some evidence on this point. The experiments were similar to those in the left-hand panels of Figure 5, except that values of n greater than 160 were omitted. They involve semiparametric bootstrap DGPs that use resampled residuals,

with errors that are either $t(4)$ or centered $\chi^2(2)$. There were 100,000 replications. However, because computational cost was an issue, B_1 and B_2 were quite small at 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 7, 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 ordinary 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 $\chi^2(2)$ errors than for the ones with $t(4)$ errors, 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.

7. Two Stage Least Squares *t* Statistics

There is a large literature on the finite-sample properties of 2SLS estimates, and it is well-known that inferences based on asymptotic theory can be poor when the instruments are weak. See Bound, Jaeger, and Baker (1995), Staiger and Stock (1997), and Stock, Wright, and Yogo (2002), among many others.

Various procedures for bootstrapping 2SLS estimates have been proposed, starting with Freedman (1984), and it is natural to conjecture that using the bootstrap may solve the problem. However, although published work on the finite-sample properties of 2SLS bootstrap procedures is scarce, it appears to be well-known that bootstrapping 2SLS often does not work very well. In this section, we investigate whether the FDB can offer an improvement.

The 2SLS estimator applies to one structural equation from a linear simultaneous equations model. The structural equation can be written as

$$\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{Y}_1\boldsymbol{\gamma} + \boldsymbol{u} = \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{u}, \tag{16}$$

where $\boldsymbol{X}_1$ is an $n \times k_1$ matrix of observations on exogenous or predetermined variables, $\boldsymbol{Y}_1$ is $n \times g_1$ matrix of observations on endogenous variables, $\boldsymbol{Z} \equiv [\boldsymbol{X}_1 \;\; \boldsymbol{Y}_1]$, and $\boldsymbol{\delta}^\top = [\boldsymbol{\beta}^\top \;\; \boldsymbol{\gamma}^\top]^\top$. The rest of the system is treated as an unrestricted reduced form,

$$\boldsymbol{Y}_1 = \boldsymbol{X}\boldsymbol{\Pi} + \boldsymbol{V}, \tag{17}$$

where there are $k = k_1 + k_2$ instruments in the matrix $\boldsymbol{X} \equiv [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$. The error terms have mean zero conditional on $\boldsymbol{X}$, and they are assumed to be IID. For every observation,

$$\mathrm{E}\left(\begin{bmatrix} u_t \\ \boldsymbol{v}_t \end{bmatrix} [\, u_t \quad \boldsymbol{v}_t^\top] \right) = \begin{bmatrix} \sigma^2 & \boldsymbol{\omega}^\top \\ \boldsymbol{\omega} & \boldsymbol{\Omega} \end{bmatrix} \equiv \boldsymbol{\Sigma}, \tag{18}$$

where $\boldsymbol{v}_t$ is the t^{th} row of $\boldsymbol{V}$ rewritten as a column vector. The matrix $\boldsymbol{\Sigma}$ is $g \times g$, where $g = g_1 + 1$.

The **2SLS**, or **generalized IV**, estimator is

$$\hat{\boldsymbol{\delta}}^{\mathrm{IV}} = (\boldsymbol{Z}^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y}_1, \tag{19}$$

where $\boldsymbol{P}_{\boldsymbol{X}} \equiv \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ is an idempotent projection matrix that projects on to $\mathcal{S}(\boldsymbol{X})$, the subspace spanned by the instruments. The estimator (19) can also be written as

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{IV}} \\ \hat{\boldsymbol{\gamma}}^{\mathrm{IV}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{Y}_1 \\ \boldsymbol{Y}_1^\top \boldsymbol{X}_1 & \boldsymbol{Y}_1^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{Y}_1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{y}_1 \\ \boldsymbol{Y}_1^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{y}_1 \end{bmatrix}.$$

The usual covariance matrix estimate is

$$\hat{\sigma}^2_{\mathrm{IV}} (\boldsymbol{Z}^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{Z})^{-1}, \tag{20}$$

where

$$\hat{\sigma}^2_{\mathrm{IV}} \equiv \frac{1}{n} \| \boldsymbol{y}_1 - \boldsymbol{Z} \hat{\boldsymbol{\delta}}^{\mathrm{IV}} \|^2.$$

In principle, bootstrapping can be used both to obtain more accurate estimates and to obtain more reliable inferences. Since this paper is concerned with testing, the experiments concern the latter. They focus on the t statistic for the first element of $\boldsymbol{\gamma}$ to equal its true value γ_{10}. This is just

$$t_\gamma = \frac{\hat{\gamma}_1^{\mathrm{IV}} - \gamma_{10}}{s(\hat{\gamma}_1^{\mathrm{IV}})}, \tag{21}$$

where $s(\hat{\gamma}_1^{\mathrm{IV}})$ is the square root of the appropriate diagonal element of the matrix (20). In the experiments, $g_1 = 1$, so that γ_1 is the only coefficient on an endogenous variable in the structural equation.

There are numerous ways to bootstrap the t statistic (21) in the model given by (16), (17), and (18). The principal procedure that I investigate works as follows:

1. Estimate the reduced form equation(s) (17) by OLS, retaining the parameter estimates $\hat{\boldsymbol{\Pi}}$ and the residuals $\hat{\boldsymbol{V}}$. Then use 2SLS to obtain t_γ.
2. Estimate the structural equation (16) by 2SLS (or OLS if appropriate) under the restriction that γ_1 equals its true value γ_{10}. This yields parameter estimates $\tilde{\boldsymbol{\delta}} = [\tilde{\boldsymbol{\beta}} \vdots \tilde{\boldsymbol{\gamma}}]$ and residuals $\tilde{\boldsymbol{u}}$.
3. Create the $n \times (g_1 + 1)$ matrix

$$\acute{\boldsymbol{U}} \equiv \begin{bmatrix} \tilde{\boldsymbol{u}} & \acute{\boldsymbol{V}} \end{bmatrix}, \quad \text{where} \quad \acute{\boldsymbol{V}} = \left(\frac{n}{n-k} \right)^{1/2} \hat{\boldsymbol{V}}.$$

4. Obtain n row vectors of bootstrap errors $\boldsymbol{U}_t^*$ by resampling from the rows of $\acute{\boldsymbol{U}}$. The bootstrap errors are then $\boldsymbol{U}^* = [\boldsymbol{u}^* \;\; \boldsymbol{V}^*]$.

5. For each bootstrap sample, generate $\boldsymbol{Y}_1^*$ as $\boldsymbol{X}\hat{\boldsymbol{\Pi}} + \boldsymbol{V}^*$ and then generate $\boldsymbol{y}_1^*$ as $\boldsymbol{X}_1\tilde{\boldsymbol{\beta}} + \boldsymbol{Y}_1^*\tilde{\boldsymbol{\gamma}} + \boldsymbol{u}^*$.
6. For each bootstrap sample, estimate the model by 2SLS and calculate the bootstrap test statistic

$$t_j^* = \frac{\hat{\gamma}_1^* - \gamma_{10}}{s(\hat{\gamma}_1^*)}. \tag{22}$$

This procedure imposes virtually no assumptions on the functional form of the joint distribution of the error terms, but it does assume that they are IID and independent of the instruments. In this respect, it is more restrictive than the procedure proposed by Freedman (1984), a variation of the pairs bootstrap, which allows for heteroskedasticity of unknown form. Moreover, it imposes the null hypothesis on the bootstrap samples, since these depend on estimates of the structural equation subject to the restriction that $\gamma_1 = \gamma_{10}$.

Especially when one wants to bootstrap more than one test statistic, a good deal of computation can be avoided by not imposing the null hypothesis; this avoids step 2 and the need to repeat all subsequent steps for each hypothesis to be tested. However, not imposing the null requires that the bootstrap test statistic be calculated as

$$t_j^* = \frac{\hat{\gamma}_1^* - \hat{\gamma}_1}{s(\hat{\gamma}_1^*)}. \tag{23}$$

Unlike the bootstrap t statistic (22), which tests a null hypothesis that is not true in the bootstrap DGP when $\gamma_1 \neq \gamma_{10}$, (23) tests a null that is always true in the bootstrap DGP. The bootstrap tests based on restricted estimates and unrestricted estimates are referred to as bootstrap-R and bootstrap-U, respectively. Theory suggests that the former should work better; see Davidson and MacKinnon (1999).

