

Wild Bootstrap Tests for IV Regression

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

email: Russell.Davidson@mcgill.ca

and

James G. MacKinnon

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

email: jgm@econ.queensu.ca

Abstract

We propose a wild bootstrap procedure for linear regression models estimated by instrumental variables. Like other bootstrap procedures that we have proposed elsewhere, it uses efficient estimates of the reduced-form equation(s). Unlike them, it takes account of possible heteroskedasticity of unknown form. We apply this procedure to t tests, including heteroskedasticity-robust t tests, and provide simulation evidence that it works far better than older methods, such as the pairs bootstrap. We also show how to obtain reliable confidence intervals by inverting bootstrap tests. An empirical example illustrates the utility of these procedures.

Keywords: Instrumental variables estimation, two-stage least squares, wild bootstrap, pairs bootstrap, residual bootstrap, weak instruments, confidence intervals

JEL codes: C12, C15, C30

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada, the Canada Research Chairs program (Chair in Economics, McGill University), and the Fonds Québécois de Recherche sur la Société et la Culture. We are grateful to Arthur Sweetman for a valuable suggestion.

August 2007

1. Introduction

It is often difficult to make reliable inferences from regressions estimated using instrumental variables. This is especially true when the instruments are weak. There is an enormous literature on this subject, much of it quite recent. Most of the papers focus on the case in which there is just one endogenous variable on the right-hand side of the regression, and the problem is to test a hypothesis about the coefficient of that variable. In this paper, we also focus on this case, but, in addition, we discuss confidence intervals, and we allow the number of endogenous variables to exceed two.

One way to obtain reliable inferences is to use statistics with better properties than those of the usual IV t statistic. These include the famous Anderson-Rubin, or AR, statistic proposed in Anderson and Rubin (1949) and extended in Dufour and Taamouti (2005, 2007), the Lagrange Multiplier, or K , statistic proposed in Kleibergen (2002), and the conditional likelihood ratio, or CLR, test proposed in Moreira (2003). A detailed analysis of several tests is found in Andrews, Moreira, and Stock (2006).

A second way to obtain reliable inferences is to use the bootstrap. This approach has been much less popular, probably because the simplest bootstrap methods for this problem do not work very well. See, for example, Flores-Laguna (2007). However, the more sophisticated bootstrap methods recently proposed in Davidson and MacKinnon (2006) work very much better than traditional bootstrap procedures, even when they are combined with the usual t statistic. Although this combination may not be quite as reliable as bootstrapping the CLR or K tests, it works perfectly well in a great many cases.

One advantage of the t statistic over the AR, K , and CLR statistics is that it can easily be modified to be asymptotically valid in the presence of heteroskedasticity of unknown form. But existing procedures for bootstrapping IV t statistics either are not valid in this case or work badly in general. The main contribution of this paper is to propose a new bootstrap data generating processes (DGP) which is valid under heteroskedasticity of unknown form and works well in finite samples even when the instruments are quite weak. This is a wild bootstrap version of one of the methods proposed in Davidson and MacKinnon (2006). Using a heteroskedasticity-robust t statistic combined with this bootstrap method generally seems to work remarkably well.

In the next section, we discuss six bootstrap methods that can be applied to the t statistic on the single right-hand side endogenous variable in a linear regression model estimated by IV. Three of these have been available for some time, two were proposed in Davidson and MacKinnon (2006), and one is a new procedure based on the wild bootstrap. In Section 3, we investigate the finite-sample performance of these bootstrap methods by simulation. Our simulation results are quite extensive and are presented graphically. In Section 4, we briefly discuss the more general case in which there are two or more endogenous variables on the right-hand side. In Section 5, we discuss how to obtain confidence intervals by inverting bootstrap tests. Finally, in Section 6, we present an empirical application that involves estimation of the return to schooling.

2. Bootstrap Methods for IV Regression

In most of this paper, we deal with the two-equation model

$$\mathbf{y}_1 = \beta \mathbf{y}_2 + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}_1 \quad (1)$$

$$\mathbf{y}_2 = \mathbf{W}\boldsymbol{\pi} + \mathbf{u}_2. \quad (2)$$

Here \mathbf{y}_1 and \mathbf{y}_2 are n -vectors of observations on endogenous variables, \mathbf{Z} is an $n \times k$ matrix of observations on exogenous variables, and \mathbf{W} is an $n \times l$ matrix of exogenous instruments with the property that $\mathfrak{S}(\mathbf{Z})$, the subspace spanned by the columns of \mathbf{Z} , lies in $\mathfrak{S}(\mathbf{W})$, the subspace spanned by the columns of \mathbf{W} . Equation (1) is a structural equation, and equation (2) is a reduced-form equation. Observations are indexed by i , so that, for example, y_{1i} denotes the i^{th} element of \mathbf{y}_1 .

We assume that $l > k$. This means that the model is either just identified or over-identified. The disturbances are assumed to be serially uncorrelated. When they are homoskedastic, they have a contemporaneous covariance matrix

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

However, we will often allow them to be heteroskedastic with unknown (but bounded) variances σ_{1i}^2 and σ_{2i}^2 and correlation coefficient ρ_i that may depend on \mathbf{W}_i , the row vector of instrumental variables for observation i .

The usual t statistic for $\beta = \beta_0$ can be written as

$$t_s(\hat{\beta}, \beta_0) = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_1 \|\mathbf{P}_W \mathbf{y}_2 - \mathbf{P}_Z \mathbf{y}_2\|}, \quad (3)$$

where $\hat{\beta}$ is the generalized IV, or 2SLS, estimate of β , \mathbf{P}_W and \mathbf{P}_Z are the matrices that project orthogonally on to the subspaces $\mathfrak{S}(\mathbf{W})$ and $\mathfrak{S}(\mathbf{Z})$, respectively, and $\|\cdot\|$ denotes the Euclidean length of a vector. In equation (3),

$$\hat{\sigma}_1 = \left(\frac{1}{n} \hat{\mathbf{u}}_1^\top \hat{\mathbf{u}}_1 \right)^{1/2} = \left(\frac{1}{n} (\mathbf{y}_1 - \hat{\beta} \mathbf{y}_2 - \mathbf{Z} \hat{\boldsymbol{\gamma}})^\top (\mathbf{y}_1 - \hat{\beta} \mathbf{y}_2 - \mathbf{Z} \hat{\boldsymbol{\gamma}}) \right)^{1/2} \quad (4)$$

is the usual 2SLS estimate of σ_1 . Here $\hat{\boldsymbol{\gamma}}$ denotes the IV estimate of $\boldsymbol{\gamma}$, and $\hat{\mathbf{u}}_1$ is the usual vector of IV residuals. Many regression packages divide by $n - k - 1$ instead of by n . Since $\hat{\sigma}_1$ as defined in (4) is not necessarily biased downwards, we do not do so.

When homoskedasticity is not assumed, the usual t statistic (3) should be replaced by the heteroskedasticity-robust t statistic

$$t_h(\hat{\beta}, \beta_0) = \frac{\hat{\beta} - \beta_0}{s_h(\hat{\beta})}, \quad (5)$$

where

$$s_h(\hat{\beta}) \equiv \frac{(\sum_{i=1}^n \hat{u}_{1i}^2 (\mathbf{P}_W \mathbf{y}_2 - \mathbf{P}_Z \mathbf{y}_2)_i)^{1/2}}{\|\mathbf{P}_W \mathbf{y}_2 - \mathbf{P}_Z \mathbf{y}_2\|^2}. \quad (6)$$

Here $(\mathbf{P}_W \mathbf{y}_2 - \mathbf{P}_Z \mathbf{y}_2)_i$ denotes the i^{th} element of the vector $\mathbf{P}_W \mathbf{y}_2 - \mathbf{P}_Z \mathbf{y}_2$. Expression (6) is what most regression packages routinely print as a heteroskedasticity-consistent standard error for $\hat{\beta}$. It is evidently the square root of a sandwich variance estimate.

The basic idea of bootstrap testing is to compare the observed value of some test statistic, say $\hat{\tau}$, with the empirical distribution of a number of bootstrap test statistics, say τ_j^* , for $j = 1, \dots, B$, where B is the number of bootstrap replications. If α is the level of the test we wish to perform, it is desirable that $\alpha(B + 1)$ should be an integer, and a commonly used value of B is 999. If we are prepared to assume that τ is symmetrically distributed around the origin, it is appropriate to use the **symmetric bootstrap P value**

$$\hat{p}_s^*(\hat{\tau}) = \frac{1}{B} \sum_{j=1}^B \mathbf{I}(|\tau_j^*| > |\hat{\tau}|). \quad (7)$$

We reject the null hypothesis whenever $\hat{p}_s^*(\hat{\tau}) < \alpha$.

In the case of IV t statistics, the probability of rejecting in one direction can be very much greater than the probability of rejecting in the other, because $\hat{\beta}$ is often biased. In such cases, we can use the **equal-tail bootstrap P value**

$$\hat{p}_{\text{et}}^*(\hat{\tau}) = 2 \min\left(\frac{1}{B} \sum_{j=1}^B \mathbf{I}(\tau_j^* < \hat{\tau}), \frac{1}{B} \sum_{j=1}^B \mathbf{I}(\tau_j^* > \hat{\tau})\right). \quad (8)$$

Here we actually perform two tests, one against values in the lower tail of the distribution and the other against values in the upper tail, and reject if either of them yields a bootstrap P value less than $\alpha/2$.

The choice of the DGP used to generate the bootstrap samples is critical, and it can dramatically affect the properties of bootstrap tests. In the remainder of this section, we discuss six different bootstrap DGPs for t tests of $\beta = \beta_0$ in the IV regression model given by (1) and (2). Three of these have been around for some time, but they often work badly. Two were proposed in Davidson and MacKinnon (2006), and they generally work very well under homoskedasticity. The last one is new. It is a wild bootstrap test that takes account of heteroskedasticity of unknown form.

The oldest and best-known method for bootstrapping the test statistics (3) and (5) is to use the **pairs bootstrap**, which was originally proposed in Freedman (1981) and applied to 2SLS regression in Freedman (1984). The idea is to resample the rows of the matrix

$$[\mathbf{y}_1 \quad \mathbf{y}_2 \quad \mathbf{W}]. \quad (9)$$

For the pairs bootstrap, the i^{th} row of each bootstrap sample is simply one of the rows of the matrix (9), chosen at random with probability $1/n$. Other variants of

the pairs bootstrap have been proposed for this problem. In particular, Moreira, Porter, and Suarez (2005) propose a variant that seems more complicated, because it involves estimating the model, but actually yields identical results when applied to both ordinary and heteroskedasticity-robust t statistics. Flores-Laguna (2007) proposes another variant that yields results very similar, but not identical, to those from the ordinary pairs bootstrap.

Because the pairs bootstrap DGP does not impose the null hypothesis, the bootstrap test statistics must be computed as

$$t(\hat{\beta}_j^*, \hat{\beta}) = \frac{\hat{\beta}_j^* - \hat{\beta}}{\text{se}(\hat{\beta}_j^*)}. \quad (10)$$

Here, $\hat{\beta}_j^*$ is the IV estimate of β from the j^{th} bootstrap sample, and $\text{se}(\hat{\beta}_j^*)$ is the standard error of $\hat{\beta}_j^*$, calculated by whatever method is used for the standard error of $\hat{\beta}$ in the t statistic that is being bootstrapped. If we used β_0 in place of $\hat{\beta}$ in (10), we would be testing a hypothesis that is not true of the bootstrap DGP.

The pairs bootstrap is fully nonparametric and is valid in the presence of heteroskedasticity of unknown form, but, as we shall see in the next section, it has little else to recommend it. The other bootstrap methods that we consider are semiparametric and require estimation of the model given by (1) and (2). We consider a number of ways of estimating this model and constructing bootstrap DGPs.

The least efficient way to estimate the model is to use OLS on the reduced-form equation (2) and IV on the structural equation (1), without imposing the restriction that $\beta = \beta_0$. This yields estimates $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\pi}$, a vector of IV residuals \hat{u}_1 , and a vector of OLS residuals \hat{u}_2 . Using these estimates, we can easily construct the DGP for the **unrestricted residual bootstrap**, or **UR bootstrap**. The UR bootstrap DGP can be written as

$$y_{i1}^* = \hat{\beta}y_{2i}^* + \mathbf{Z}_i\hat{\gamma} + \hat{u}_{1i}^* \quad (11)$$

$$y_{2i}^* = \mathbf{W}_i\hat{\pi} + \hat{u}_{2i}^*, \quad (12)$$

where

$$\begin{bmatrix} \hat{u}_{1i}^* \\ \hat{u}_{2i}^* \end{bmatrix} \sim \text{EDF} \left(\begin{array}{c} \hat{u}_{1i} \\ (n/(n-l))^{1/2}\hat{u}_{2i} \end{array} \right). \quad (13)$$

Equations (11) and (12) are simply the structural and reduced-form equations evaluated at the unrestricted estimates. Note that we could omit $\mathbf{Z}_i\hat{\gamma}$ from equation (11), since the t statistics are invariant to the true value of γ .

According to (13), the bootstrap disturbances are drawn in pairs from the joint empirical distribution of the unrestricted residuals, with the residuals from the reduced-form equation rescaled so as to have variance equal to the OLS variance estimate. This rescaling is not essential. It would also be possible to rescale the residuals from the

structural equation, but it is unclear what benefit might result. The bootstrap DGP given by (11), (12), and (13) ensures that, asymptotically, the joint distribution of the bootstrap disturbances is the same as the joint distribution of the actual disturbances if the model is correctly specified and the disturbances are homoskedastic.

Since the UR bootstrap DGP does not impose the null hypothesis, the bootstrap test statistics must be calculated in the same way as for the pairs bootstrap, using equation (10), so as to avoid testing a hypothesis that is not true of the bootstrap DGP.

Whenever possible, it is desirable to impose the null hypothesis of interest on the bootstrap DGP. This is because imposing a (true) restriction makes estimation more efficient, and using more efficient estimates in the bootstrap DGP should reduce the error in rejection probability associated with the bootstrap test. In some cases, it can even improve the rate at which the error in rejection frequency shrinks as the sample size increases; see Davidson and MacKinnon (1999). All of the remaining bootstrap methods that we discuss impose the null hypothesis.

The DGP for the **restricted residual bootstrap**, or **RR bootstrap**, is very similar to the one for the UR bootstrap, but it imposes the null hypothesis on both the structural equation and the bootstrap disturbances. We replace equation (11) by

$$y_{i1}^* = \tilde{u}_{1i}^*, \quad (14)$$

since the value of γ does not matter. Equation (12) is used unchanged, and equation (13) is replaced by

$$\begin{bmatrix} \tilde{u}_{1i}^* \\ \hat{u}_{2i}^* \end{bmatrix} \sim \text{EDF} \left(\begin{array}{c} (n/(n-k))^{1/2} \tilde{u}_{1i} \\ (n/(n-l))^{1/2} \hat{u}_{2i} \end{array} \right).$$

Since \tilde{u}_{1i} is just an OLS residual, it makes sense to rescale it.

As we shall see in the next section, the RR bootstrap outperforms the pairs and UR bootstraps, but, like them, it does not work at all well when the instruments are weak. The problem is that $\hat{\boldsymbol{\pi}}$ is not an efficient estimator of $\boldsymbol{\pi}$, and, when the instruments are weak, $\hat{\boldsymbol{\pi}}$ may be very inefficient indeed. Therefore, Davidson and MacKinnon (2006) suggested using a more efficient estimator, which was also used by Kleibergen (2002) in constructing the K statistic. This estimator is asymptotically equivalent to the ones that would be obtained by using either 3SLS or FIML on the system consisting of equations (1) and (2). It may be obtained by running the regression

$$\mathbf{y}_2 = \mathbf{W}\boldsymbol{\pi} + \delta \mathbf{M}_Z \mathbf{y}_1 + \text{residuals}. \quad (15)$$

This is just the reduced-form equation (2) augmented by the residuals from restricted estimation of the structural equation (1). It yields estimates $\tilde{\boldsymbol{\pi}}$ and $\tilde{\delta}$ and residuals

$$\tilde{\mathbf{u}}_2 \equiv \mathbf{y}_2 - \mathbf{W}\tilde{\boldsymbol{\pi}}.$$

These are not the OLS residuals from (15), which would be too small, but the OLS residuals plus $\tilde{\delta}\mathbf{M}_Z\mathbf{y}_1$.

This procedure provides all the ingredients for what Davidson and MacKinnon (2006) call the **restricted efficient residual bootstrap**, or **RE bootstrap**. The DGP uses equation (14) as the structural equation and

$$y_{2i}^* = \mathbf{W}_i \tilde{\boldsymbol{\pi}} + \tilde{u}_{2i}^* \quad (16)$$

as the reduced-form equation, and the bootstrap disturbances are generated by

$$\begin{bmatrix} \tilde{u}_{1i}^* \\ \tilde{u}_{2i}^* \end{bmatrix} \sim \text{EDF} \left(\begin{array}{c} (n/(n-k))^{1/2} \tilde{u}_{1i} \\ (n/(n-l))^{1/2} \tilde{u}_{2i} \end{array} \right). \quad (17)$$

Here the residuals are rescaled in exactly the same way as for the RR bootstrap. This rescaling, which is optional, should have only a slight effect unless k and/or l is large relative to n .