In the simulation experiments, there are just two endogenous variables, so that $g_1 = 1$. The only exogenous variable in the structural equation is a constant, so that $k_1 = 1$. There are k_2 instruments, which are IID standard normal, and the number of overidentifying restrictions is $r = k_2 - 1 = k - 2$. All experiments had 100,000 replications and 999 bootstraps. Bootstrap P values were based on the symmetric bootstrap P value (2), because equal-tail bootstrap P values generally worked poorly.

The key parameters in the experiments are n, the sample size, r, the number of overidentifying restrictions, ρ, the correlation between the error terms of the structural and reduced form equations, and R^2_∞, the asymptotic R^2 for the reduced form equation. The quantity nR^2_∞ measures the strength of the instruments. Note that, if there were exogenous variables other than a constant in the structural equation, nR^2_∞ would no longer be an appropriate measure of instrument strength. Instead, we would need a measure of the explanatory power of $\boldsymbol{X}_2$ when it is added to the reduced form equation.

Certain parameters were not varied. In particular, σ^2 and γ_{10} were not varied because, under the null, t_γ is invariant to them. Moreover, because the variance of each column of $\boldsymbol{X}$, except the first, is 1, the variance of y_{t2} is just

$$\sigma_2^2 = \sum_{j=2}^{k} \pi_j^2 + \sigma_v^2,$$

where π_j is the coefficient on the j^{th} exogenous variable in the reduced form equation for y_2. This value was set to 1 in all the experiments. This implies that each individual instrument becomes weaker as the number of instruments is increased.

Figure 8 shows rejection frequencies for tests at the .05 level as a function of r for four sample sizes: 25, 50, 100, and 200. The values of ρ and R_∞^2 were 0.9 and 0.2, respectively. Thus the values of nR_∞^2 in the four panels of the figure are 5, 10, 20, and 40. As will be apparent from subsequent figures, $\rho = 0.9$ is a value for which both the asymptotic and bootstrap tests perform quite poorly. It is evident from Figure 8 that overrejection increases sharply with the number of overidentifying restrictions. This is true for all five tests. Both bootstrap tests always overreject less severely than the asymptotic test, but the bootstrap-R test performs much better than the bootstrap-U test, especially for smaller sample sizes. Using the FDB generally provides a modest further improvement. As the sample size increases, all the tests improve, but the bootstrap tests improve more rapidly than the asymptotic one, as theory suggests. The FDB does not improve any faster than the ordinary bootstrap, however, and it offers only a modest improvement for $n = 200$.

Figure 9 shows rejection frequencies for tests at the .05 level as a function of R_∞^2 for the same four sample sizes. In these experiments, $\rho = 0.9$ and $r = 7$, so the performance of all tests is relatively poor. For all sample sizes, the asymptotic test overrejects very severely when R_∞^2 is small. The speed at which this overrejection diminishes as R_∞^2 increases depends on the sample size. The bootstrap-R test always greatly outperforms the asymptotic test, and the FDB provides a further gain which is most noticeable for smaller sample sizes and smaller values of R_∞^2. Once again, the gain from using the R instead of the U variant of the bootstrap is generally greater than any further gain from using the FDB.

Figure 10 shows rejection frequencies as a function of ρ for the same four sample sizes. In these experiments, $R_\infty^2 = 0.2$ and $r = 7$. The role of ρ is evidently very important. When it is small, all the tests actually underreject. For small sample sizes, the bootstrap-U test underrejects more severely than the bootstrap-R test, which in turn underrejects more severely than the asymptotic test. The fact that both bootstrap and asymptotic tests can underreject when the instruments are very weak is interesting.

For the bootstrap-R procedure, using the FDB generally improves matters. It leads to less severe underrejection for small values of ρ and to less severe overrejection for large values. The gains are quite modest, however, especially when $n = 200$. For bootstrap-U, the benefits of using the FDB procedure are much less clear, however. For values of ρ that are neither very large nor very small, the FDB variant actually overrejects

more severely than the ordinary bootstrap. In some cases, it even overrejects more severely than the asymptotic test.

Figure 11 shows the relationship between rejection frequencies and sample size in two different ways. In all panels, $\rho = 0.9$. In the two left-hand panels, R^2_∞ is held constant (at either 0.1 or 0.2) as n is increased from 20 to 1280 by factors of $\sqrt{2}$. In the two right-hand panels, nR^2_∞ is held constant (at either 10 or 20) for the same values of n (except that $n = 20$ has to be omitted when $nR^2_\infty = 20$). The left-hand panels show that the bootstrap tests improve more rapidly with the sample size than the asymptotic test and that the further gain from using the FDB is moderate for small values of n and negligible for large ones. The right-hand panels show that, in accordance with theory, it is nR^2_∞ and not simply n that must tend to ∞ if asymptotic results are to hold. As $n \to \infty$, all the tests converge to rejection frequencies that are very much greater than .05.

Figure 12 is similar to Figure 11, but ρ is either 0.5 (in the top two panels) or zero (in the bottom two panels). The bootstrap tests based on restricted estimates perform very well, with the FDB procedure being particularly valuable when $\rho = 0$. However, the ones based on unrestricted estimates perform poorly, and FDB-U is actually worse than bootstrap-U when $\rho = 0.5$. Moreover, the asymptotic tests outperform the bootstrap-U tests for large n in both the right-hand panels. Failing to impose the restrictions of the null hypothesis on the bootstrap DGP can evidently have very severe consequences when the instruments are weak.

The convergence results that are evident in the right-hand panels of Figures 11 and 12 may be of interest to applied workers who have very large samples with very weak instruments and who wish to perform simulation experiments to evaluate the performance of IV test statistics. Instead of using a DGP with, say, 50,000 observations and $R^2_\infty = .0004$, one could probably get almost the same results by using a DGP with 1000 observations and $R^2_\infty = .02$. It might even be possible to use a similar trick with the bootstrap, although this would require some care in specifying the bootstrap DGP.

The mediocre performance of all the tests in many of the experiments reported here confirms the conventional wisdom that t tests in models estimated by 2SLS often have very poor finite-sample properties, even if they are bootstrapped. Using the FDB generally improves matters, but not by very much. Instead of trying to fix a test that is fundamentally flawed, it would make more sense to start with a test that has better finite-sample properties, such as the likelihood ratio test or the test proposed by Kleibergen (2002). See Andrews, Moreira, and Stock (2004).

8. Conclusion

This paper has studied the performance of fast double bootstrap tests for three cases of interest in applied econometrics, namely, tests for serial correlation, tests for ARCH errors, and t tests in models estimated by two-stage least squares. In the experiments reported here, the fast double bootstrap never transforms a bootstrap test that performs poorly into a test that performs perfectly across a wide range of parameter

values. However, it never worsens the performance of a bootstrap test by very much, and it sometimes improves performance quite noticeably. In general, the improvement tends to be largest when the gain from using the single-level bootstrap is relatively great. The FDB works least well for the bootstrap-U test in a regression model estimated by 2SLS, which is a remarkably poor bootstrap test to begin with.

It is important that B be reasonably large when using the fast double bootstrap because, as we saw in Section 4, the procedure tends to overreject when B is small. This problem mainly affects simulation experiments like the ones reported here. For most bootstrap tests, one can safely use a value of B like 99 or 199 when investigating test performance under the null by simulation, but that is not true for FDB tests. This is not really a problem in practice, however, because investigators will generally use quite large values of B whenever there is any doubt about the outcome of a test.

Several of the most interesting results in this paper are not specifically related to the FDB procedure. For the experiments discussed in Section 5, 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 result suggests that equal-tail bootstrap tests deserve closer investigation for a variety of problems where two-tailed tests are commonly used.