One of several possible measures of how strong the instruments are is the **concentration parameter**, which can be written as

$$a^2 \equiv \frac{1}{\sigma_2^2} \boldsymbol{\pi}^\top \mathbf{W}^\top \mathbf{M}_Z \mathbf{W} \boldsymbol{\pi}. \quad (18)$$

Evidently, the concentration parameter is large when the ratio of the error variance in the reduced-form equation to the variance explained by the part of the instruments that is orthogonal to the exogenous variables in the structural equation is small. We can estimate a^2 using either OLS estimates of equation (2) or the more efficient estimates $\tilde{\boldsymbol{\pi}}_1$ and $\tilde{\sigma}$ obtained from regression (15). However, both estimates are biased upwards, because of the tendency for OLS estimates to fit too well. Davidson and MacKinnon (2006) therefore proposes the bias-corrected estimator

$$\tilde{a}_{\text{BC}}^2 \equiv \max(0, \tilde{a}^2 - (l-k)(1 - \tilde{\rho}^2)),$$

where $\tilde{\rho}$ is the sample correlation between the elements of $\tilde{\mathbf{u}}_1$ and $\tilde{\mathbf{u}}_2$. The bias-corrected estimator can be used in a modified version of the RE bootstrap, called the REC bootstrap by Davidson and MacKinnon. It uses

$$y_{2i}^* = \mathbf{W}_{1i} \ddot{\boldsymbol{\pi}}_1 + \tilde{u}_{2i}^*, \quad \text{where } \ddot{\boldsymbol{\pi}}_1 = (\tilde{a}_{\text{BC}}/\tilde{a})\tilde{\boldsymbol{\pi}}_1,$$

instead of equation (16) as the reduced-form equation in the bootstrap DGP. The bootstrap disturbances are still generated by (17). Simulation experiments not reported here, in addition to those in the original paper, show that, when applied to t statistics, the performance of the RE and REC bootstraps tends to be very similar. Either one of them may perform better in any particular case, but neither appears to be superior overall. We therefore do not discuss the REC bootstrap further.

As shown in Davidson and MacKinnon (2006), and as we will see in the next section, the RE bootstrap, based on efficient estimates of the reduced form, generally works very much better than earlier methods. However, like the RR and UR bootstraps (and unlike the pairs bootstrap), it takes no account of possible heteroskedasticity. We now propose a new bootstrap method which does so. It is a wild bootstrap version of the RE bootstrap.

The wild bootstrap was originally proposed in Wu (1986) in the context of OLS regression. It can be generalized quite easily to the IV case studied in this paper. The idea of the wild bootstrap is to use for the bootstrap disturbance(s) associated with the i^{th} observation the actual residual(s) for that observation, possibly transformed in some way, and multiplied by a random variable, independent of the data, with mean 0 and variance 1. Often, a binary random variable is used for this purpose. We propose the **wild restricted efficient residual bootstrap**, or **WRE bootstrap**. The DGP uses (14) and (16) as the structural and reduced form equations, respectively, with

$$\begin{bmatrix} \tilde{u}_{1i}^* \\ \tilde{u}_{2i}^* \end{bmatrix} = \begin{bmatrix} (n/(n-k))^{1/2} \tilde{u}_{1i} v_i^* \\ (n/(n-l))^{1/2} \tilde{u}_{2i} v_i^* \end{bmatrix}, \quad (19)$$

where v_i^* is a random variable that has mean 0 and variance 1. Until recently, the most popular choice for v_i^* has been

$$v_i^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}); \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}). \end{cases}$$

However, Davidson and Flachaire (2001) have shown that, when the disturbances are not too asymmetric, it is better to use the **Rademacher distribution**, according to which

$$v_i^* = 1 \text{ with probability } \frac{1}{2}; \quad v_i^* = -1 \text{ with probability } \frac{1}{2}. \quad (20)$$

Notice that, in equation (19), both rescaled residuals are multiplied by the same value of v_i^* . This preserves the correlation between the two disturbances, at least when they are symmetrically distributed. Using the Rademacher distribution (20) imposes symmetry on the bivariate distribution of the bootstrap disturbances, and this may affect the correlation when they are not actually symmetric.

There is a good deal of evidence that the wild bootstrap works reasonably well for univariate regression models, even when there is quite severe heteroskedasticity. See, among others, Gonçalves and Kilian (2004) and MacKinnon (2006). Although the wild bootstrap cannot be expected to work quite as well as a comparable residual bootstrap method when the disturbances are actually homoskedastic, the cost of insuring against heteroskedasticity generally seems to be very small; see Section 3.

Of course, it is straightforward to create wild bootstrap versions of the RR and REC bootstraps that are analogous to the WRE bootstrap. In our simulation experiments, we studied these methods, which it is natural to call the WRR and WREC bootstraps,

respectively. However, we do not report results for either of them. The performance of WRR is very similar to that of RR when the disturbances are homoskedastic, and the performance of WREC is generally quite similar to that of WRE.

3. Finite-Sample Properties of Competing Bootstrap Methods

In this section, we graphically report the results of a number of large-scale sampling experiments. These were designed to investigate several important issues, not all of which are specific to the new wild bootstrap procedures that we propose in this paper.

In many of our experiments, there is no heteroskedasticity, and the data are generated by the simplified model

$$\mathbf{y}_1 = \beta \mathbf{y}_2 + \mathbf{u}_1 \tag{21}$$

$$\mathbf{y}_2 = a \mathbf{w}_1 + \mathbf{u}_2, \quad \mathbf{u}_2 = \rho \mathbf{u}_1 + r \mathbf{v}, \tag{22}$$

where $r \equiv (1 - \rho^2)^{1/2}$. Here the n -vectors \mathbf{u}_1 and \mathbf{v} contain independent, standard normal elements. The elements of \mathbf{u}_1 and \mathbf{u}_2 are contemporaneously correlated, but serially uncorrelated, standard normal random variables with correlation ρ . The instrument vector \mathbf{w}_1 is normally distributed and scaled so that $\|\mathbf{w}_1\| = 1$. This, together with the way the disturbances are constructed, ensures that the square of the coefficient a in (22) is the concentration parameter a^2 defined in (18).

Although there is just one instrument in equation (22), the model that is actually estimated, namely (1) and (2), includes l of them, of which one is \mathbf{w}_1 , $l - 2$ are standard normal random variables that have no explanatory power, and the last is a constant term, which is also the sole column of the matrix \mathbf{Z} of exogenous explanatory variables in the structural equation, so that $k = 1$. Including a constant term ensures that the residuals have mean zero and do not have to be recentered for the residual bootstraps.

In the context of the DGP given by (21) and (22), there are only four parameters that influence the finite-sample performance of the tests, whether asymptotic or bootstrap. The four parameters are the sample size n , $l - k$, which is one more than the number of overidentifying restrictions, a (or, equivalently, a^2), and ρ . In most of our experiments, we hold a fixed as we vary n . This implies a version of the weak-instrument asymptotics of Staiger and Stock (1997). Consequently, we do not expect any method to work perfectly, even as $n \rightarrow \infty$. By allowing n and a to vary independently, we are able to separate the effects of sample size *per se* from the effects of instrument weakness.

All experiments use 100,000 replications for each set of parameter values, and all bootstrap tests are based on $B = 399$ bootstrap replications. This is a smaller number than should generally be used in practice, but it is perfectly satisfactory for simulation experiments, because experimental randomness in the bootstrap P values tends to average out across replications. Unless otherwise noted, bootstrap tests are based on the equal-tail P value (8) rather than the symmetric P value (7). In some cases, as

we discuss later, using the latter would have produced noticeably different results. We focus on rejection frequencies for tests at the .05 level. Results for rejection frequencies at other common levels are qualitatively similar.

Figure 1 shows the effects of varying a from 1 (instruments very weak) to 64 (instruments extremely strong) by factors of $\sqrt{2}$. In these experiments, $n = 400$ and $l - k = 11$. The reasons for choosing these values will be discussed below. In the top two panels, $\rho = 0.9$, and, in the bottom two, $\rho = 0.1$. The left-hand panels show rejection frequencies for the asymptotic test and the pairs, UR, and RR bootstraps. The right-hand panels show rejection frequencies for the RE and WRE bootstraps, as well as partial ones for the RR bootstrap. Notice that the vertical axis is different in every panel and has a much larger range in the left-hand panels than in the right-hand ones. Results are shown for both the usual t statistic (3) and the heteroskedasticity-robust t statistic (5). The former are shown as solid, dashed, or dotted lines, and the latter are shown as symbols.

Several striking results emerge from Figure 1. In all cases, there is generally not much to choose between the results for $t_s(\hat{\beta}, \beta_0)$ and the results for $t_h(\hat{\beta}, \beta_0)$. This is not surprising, since the disturbances are actually homoskedastic. Everything else we say about these results applies equally to both test statistics.

It is clear from the top left-hand panel that the older bootstrap methods (namely, the pairs, UR, and RR bootstraps) can overreject very severely when ρ is large and a is not large, although, in this case, they do always work better than the asymptotic test. In contrast, the top right-hand panel shows that the new, efficient bootstrap methods (namely, the RE and WRE bootstraps) all tend to underreject slightly in the same case. This problem is more pronounced for RE than for WRE.

The two bottom panels show that, when ρ is small, things can be very different. The asymptotic test now underrejects modestly for small values of a , the pairs and UR bootstraps overreject quite severely, and the RR bootstrap underrejects a bit less than the asymptotic test. This is a case in which bootstrap tests can evidently be much less reliable than asymptotic ones. As can be seen from the bottom right-hand panel, the efficient bootstrap methods generally perform much better than the older ones. There are only modest differences between the rejection frequencies for WRE and RE, with the former being slightly less prone to underreject for small values of a .

It is evident from the bottom right-hand panel of Figure 1 that the RR, RE, and WRE bootstraps perform almost the same when $\rho = 0.1$, even when the instruments are weak. This makes sense, because there is little efficiency to be gained by running regression (15) instead of regression (2) when ρ is small. Thus we can expect the RE and RR bootstrap DGPs to be quite similar whenever the correlation between the reduced-form and structural disturbances is small.

Figure 2 shows the effects of varying ρ from 0 to 0.95 by increments of 0.05. In the top two panels, $a = 2$, so that the instruments are quite weak, and, in the bottom

two panels, $a = 8$, so that they are moderately strong. As in Figure 1, the two left-hand panels show rejection frequencies for older methods that often work poorly. We see that the asymptotic test tends to overreject severely, except when ρ is close to 0, that the pairs and UR bootstraps always overreject, and that the RR bootstrap almost always performs better than the pairs and UR bootstraps. However, even it overrejects severely when ρ is large.

As in Figure 1, the two right-hand panels in Figure 2 show results for the new, efficient bootstrap methods, as well as partial ones for the RR bootstrap for purposes of comparison. Note the different vertical scales. The new methods all work reasonably well when $a = 2$ and very well, although not quite perfectly, when $a = 8$. Once again, it seems that WRE works a little bit better than RE.

In the first two sets of experiments, the number of instruments was fairly large, with $l - k = 11$. Of course, different choices for this number would have produced different results. In Figure 3, $l - k$ varies from 1 to 21. In the top two panels, $a = 2$ and $\rho = 0.9$; in the bottom two, $a = 2$ and $\rho = 0.1$. Since a is quite small, all the tests perform relatively poorly. As before, the new bootstrap tests generally perform very much better than the older ones, although, as expected, RR is almost indistinguishable from RE when $\rho = 0.1$.

When $\rho = 0.9$, the performance of the asymptotic test and the older bootstrap tests deteriorates dramatically as $l - k$ increases. This is not evident when $\rho = 0.1$, however. In contrast, the performance of the efficient bootstrap tests actually tends to improve as $l - k$ increases. The only disturbing result is in the top right-hand panel, where the RE and WRE bootstrap tests underreject fairly severely when $l = k \leq 3$, that is, when there are two or fewer overidentifying restrictions. The rest of our experiments do not deal with this case, and so they may not accurately reflect what happens when the number of instruments is very small.

In all the experiments discussed so far, $n = 400$. It makes sense to use a reasonably large number, because cross-section data sets with weak instruments are often fairly large. However, using a very large number would have greatly raised the cost of the experiments. Using larger values of n while holding a fixed would not necessarily cause any of the tests to perform better, because, in theory, rejection frequencies only approach nominal levels as both n and a tend to infinity. Nevertheless, it is of interest to see what happens as n changes while we hold a fixed.

Figure 4 shows how the efficient bootstrap methods perform in four cases ($a = 2$ or $a = 8$, and $\rho = 0.1$ or $\rho = 0.9$) for sample sizes that increase from 25 to 1600 by factors of approximately $\sqrt{2}$. Note that, as n increases, the instruments become very weak indeed when $a = 2$. For $n = 1600$, the R^2 of the reduced-form regression (22) in the DGP, evaluated at the true parameter values, is just 0.0025. Even when $a = 8$, it is just 0.0385.

The results in Figure 4 are striking. Both the efficient bootstrap methods perform better for $n = 1600$ than for $n = 25$, often very much better. As n increases from 25

to about 200, the performance of the tests often changes quite noticeably. However, their performance never changes very much as n is increased beyond 400, which is why we used that figure in most of the experiments. When possible, the figure includes rejection frequencies for RR. Interestingly, when $\rho = 0.1$, it actually outperforms RE for very small sample sizes, although its performance is almost indistinguishable from that of RE for $n \geq 70$.

Up to this point, we have reported results only for equal-tail bootstrap tests, that is, ones based on the equal-tail P value (8). We believe that these are more attractive in the context of IV estimation than tests based on the symmetric P value (7), because IV estimates can be severely biased when the instruments are weak. However, it is important to point out that results for symmetric bootstrap tests would have differed, in some ways substantially, from the ones reported for equal-tail tests.

Figure 5 is comparable to Figure 2. Like them, it shows rejection frequencies as functions of ρ for $n = 400$ and $l - k = 11$, with $a = 2$ in the top row and $a = 8$ in the bottom row, but this time for symmetric bootstrap tests. Comparing the top left-hand panels of Figure 5 and Figure 2, we see that, instead of overrejecting, symmetric bootstrap tests based on the pairs and UR bootstraps underreject severely when the instruments are weak and ρ is small, although they overreject even more severely than equal-tail tests when ρ is very large. Results for the RR bootstrap are much less different, but the symmetric version underrejects a little bit more than the equal-tail version for small values of ρ and overrejects somewhat more for large values.

As one would expect, the differences between symmetric and equal-tail tests based on the new, efficient bootstrap methods are much less dramatic than the differences for the pairs and UR bootstraps. At first glance, this statement may appear to be false, because the two right-hand panels in Figure 5 look quite different from the corresponding ones in Figure 2. However, it is important to bear in mind that the vertical axes in the right-hand panels are highly magnified. The actual differences in rejection frequencies are fairly modest. Overall, the equal-tail tests seem to perform better than the symmetric ones, and they are less sensitive to the values of ρ , which further justifies our choice to focus on them.

Next, we turn our attention to heteroskedasticity. One big advantage of the t test over competing tests such as the AR, K , and CLR tests is that it can easily be modified to be robust to heteroskedasticity of unknown form; recall (5). The major advantage of the WRE over the RE bootstrap is that the former accounts for heteroskedasticity in the bootstrap DGP and the latter does not. Thus it is of considerable interest to see how the various tests perform when there is heteroskedasticity.

In principle, heteroskedasticity can manifest itself in a number of ways. However, because there is only one exogenous variable that actually matters in the DGP given

by (21) and (22), there are not many obvious ways to model it without using a more complicated model. In our first set of experiments, we used the DGP

$$\mathbf{y}_1 = \beta \mathbf{y}_2 + n^{1/2} |\mathbf{w}_1| * \mathbf{u}_1 \quad (23)$$

$$\mathbf{y}_2 = a \mathbf{w}_1 + \mathbf{u}_2, \quad \mathbf{u}_2 = \rho n^{1/2} |\mathbf{w}_1| * \mathbf{u}_1 + r \mathbf{v}, \quad (24)$$

where, as before, \mathbf{u}_1 and \mathbf{v} contain independent, standard normal elements. The purpose of the factor $n^{1/2}$ is to rescale the instrument so that its squared length is n instead of 1. Thus each element of \mathbf{u}_1 is multiplied by the absolute value of the corresponding element of \mathbf{w}_1 , appropriately rescaled.

We investigated rejection frequencies as a function of ρ for this DGP for two values of a , namely, $a = 2$ and $a = 8$. These results are comparable to those in Figure 2. Results for the new, efficient bootstrap methods only are reported in Figure 6. The left-hand panels contain results for $a = 2$, and the right-hand panels for $a = 8$. The top two panels show results for both bootstrap methods for both $t_s(\hat{\beta}, \beta_0)$ and $t_h(\hat{\beta}, \beta_0)$. The bottom two panels omit the results for the RE bootstrap applied to $t_s(\hat{\beta}, \beta_0)$. This allows the vertical axis to have a much larger scale and makes it far easier to distinguish among the results for the other methods.

The most striking result in Figure 6 is that using RE, the bootstrap method which does not allow for heteroskedasticity, along with the test statistic $t_s(\hat{\beta}, \beta_0)$, which requires homoskedasticity, often leads to severe overrejection. Of course, this is hardly a surprise. But the result is a bit more interesting if we express it in another way. Using *either* WRE, the bootstrap method which allows for heteroskedasticity, *or* the test statistic $t_h(\hat{\beta}, \beta_0)$, which is valid in the presence of heteroskedasticity of unknown form, generally seems to produce rejection frequencies that are reasonably close to nominal levels.

It seems plausible that the best results would be obtained when we combine WRE with $t_h(\hat{\beta}, \beta_0)$. This is clearly so when $a = 8$. In this case, WRE performs exceedingly well, overrejecting just slightly for all values of ρ . However, when $a = 2$, it is difficult to draw any strong conclusions. Both bootstrap methods perform quite well when applied to $t_h(\hat{\beta}, \beta_0)$, and WRE also performs quite well when applied to $t_s(\hat{\beta}, \beta_0)$.

We also performed a second set of experiments in which the DGP was similar to (23) and (24), except that each element of \mathbf{u}_1 was multiplied by $n^{1/2} w_{1i}^2$ instead of by $n^{1/2} |w_{1i}|$. Thus the heteroskedasticity was considerably more extreme. Results are not shown, because they were quite similar to those in Figure 6. WRE performed exceedingly well when applied to $t_h(\hat{\beta}, \beta_0)$, and RE also performed quite well in that case. However, RE performed very poorly indeed when applied to $t_s(\hat{\beta}, \beta_0)$.

Taken together, our results for both the homoskedastic and heteroskedastic cases suggest that, at least when the sample size is moderate to large (say, 200 or more), the safest approach is probably to use the WRE bootstrap with the robust statistic $t_h(\hat{\beta}, \beta_0)$. In the absence of heteroskedasticity, this should yield results very similar to using the RE bootstrap with $t_s(\hat{\beta}, \beta_0)$, but the latter approach can be very seriously misleading when heteroskedasticity is present.

4. More than Two Endogenous Variables

Up to this point, as in Davidson and MacKinnon (2006), we have focused on the case in which there is just one endogenous variable on the right-hand side. The K and CLR tests are designed to handle only this special case. However, there is no such restriction for t statistics, and the RE and WRE bootstraps can easily be extended to handle more general situations.