For the experiments discussed in Section 6, the performance of both asymptotic and bootstrap tests for ARCH errors is very sensitive to the distribution of the error terms. Although using resampled residuals works asymptotically, it does not always work well in finite samples. Surprisingly, smoothing the resampled residuals has little or no impact on the rejection frequencies for bootstrap tests when the error terms are nonnormal.

For the experiments discussed in Section 7, using a bootstrap DGP that imposes the null hypothesis often works dramatically better than using a bootstrap DGP that is based on unrestricted estimates. No procedure works really well when the instruments are very weak and the correlation between the structural and reduced form errors is large. But FDB bootstrap t tests based on a bootstrap DGP that uses restricted estimates are, in many cases, considerably more reliable than asymptotic t tests or bootstrap t tests based on bootstrap DGPs that use unrestricted estimates.

References

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2004). "Optimal invariant tests for instrumental variables regression", Cowles Foundation Discussion Paper No. 1476, Yale University.

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.

Bound, J., A. A. Jaeger, and R. M. Baker (1995). "Problems with instrumental variables estimation when the correlation between instruments and the endogenous explanatory variables is weak," *Journal of the American Statistical Association*, **90**, 443–450.

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

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

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

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

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

Davidson, R., and J. G. MacKinnon (2002). "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.

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.

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

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

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

Kleibergen, F. (2002). "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression", *Econometrica*, **70**, 1781–1803.

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

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

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

Staiger, D., and J. H. Stock (1997). "Instrumental variables regressions with weak instruments," *Econometrica*, **65**, 557–586.

Stock, J. H., J. H. Wright, and M. Yogo (2002). "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics*, **20**, 518–529.

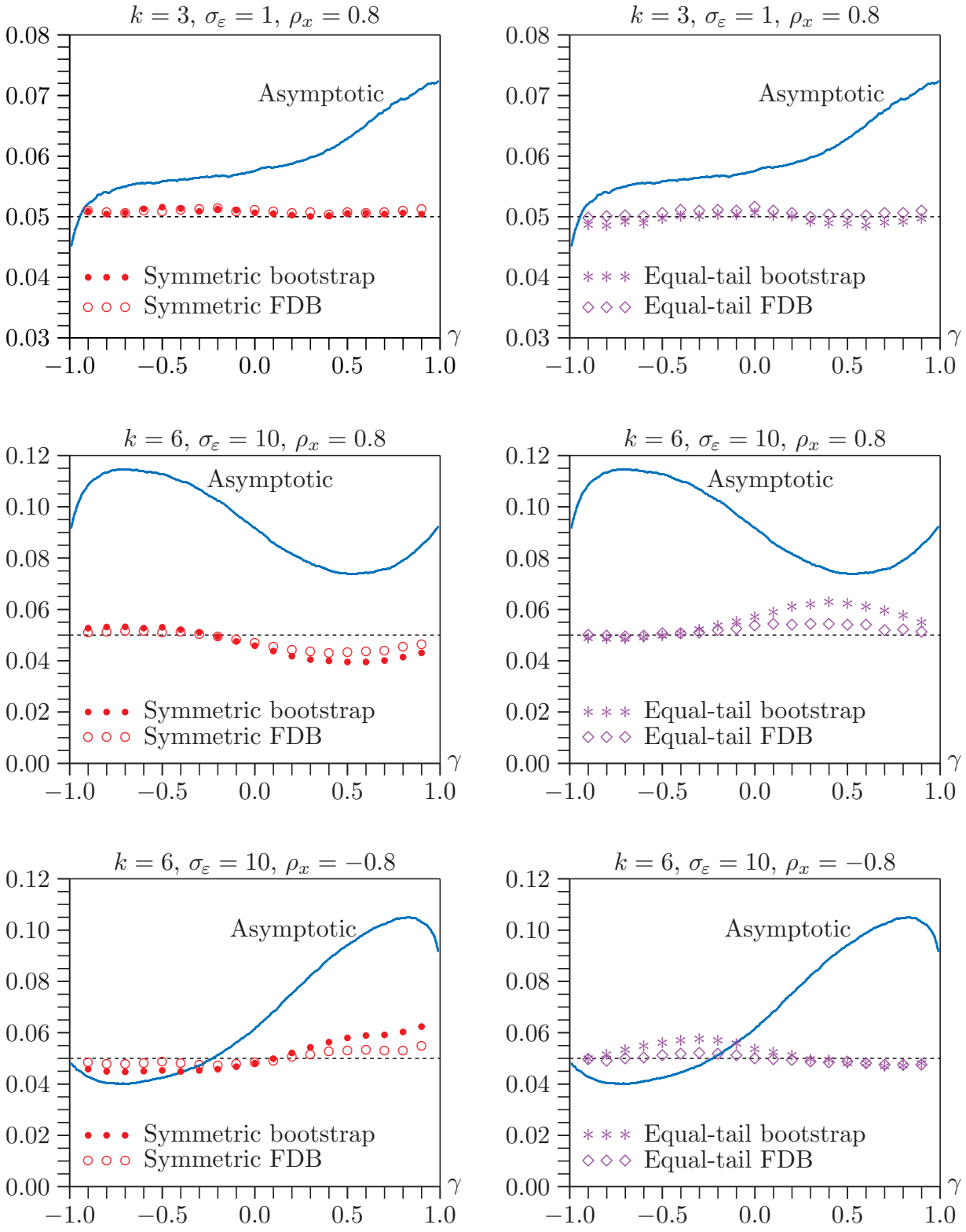

Figure 1. Durbin-Godfrey test rejection frequencies at .05 level under the null, $n = 20$

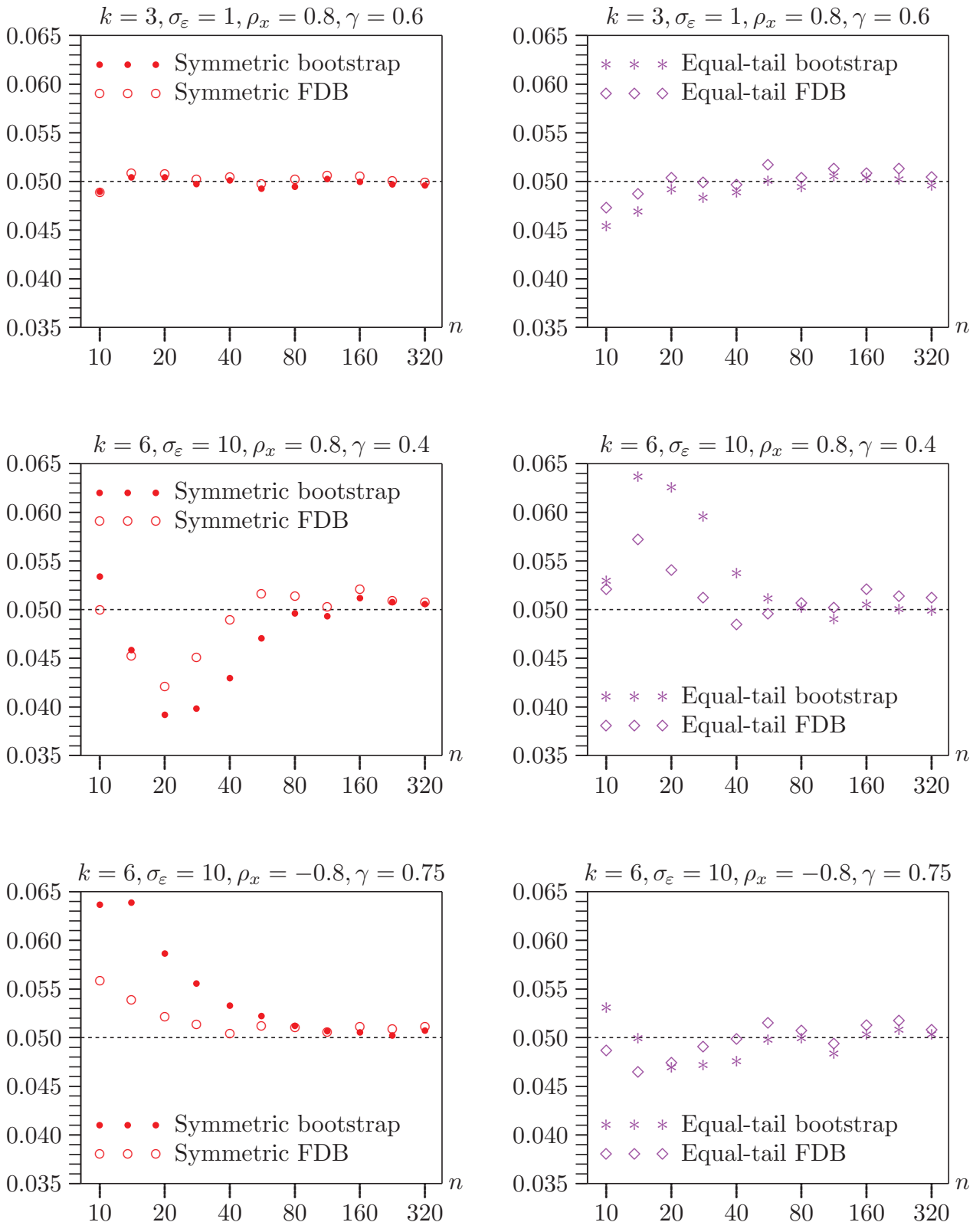

Figure 2. Durbin-Godfrey test rejection frequencies at .05 level under the null

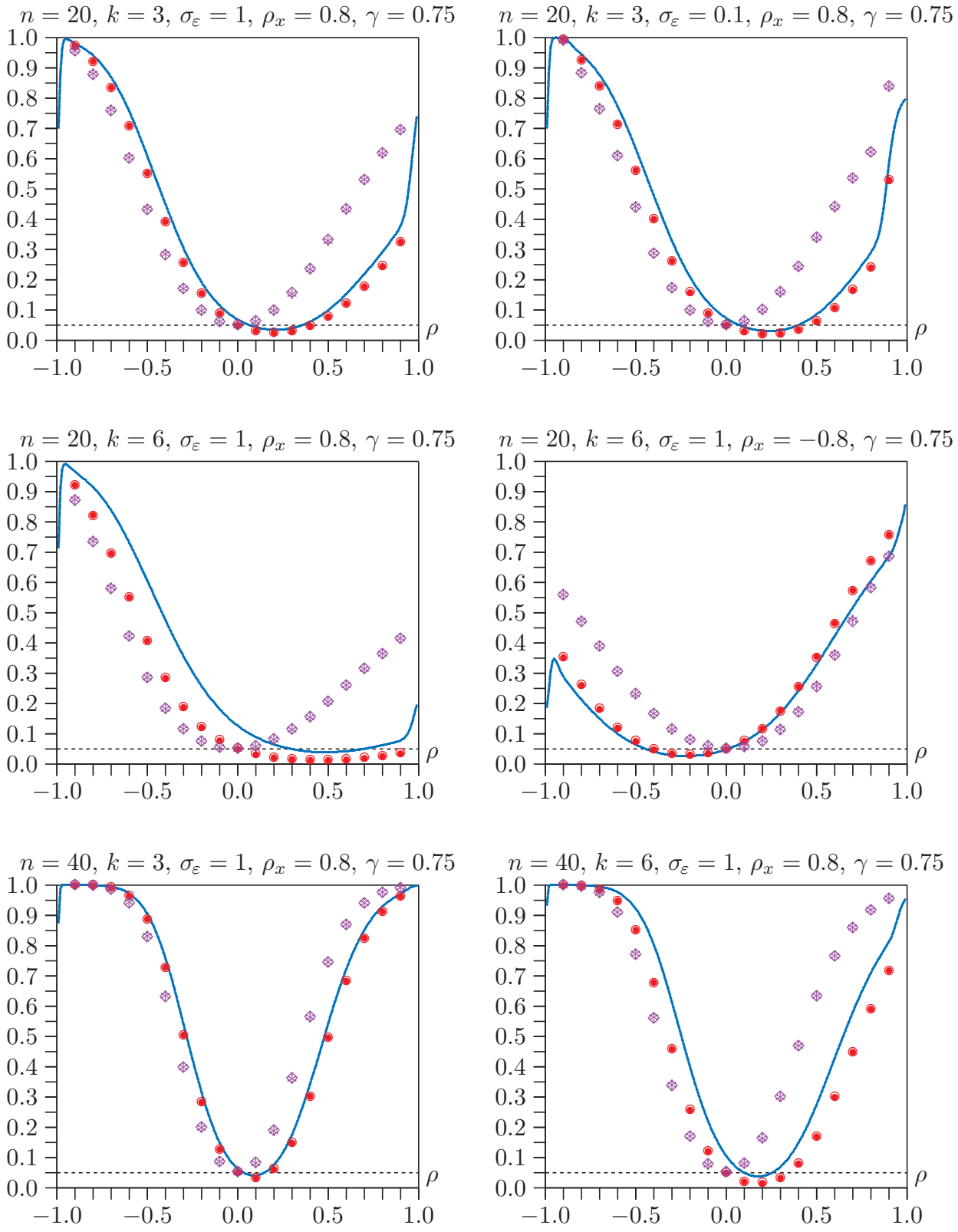

Figure 3. Power of Durbin-Godfrey tests at .05 level

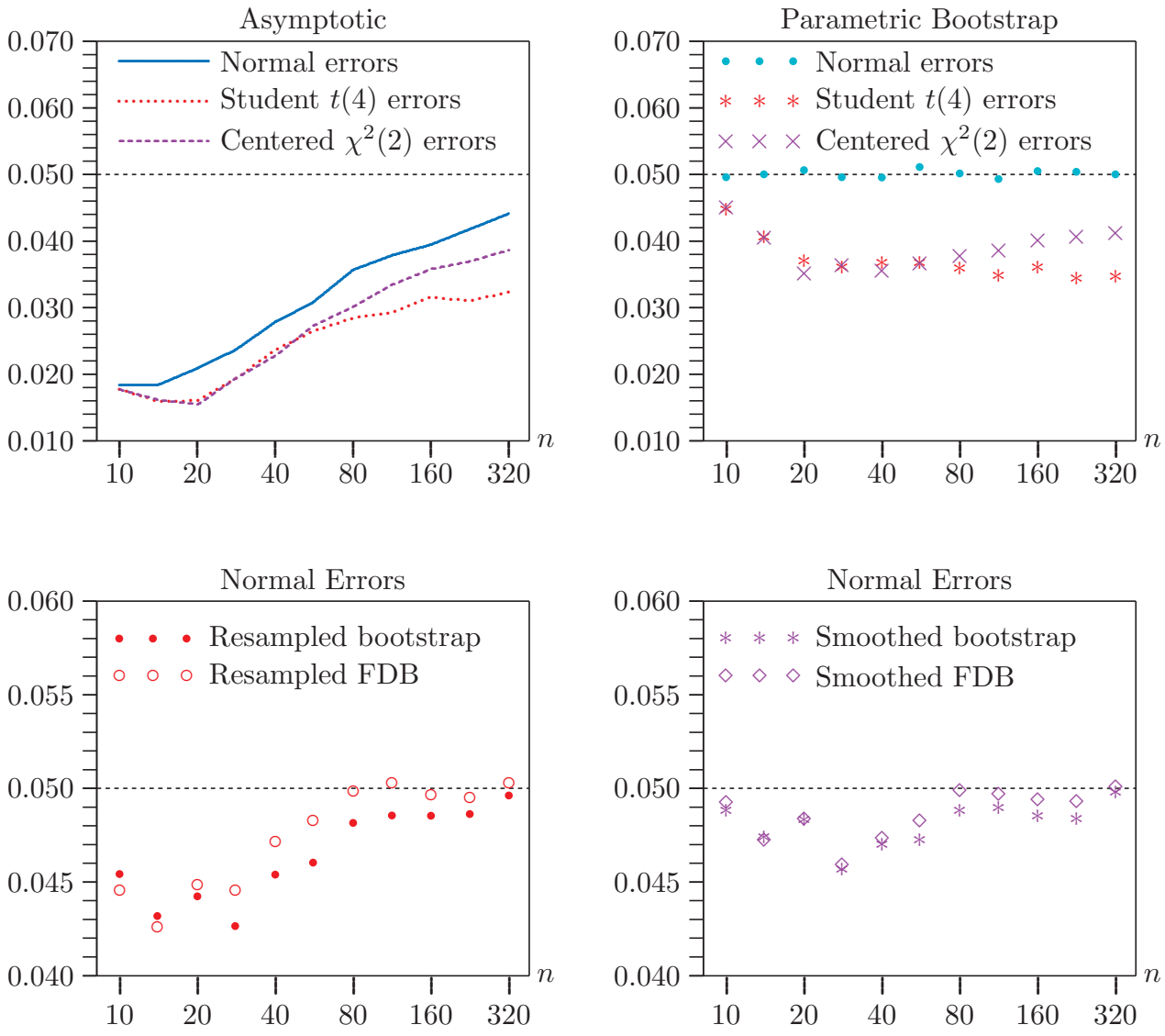