For notational simplicity, we deal with the case in which there are just two endogenous variables on the right-hand side. It is trivial to extend the analysis to handle any number of them. The model of interest is

$$\mathbf{y}_1 = \beta_2 \mathbf{y}_2 + \beta_3 \mathbf{y}_3 + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}_1 \quad (25)$$

$$\mathbf{y}_2 = \mathbf{W}\boldsymbol{\pi}_2 + \mathbf{u}_2 \quad (26)$$

$$\mathbf{y}_3 = \mathbf{W}\boldsymbol{\pi}_3 + \mathbf{u}_3, \quad (27)$$

where the notation should be obvious. As before, \mathbf{Z} and \mathbf{W} are, respectively, an $n \times k$ and an $n \times l$ matrix of exogenous variables with the property that $\mathcal{S}(\mathbf{Z})$ lies in $\mathcal{S}(\mathbf{W})$. For identification, we require that $l \geq k + 2$.

The pairs and UR bootstraps require no discussion. The RR bootstrap is also quite easy to implement in this case. To test the hypothesis that, say, $\beta_2 = \beta_{20}$, we need to estimate by 2SLS a restricted version of equation (25),

$$\mathbf{y}_1 - \beta_{20} \mathbf{y}_2 = \beta_3 \mathbf{y}_3 + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}_1, \quad (28)$$

in which \mathbf{y}_3 is the only endogenous right-hand side variable, so as to yield restricted estimates $\tilde{\beta}_3$ and $\tilde{\boldsymbol{\gamma}}$ and 2SLS residuals $\tilde{\mathbf{u}}_1$. We also estimate equations (26) and (27) by OLS, as usual. Then the bootstrap DGP is

$$\begin{aligned} y_{i1}^* - \beta_{20} y_{i2}^* &= \tilde{\beta}_3 y_{i3}^* + \mathbf{Z}_i \tilde{\boldsymbol{\gamma}} + \tilde{u}_{1i}^* \\ y_{2i}^* &= \mathbf{W}_i \hat{\boldsymbol{\pi}}_2 + \hat{u}_{2i}^* \\ y_{3i}^* &= \mathbf{W}_i \hat{\boldsymbol{\pi}}_3 + \hat{u}_{3i}^*, \end{aligned} \quad (29)$$

where the bootstrap disturbances are generated as follows:

$$\begin{bmatrix} \tilde{u}_{1i}^* \\ \hat{u}_{2i}^* \\ \hat{u}_{3i}^* \end{bmatrix} \sim \text{EDF} \left(\begin{array}{c} \tilde{u}_{1i} \\ (n/(n-l))^{1/2} \hat{u}_{2i} \\ (n/(n-l))^{1/2} \hat{u}_{3i} \end{array} \right). \quad (30)$$

As before, we may omit the term $\mathbf{Z}_i \tilde{\boldsymbol{\gamma}}$ from the first of equations (29). In (30), we rescale the OLS residuals from the two reduced-form equations but not the 2SLS ones from equation (28), although this is not essential.

For the RE and WRE bootstraps, we need to re-estimate equations (26) and (27) so as to obtain more efficient estimates that are asymptotically equivalent to 3SLS. We do so by estimating the analogs of regression (15) for these two equations, which are

$$\begin{aligned} \mathbf{y}_2 &= \mathbf{W}\boldsymbol{\pi}_2 + \delta_2\tilde{\mathbf{u}}_1 + \text{residuals, and} \\ \mathbf{y}_3 &= \mathbf{W}\boldsymbol{\pi}_3 + \delta_3\tilde{\mathbf{u}}_1 + \text{residuals.} \end{aligned}$$

We then use the OLS estimates $\tilde{\boldsymbol{\pi}}_2$ and $\tilde{\boldsymbol{\pi}}_3$ and the residuals $\tilde{\mathbf{u}}_2 \equiv \mathbf{y}_2 - \mathbf{W}\tilde{\boldsymbol{\pi}}_2$ and $\tilde{\mathbf{u}}_3 \equiv \mathbf{y}_3 - \mathbf{W}\tilde{\boldsymbol{\pi}}_3$ in the RE and WRE bootstrap DGPs:

$$\begin{aligned} y_{i1}^* - \beta_{20}y_{i2}^* &= \tilde{\beta}_3 y_{i3}^* + \mathbf{Z}_i \tilde{\boldsymbol{\gamma}} + \tilde{u}_{1i}^* \\ y_{2i}^* &= \mathbf{W}_i \tilde{\boldsymbol{\pi}}_2 + \tilde{u}_{2i}^* \\ y_{3i}^* &= \mathbf{W}_i \tilde{\boldsymbol{\pi}}_3 + \tilde{u}_{3i}^*. \end{aligned} \tag{31}$$

Only the second and third equations of (31) differ from the corresponding equations of (29) for the RR bootstrap. In the case of the RE bootstrap, we resample from triples of (rescaled) residuals:

$$\begin{bmatrix} \tilde{u}_{1i}^* \\ \tilde{u}_{2i}^* \\ \tilde{u}_{3i}^* \end{bmatrix} \sim \text{EDF} \left(\begin{array}{c} \tilde{u}_{1i} \\ (n/(n-l))^{1/2} \tilde{u}_{2i} \\ (n/(n-l))^{1/2} \tilde{u}_{3i} \end{array} \right).$$

In the case of the WRE bootstrap, we use the analog of (19), which is

$$\begin{bmatrix} \tilde{u}_{1i}^* \\ \tilde{u}_{2i}^* \\ \tilde{u}_{3i}^* \end{bmatrix} = \begin{bmatrix} \tilde{u}_{1i} v_i^* \\ (n/(n-l))^{1/2} \tilde{u}_{2i} v_i^* \\ (n/(n-l))^{1/2} \tilde{u}_{3i} v_i^* \end{bmatrix},$$

where v_i^* is a suitable random variable with mean 0 and variance 1.

5. Bootstrap Confidence Intervals

A confidence interval for a parameter is constructed, implicitly or explicitly, by inverting a test. We may always test whether any given parameter value is the true value. The upper and lower limits of the confidence interval are those values for which the test statistic equals its critical value. Equivalently, for an interval with nominal coverage $1 - \alpha$, they are the parameter values for which the P value of the test equals α for a two-tailed test. For an elementary exposition, see Davidson and MacKinnon (2004, Chapter 5).

There is a vast literature on bootstrap confidence intervals; Davison and Hinkley (1997) provides a good introduction. The type that is most widely used in econometrics is the **percentile t** , or **bootstrap t** , interval. Percentile t intervals could easily be constructed

using the pairs or UR bootstraps, for which the bootstrap DGP does not impose the null hypothesis, but they would certainly work badly whenever bootstrap tests based on these methods work badly, that is, whenever ρ is not small and a is not large; see Figures 1, 2, 3, and 5.

It is conceptually easy, although perhaps computationally demanding, to construct confidence intervals using bootstrap methods that do impose the null hypothesis. We now explain precisely how to construct such an interval with nominal coverage $1 - \alpha$. The method we propose can be used with any bootstrap DGP that imposes the null hypothesis, including the RE and WRE bootstraps. It can be expected to work well whenever the rejection frequencies for tests at level α based on the relevant bootstrap method are in fact close to α .

1. Estimate the model (1) and (2) by 2SLS so as to obtain the IV estimate $\hat{\beta}$ and either the ordinary standard error $s_s(\hat{\beta})$ or the heteroskedasticity-robust standard error $s_h(\hat{\beta})$ defined in (6). Our simulation results suggest that there is no significant cost to using the latter, even when the disturbances are homoskedastic, for the sample sizes typically encountered with cross-section data.
2. Write a routine that, for any value of β , say β_0 , constructs the test statistic $(\hat{\beta} - \beta_0)/s_h(\hat{\beta})$, or possibly $(\hat{\beta} - \beta_0)/s_s(\hat{\beta})$, and bootstraps it under the null hypothesis that $\beta = \beta_0$. This routine will perform B bootstrap replications using a random number generator that depends on a seed m to calculate an equal-tail bootstrap P value, say $p^*(\beta_0)$, using equation (8).
3. Choose a reasonably large value of B such that $\alpha(B + 1)$ is an integer, and also choose m . The same values of m and B must be used each time $p^*(\beta_0)$ is calculated. This is very important, since otherwise a given value of β_0 would yield different values of $p^*(\beta_0)$ each time it is evaluated.
4. For the lower limit of the confidence interval, find two values of β , say β_{1-} and β_{1+} , with $\beta_{1-} < \beta_{1+}$, such that $p^*(\beta_{1-}) < \alpha$ and $p^*(\beta_{1+}) > \alpha$. Since both values will normally be less than $\hat{\beta}$, one obvious way to do this is to start at the lower limit of an asymptotic confidence interval, say β_1^∞ , and see whether $p^*(\beta_1)$ is greater or less than α . If it is less than α , then β_1^∞ can serve as β_{1-} ; if it is greater, then β_1^∞ can serve as β_{1+} . Whichever of β_{1-} and β_{1+} has not been found in this way can then be obtained by moving a moderate distance, perhaps $s_h(\hat{\beta})$, in the appropriate direction as many times as necessary, each time checking whether the bootstrap P value is on the desired side of α .
5. Similarly, find two values of β , say β_{u-} and β_{u+} , with $\beta_{u-} < \beta_{u+}$, such that $p^*(\beta_{u-}) > \alpha$ and $p^*(\beta_{u+}) < \alpha$.
6. Find the lower limit of the confidence interval, β_1^* . This is a value between β_{1-} and β_{1+} which is such that $p^*(\beta_1^*) \cong \alpha$. One way to find β_1^* is to minimize the function $(p^*(\beta) - \alpha)^2$ with respect to β in the interval $[\beta_{1-}, \beta_{1+}]$ by using golden section search. This method is attractive because it is guaranteed to converge to a local minimum and does not require derivatives

7. In the same way, find the upper limit of the confidence interval, β_u^* . This is a value between β_{u-} and β_{u+} which is such that $p^*(\beta_u^*) \cong \alpha$.