Figure 4. ARCH test rejection frequencies at .05 level under the null

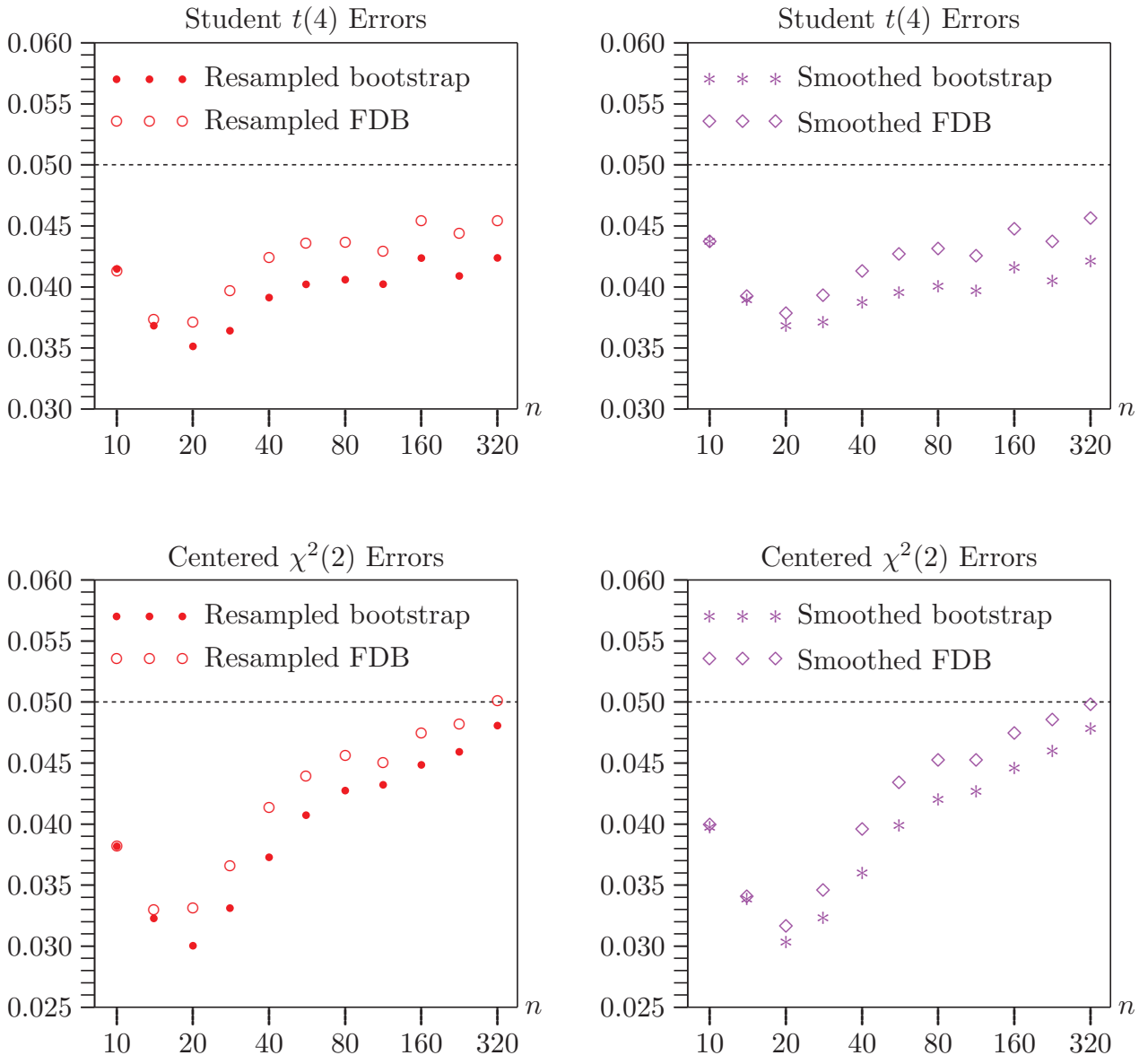

Figure 5. ARCH test rejection frequencies at .05 level under the null

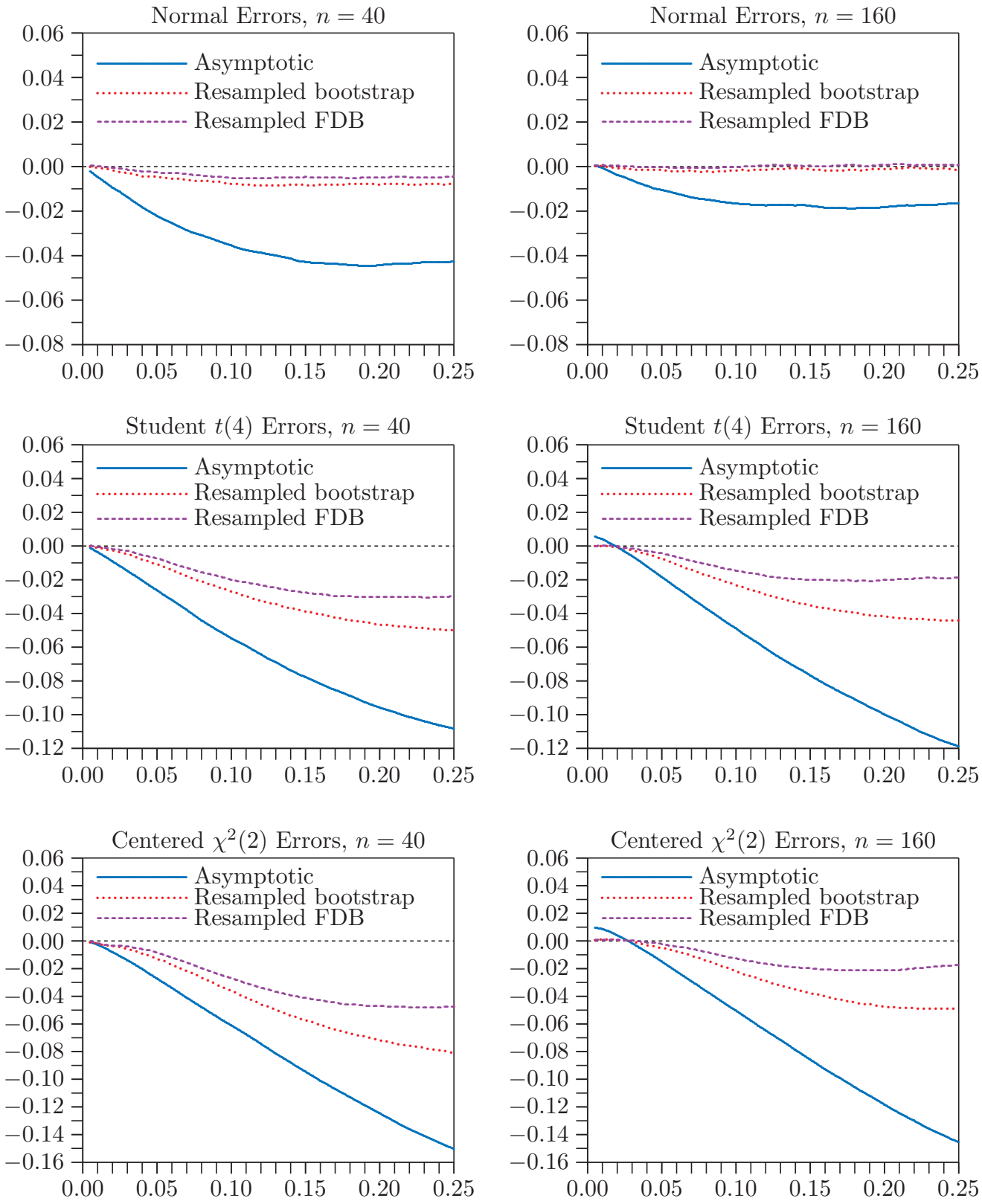

Figure 6. Rejection frequency discrepancy plots for ARCH tests