When a confidence interval is constructed in this way, the limits of the interval have the property that $p^*(\beta_l^*) \cong p^*(\beta_u^*) \cong \alpha$. The approximate equalities here would become exact, subject to the termination criterion for the golden search routine, if B were allowed to tend to infinity. The problem is that $p^*(\beta)$ is a step function, the value of which changes by precisely $1/B$ at certain points as its argument varies. This suggests that B should be fairly large, if possible.

6. An Empirical Example

The method of instrumental variables is routinely used to answer empirical questions in labor economics. In such applications, it is common to employ fairly large cross-section datasets for which the instruments are very weak. In this section, we apply our methods to an empirical example of this type. It uses the same data as Card (1995). The dependent variable in the structural equation is the log of wages for young men in 1976, and the other endogenous variable is years of schooling. There are 3610 observations, which originally came from the Young Men Cohort of the National Longitudinal Survey.

Although we use Card's data, the equation we estimate is not identical to any of the ones he estimates. We simplify the specification by omitting a large number of exogenous variables having to do with location and family characteristics, which appear to be collectively insignificant, at least in the IV regression. We also use age and age squared instead of experience and experience squared in the wage equation. As Card notes, experience is endogenous if schooling is endogenous. In some specifications, he therefore uses age and age squared as instruments. For purposes of illustrating the methods discussed in this paper, it is preferable to have just two endogenous variables in the model, and so we do not use experience as an endogenous regressor. This slightly improves the fit of the IV regression, but it also has some effect on the coefficient of interest. In addition to age and age squared, the structural equation includes a constant term and dummies for race, living in a southern state, and living in an SMSA as exogenous variables.

We use four instruments, all of which are dummy variables. The first is 1 if there is a two-year college in the local labor market, the second if there is either a two-year college or a four-year college, the third if there is a public four-year college, and the fourth if there is a private four-year college. The second instrument was not used by Card, although it is computed as the product of two instruments that he did use. The instruments are fairly weak, but apparently not as weak as they were in many of our simulations. The concentration parameter is estimated to be just 19.92, which is equivalent to $a = 4.46$. Of course, this is just an estimate, and a fairly noisy one.

Our estimate of the coefficient β , which is the effect of an additional year of schooling on the log wage, is 0.1150. This is higher than some of the results reported by Card and

Table 1. Confidence Intervals for β

Method	Test Statistic	Lower Limit	Upper Limit
Asymptotic	t_s	0.0399	0.1901
	t_h	0.0388	0.1913
RE Bootstrap	t_s	0.0497	0.3200
WRE Bootstrap	t_h	0.0500	0.3438

lower than others. The standard error is either 0.0384 (assuming homoskedasticity) or 0.0389 (robust to heteroskedasticity). Thus the t statistics for the coefficient β to be zero and the corresponding asymptotic P values are:

$$t_s = 2.999 \quad (p = 0.0027) \quad \text{and} \quad t_h = 2.958 \quad (p = 0.0031).$$

Equal-tail bootstrap P values are very similar to the asymptotic ones. Based on $B = 99,999$, the P value is 0.0021 for both the RE bootstrap using t_s and the WRE bootstrap using t_h .

Up to this point, our bootstrap results merely confirm the asymptotic ones, which suggest that the coefficient on schooling is almost certainly positive. Thus they might incorrectly be taken to show that asymptotic inference is reliable in this case. In fact, it is not. With a fairly low value of a and a reasonably large value of ρ (the correlation between the residuals from the structural and reduced-form equations is -0.474), our simulation results suggest that asymptotic theory should not perform very well in this case. Indeed, it does not, as becomes clear when we examine bootstrap confidence intervals.

We constructed two different 0.95 confidence intervals for β using the procedure of the previous section. The interval based on the RE bootstrap used t_s , and the interval based on the WRE bootstrap used t_h . In order to minimize the impact of the specific random numbers that were used, both intervals were based on $B = 99,999$. Each of them required the calculation of 40 bootstrap P values, mostly during the golden section search. Computing each bootstrap interval took about 26 minutes on a Linux machine with an Intel Core 2 Duo E6600 processor.

It can be seen from Table 1 that the lower limits of the bootstrap intervals are moderately higher than the lower limits of the asymptotic intervals, and the upper limits are very much higher. What seems to be happening is that $\hat{\beta}$ is biased downwards, because $\rho < 0$, and the standard errors are also too small. These two effects almost offset each other when we test the hypothesis that $\beta = 0$, which is why the asymptotic and bootstrap tests yield such similar results. However, they do not fully offset each other for the tests that determine the lower limit of the confidence interval, and they reinforce each other for the tests that determine the upper limit.

It may seem disappointing that the bootstrap confidence intervals in Table 1 are so wide. This is a consequence of the model and the data, not the bootstrap methods themselves. With stronger instruments, the estimates would be more precise, the confidence intervals would be narrower, and the differences between bootstrap and asymptotic intervals would be less pronounced.

7. Conclusion

In this paper, we propose a new bootstrap method for models estimated by instrumental variables. It is a wild bootstrap variant of the RE bootstrap proposed in Davidson and MacKinnon (2006). The most important features of this method are that it uses efficient estimates of the reduced-form equation(s) and that it allows for heteroskedasticity of unknown form.

In an extensive simulation study, we apply the new and existing bootstrap methods to t statistics, which may be robust to heteroskedasticity of unknown form, for the coefficient of a single endogenous variable. The WRE bootstrap could also be used with other statistics as well, but we do not investigate this possibility because most of them are not robust to heteroskedasticity. We find that, like the RE bootstrap, the new WRE bootstrap performs very much better than earlier bootstrap methods, especially when the instruments are weak.

We also show how to apply the RE and WRE bootstraps to models with two or more endogenous variables on the right-hand side, but their performance in this context remains a topic for future research. In addition, we discuss how to construct confidence intervals by inverting bootstrap tests based on bootstrap DGPs that impose the null hypothesis, such as the RE and WRE bootstraps.

Finally, we apply the efficient bootstrap methods discussed in this paper to an empirical example that involves a fairly large sample but weak instruments. When used to test the null hypothesis that years of schooling do not affect wages, the new bootstrap tests merely confirm the results of asymptotic tests. However, when used to construct confidence intervals, they yield intervals that differ radically from conventional ones based on asymptotic theory.

References

- Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). “Optimal two-sided invariant similar tests for instrumental variables regression”, *Econometrica*, 74, 715–752.
- Anderson, T. W. and H. Rubin (1949). “Estimation of the parameters of a single equation in a complete system of stochastic equations”, *Annals of Mathematical Statistics*, 20, 46–63.
- Beran, R. (1988). “Prepivoting test statistics: A bootstrap view of asymptotic refinements”, *Journal of the American Statistical Association*, 83, 687–697.
- Card, D. (1995). “Using geographic variation in college proximity to estimate the return to schooling”, in Christofides, L. N., E. K. Grant, and R. Swidinsky (eds.), *Aspects of Labour Market Behaviour: Essays in Honour of J. Vanderkamp*, Toronto, University of Toronto Press.
- Davidson, R., and E. Flachaire (2001). “The wild bootstrap, tamed at last”, Queen’s Economics Department Working Paper No. 1000.
- Davidson, R., and J. G. MacKinnon (1999). “The size distortion of bootstrap tests”, *Econometric Theory*, 15, 361–376.
- Davidson, R., and J. G. MacKinnon (2004). *Econometric Theory and Methods*, New York, Oxford University Press.
- Davidson, R., and J. G. MacKinnon (2006). “Bootstrap inference in a linear equation estimated by instrumental variables”, Queen’s Economics Department Working Paper No. 1024.
- Davison, A. C., and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*, Cambridge, Cambridge University Press.
- Dufour, J.-M., and M. Taamouti (2005). “Projection-based statistical inference in linear structural models with possibly weak instruments”, *Econometrica*, 73, 1351–1365.
- Dufour, J.-M., and M. Taamouti (2007). “Further results on projection-based inference in IV regressions with weak, collinear or missing instruments”, *Journal of Econometrics*, 139, 133–153.
- Flores-Lagunes, A. (2007). “Finite sample evidence of IV estimators under weak instruments”, *Journal of Applied Econometrics*, 22, 677–694.
- Freedman, D. A. (1981). “Bootstrapping regression models”, *Annals of Statistics*, 9, 1218–1228.
- Freedman, D. A. (1984). “On bootstrapping two-stage least-squares estimates in stationary linear models”, *Annals of Statistics*, 12, 827–842.