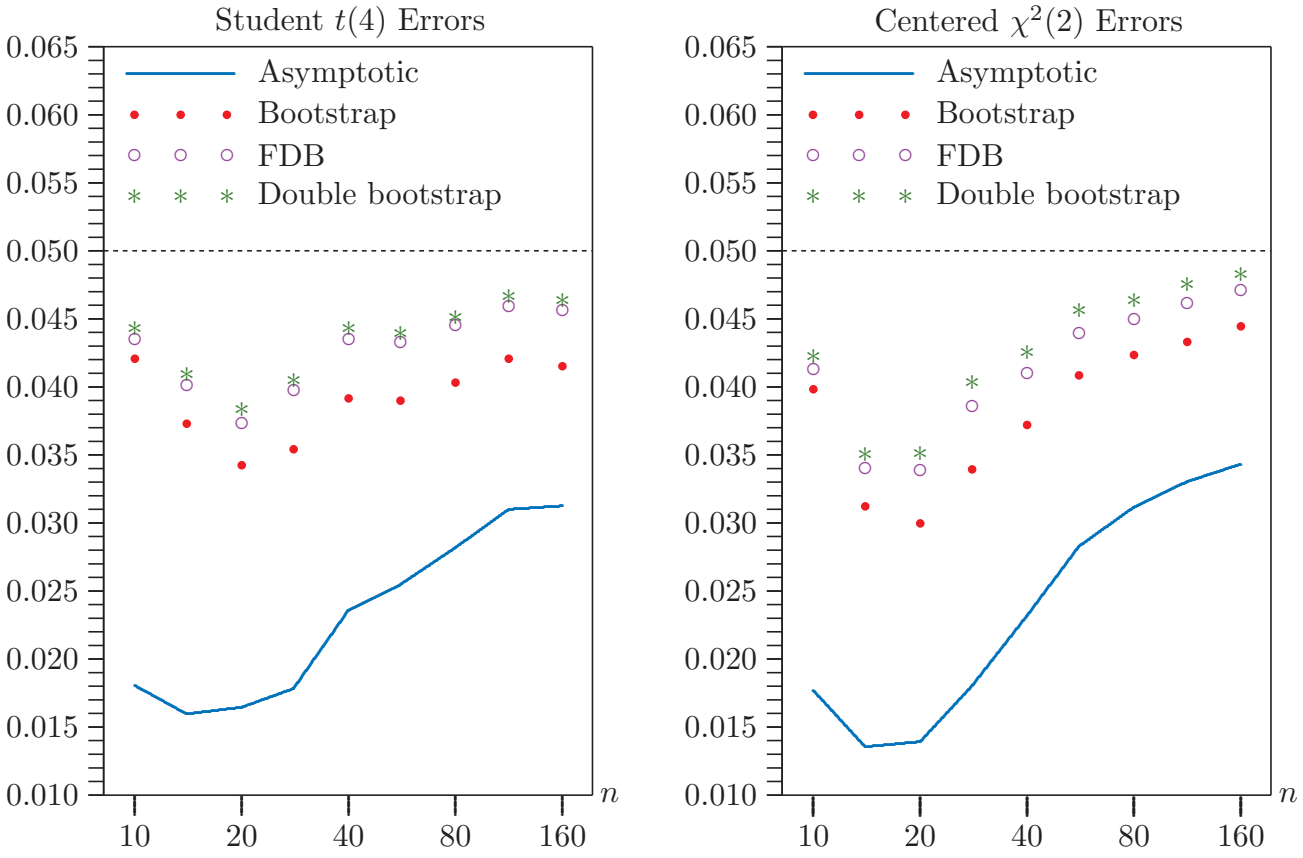

Figure 7. ARCH test rejection frequencies at .05 level under the null (resampled residuals)

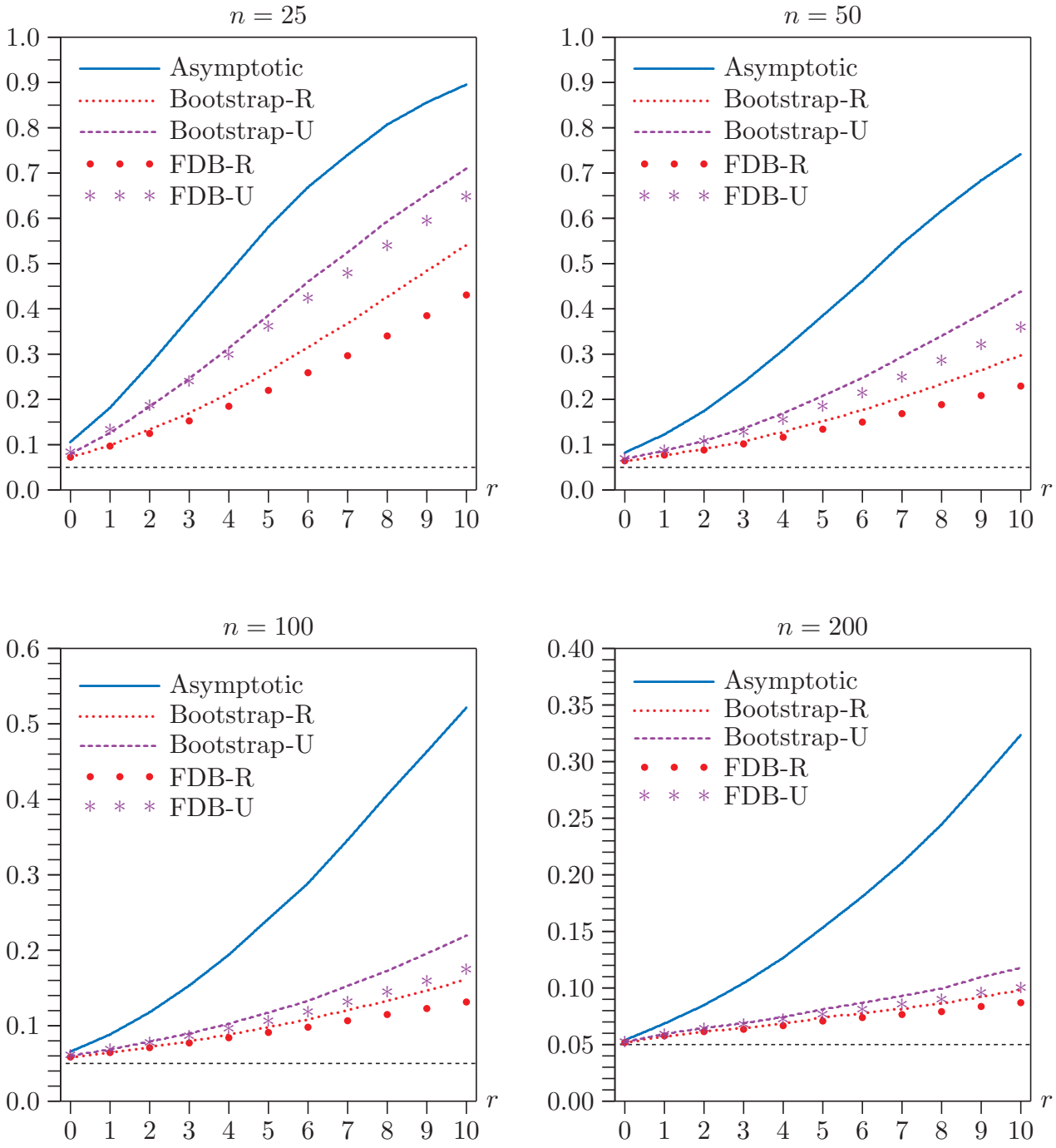

Figure 8. Rejection frequencies for 2SLS t tests at .05 level, $R^2_\infty = 0.2$, $\rho = 0.9$

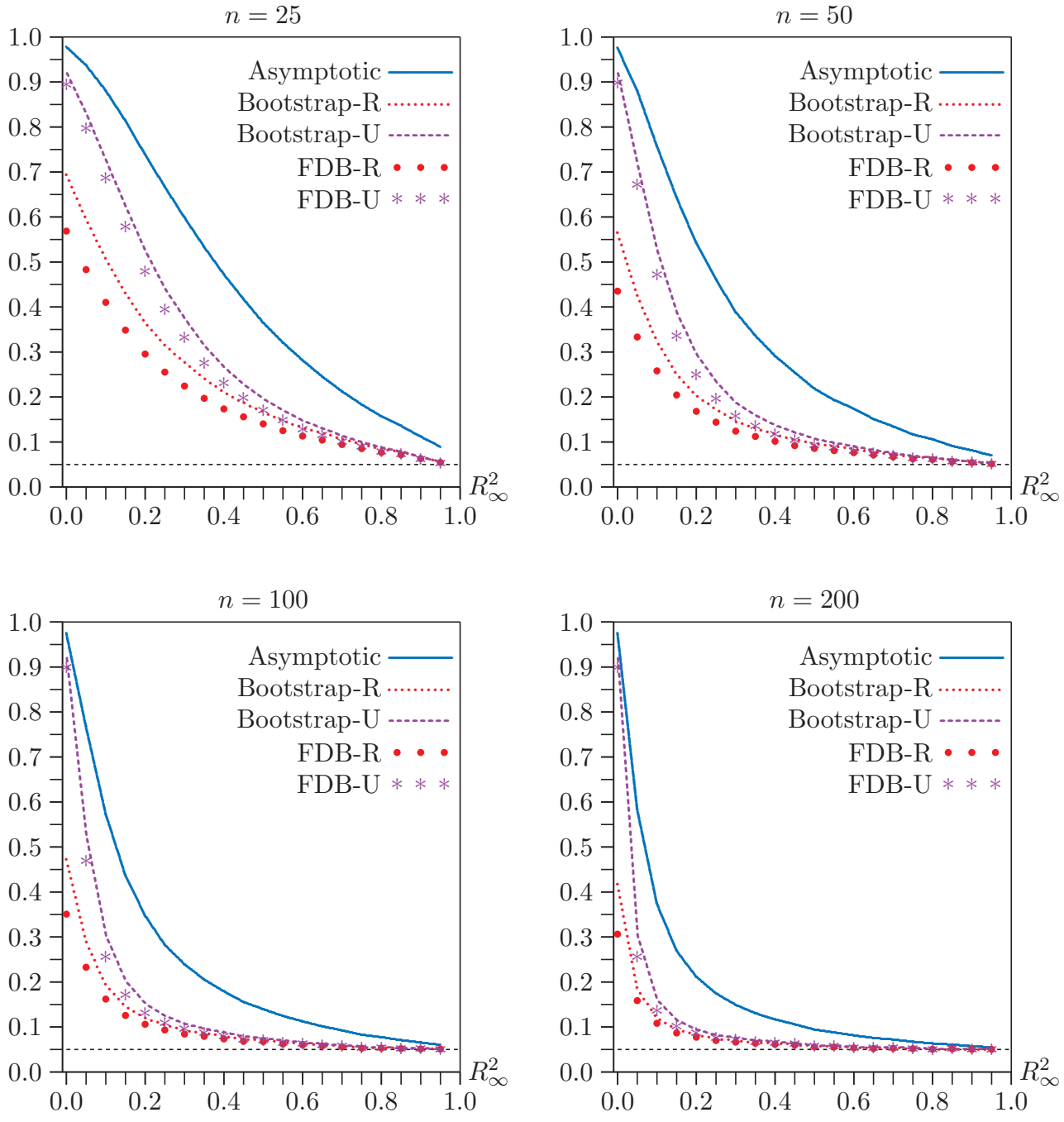

Figure 9. Rejection frequencies for 2SLS t tests at .05 level, $r = 7$, $\rho = 0.9$

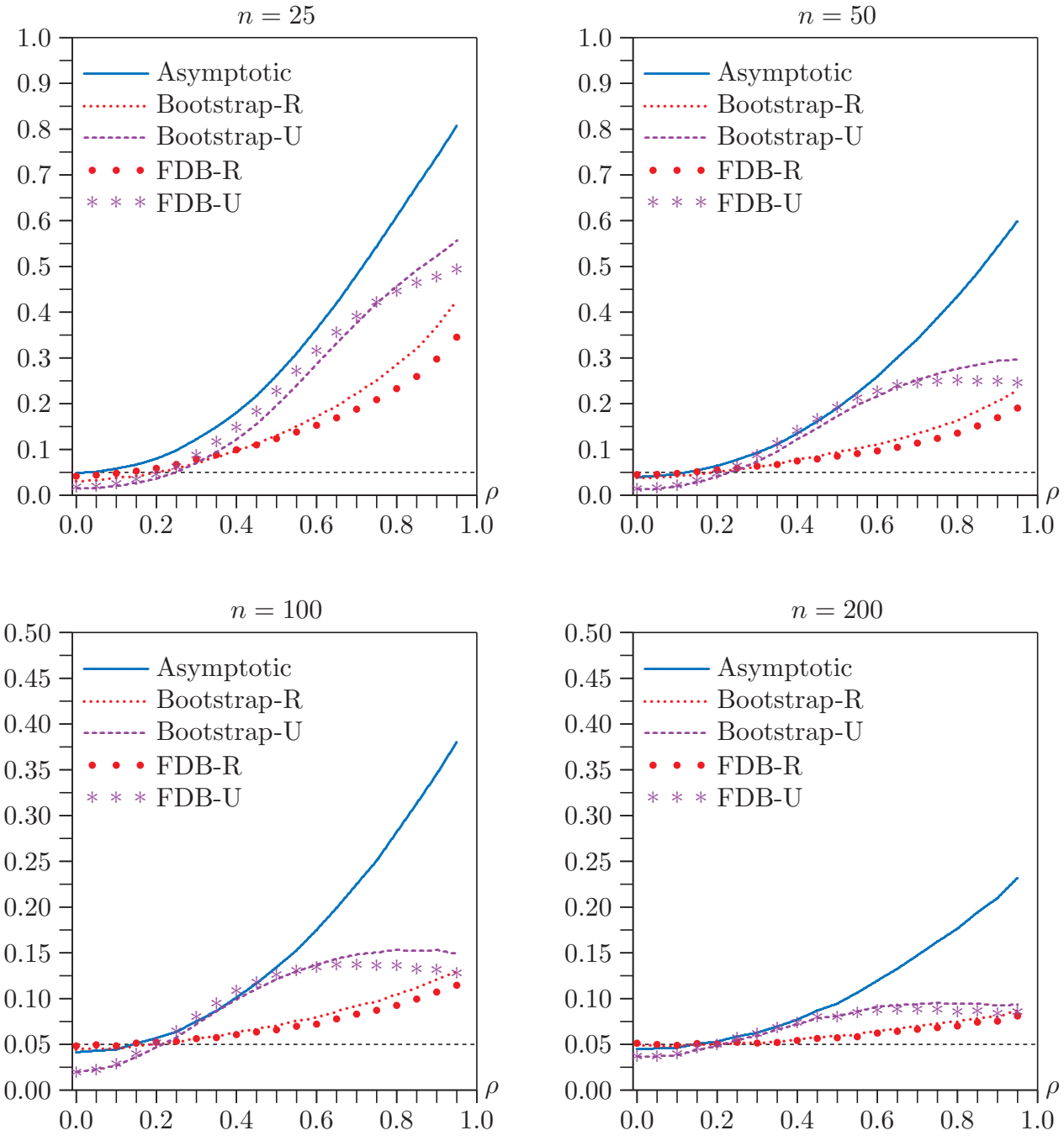

Figure 10. Rejection frequencies for 2SLS t tests at .05 level, $r = 7$, $R^2_\infty = 0.2$