- Gonçalves, S., and L. Kilian (2004). “Bootstrapping autoregressions with heteroskedasticity of unknown form”, *Journal of Econometrics*, 123, 89–120.
- Kleibergen, F. (2002). “Pivotal statistics for testing structural parameters in instrumental variables regression”, *Econometrica*, 70, 1781–1803.
- MacKinnon, J. G. (2006). “Bootstrap methods in econometrics”, *Economic Record*, 82, S2-S18.
- Moreira, M. J. (2003). “A conditional likelihood ratio test for structural models”, *Econometrica*, 71, 1027–1048.
- Moreira, M. J., J. R. Porter, and G. A. Suarez (2005). “Bootstrap and higher-order expansion validity when instruments may be weak”, NBER Working Paper No. 302, revised.
- Staiger, D., and J. H. Stock (1997). “Instrumental Variables Regression with Weak Instruments”, *Econometrica*, 65, 557–586.
- Wu, C. F. J. (1986). “Jackknife, bootstrap and other resampling methods in regression analysis”, *Annals of Statistics*, 14, 1261–1295.

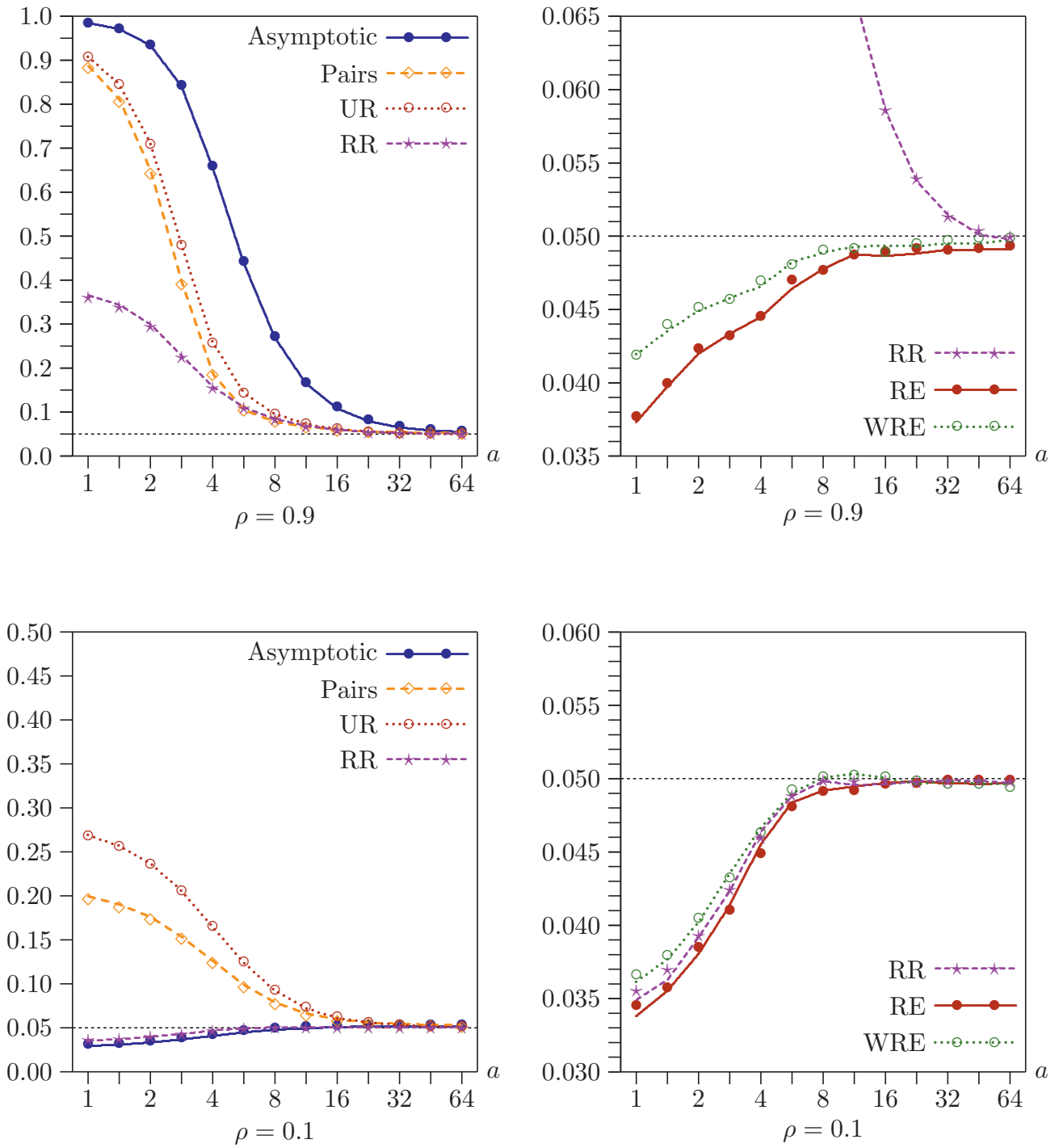


Figure 1. Rejection frequencies as functions of a for $l - k = 11$, $n = 400$

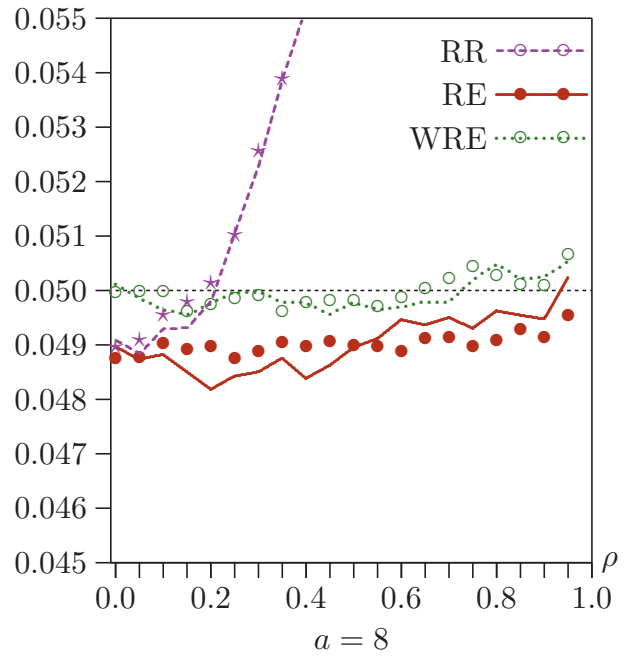
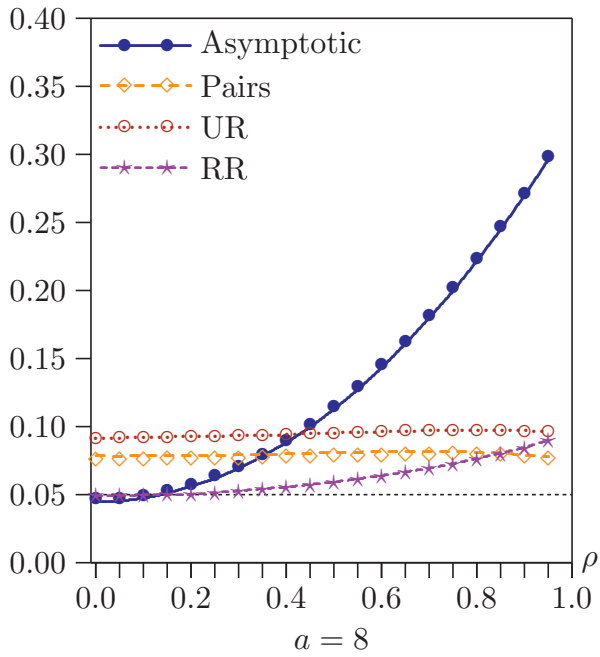
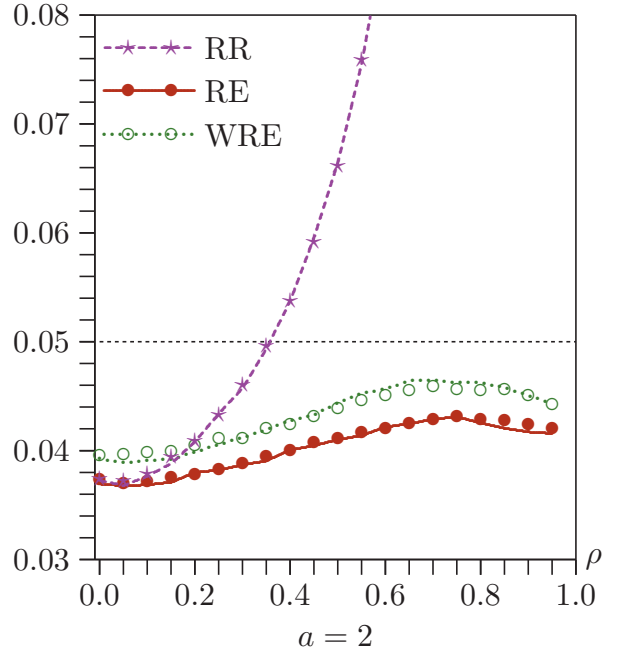
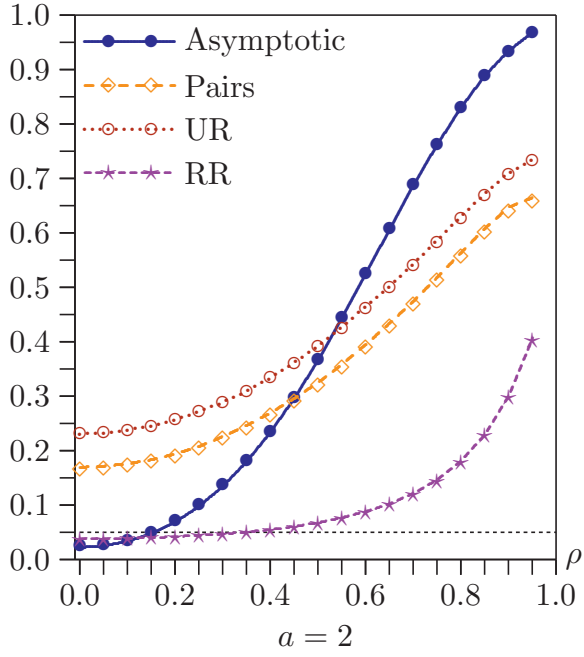