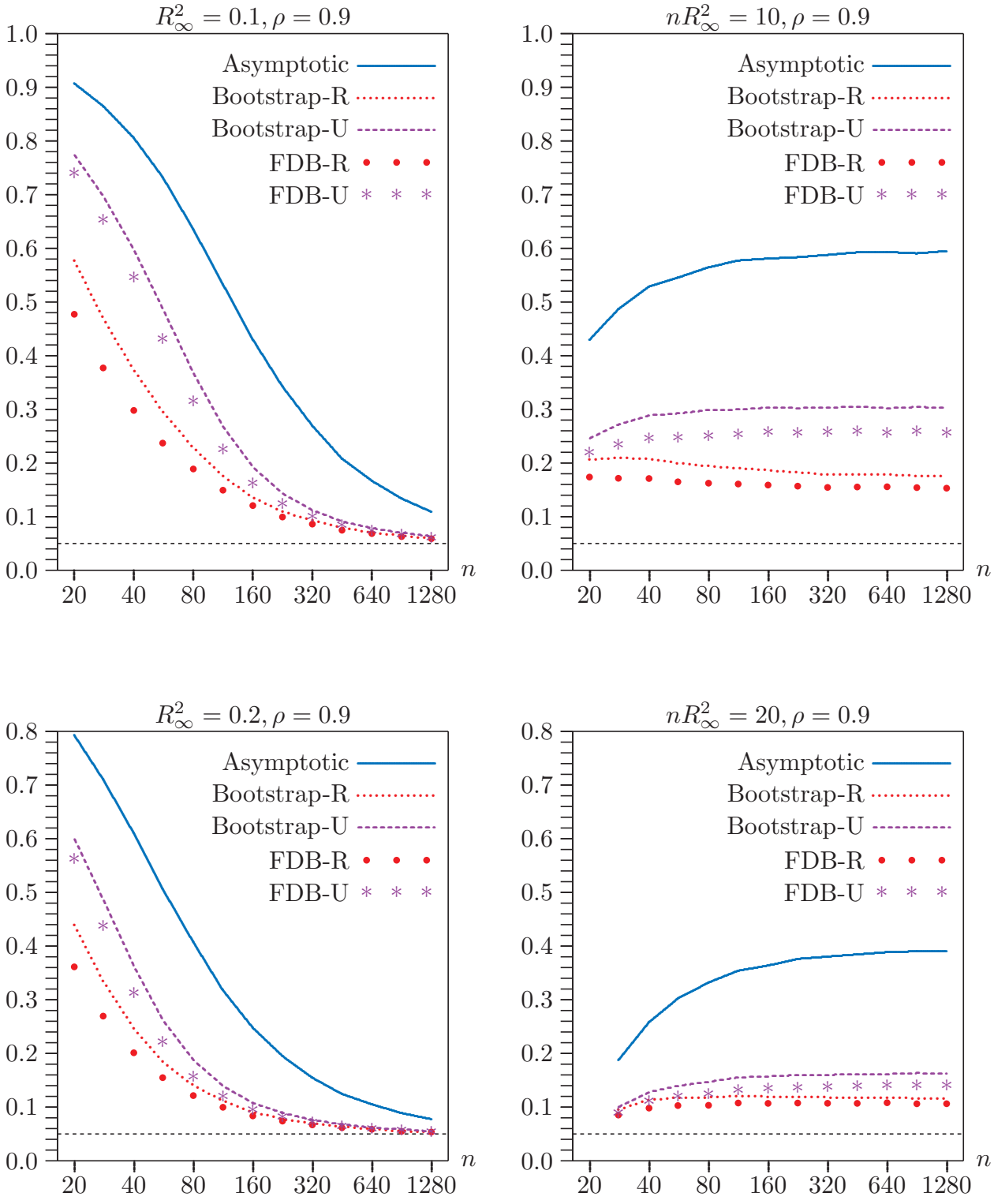

Figure 11. Rejection frequencies for 2SLS t tests at .05 level

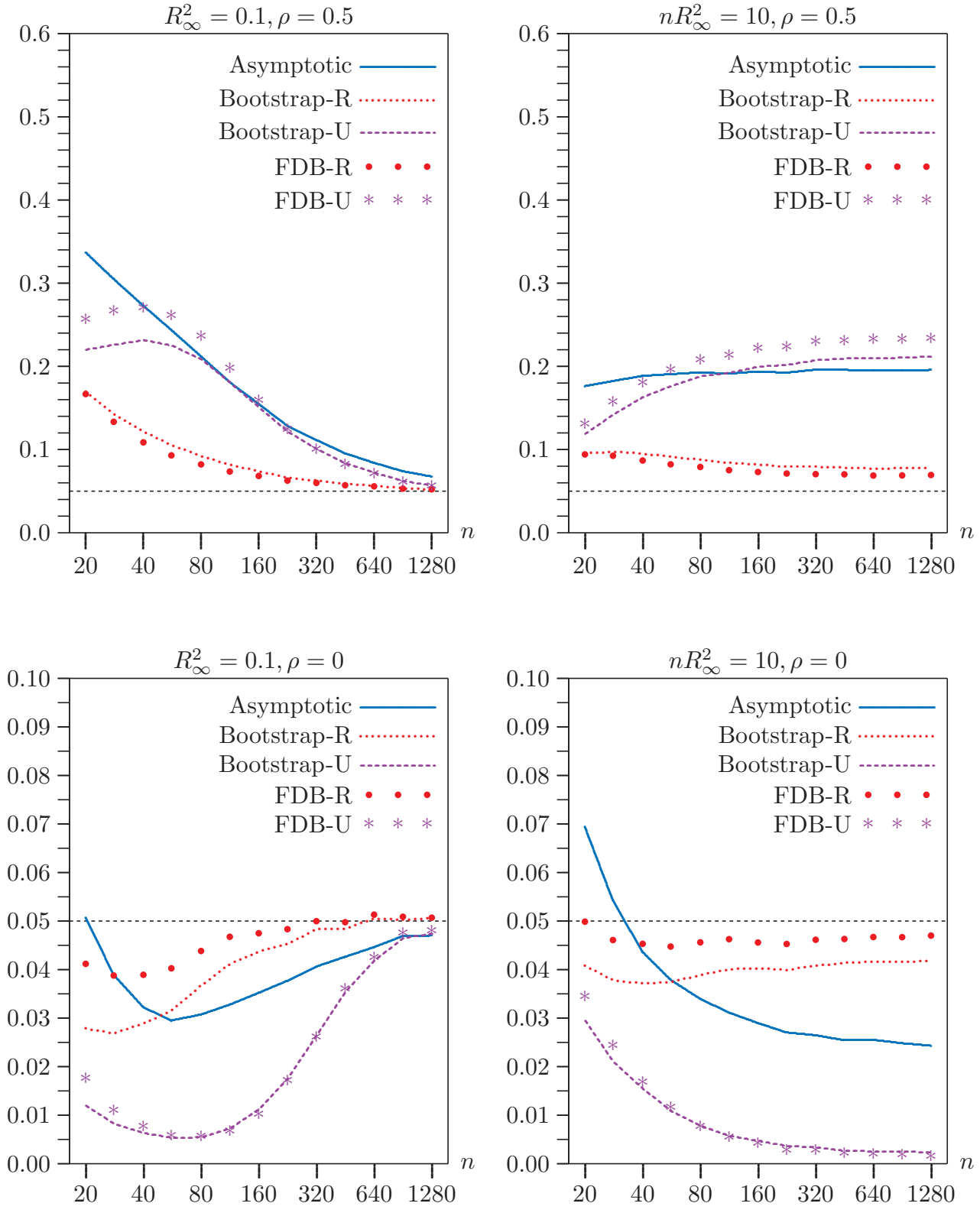

Figure 12. Rejection frequencies for 2SLS t tests at .05 level