Figure 2. Rejection frequencies as functions of ρ for $l - k = 11$, $n = 400$

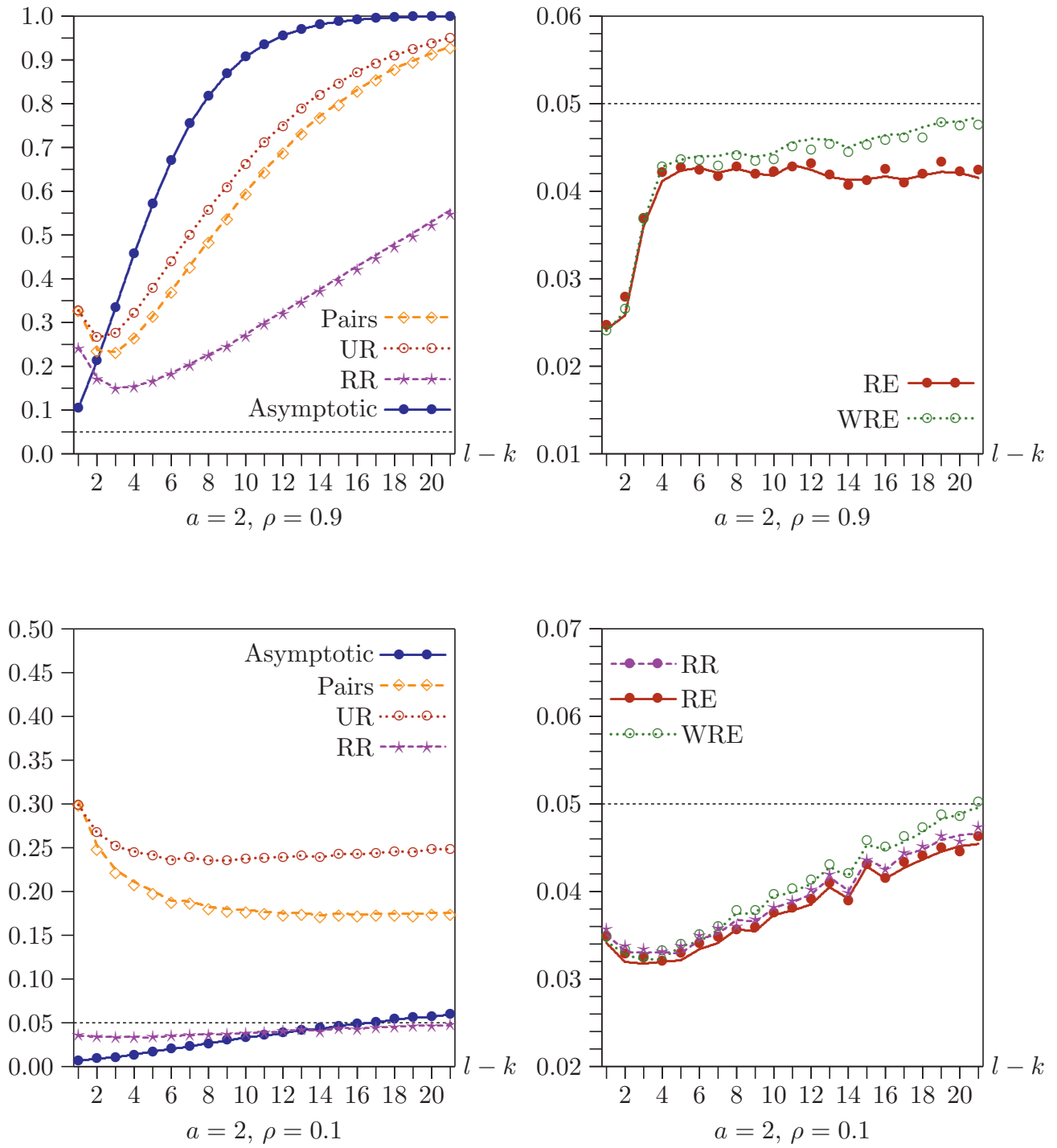


Figure 3. Rejection frequencies as functions of $l - k$ for $n = 400$

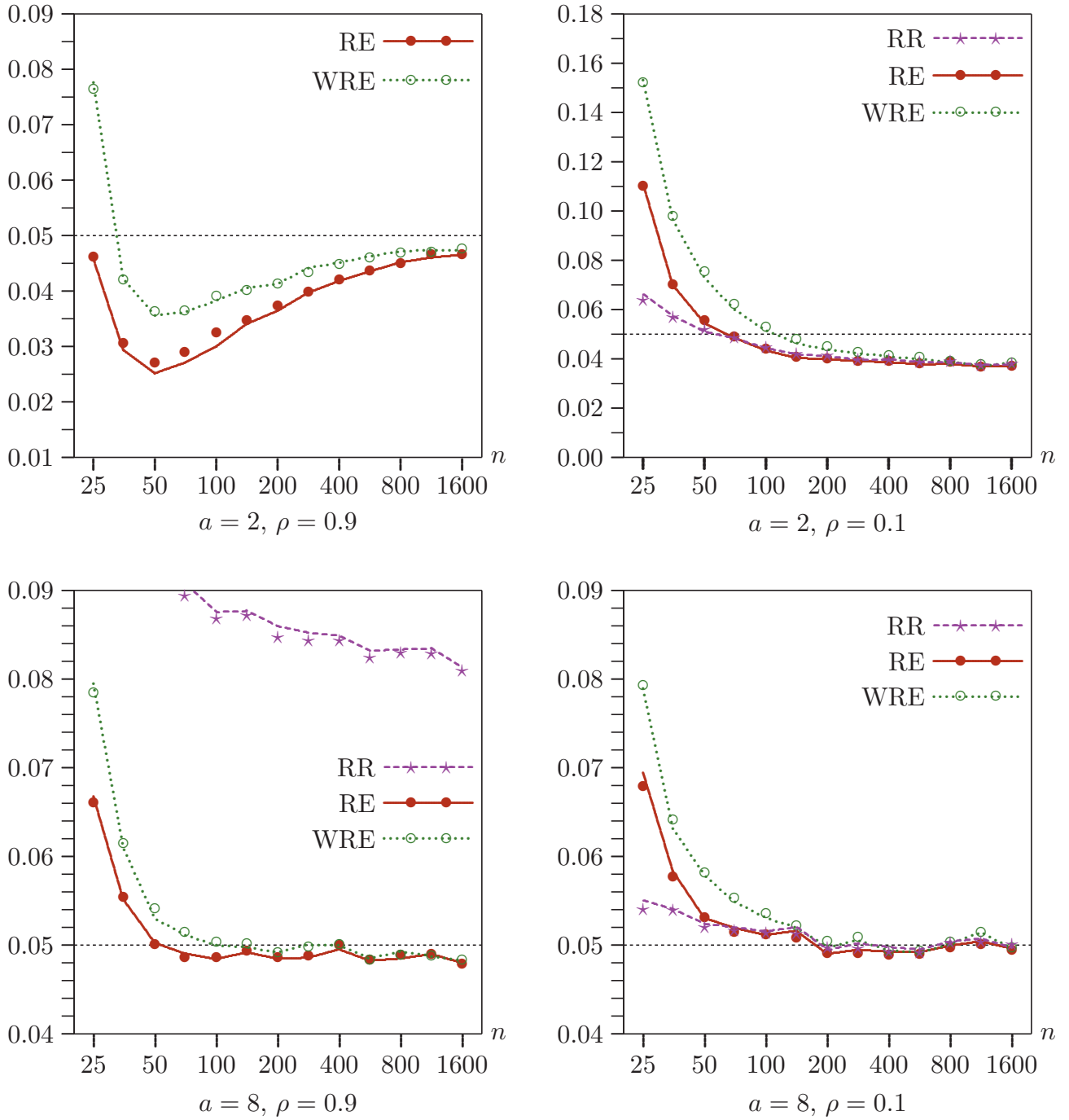


Figure 4. Rejection frequencies as functions of n for $k-l=11$

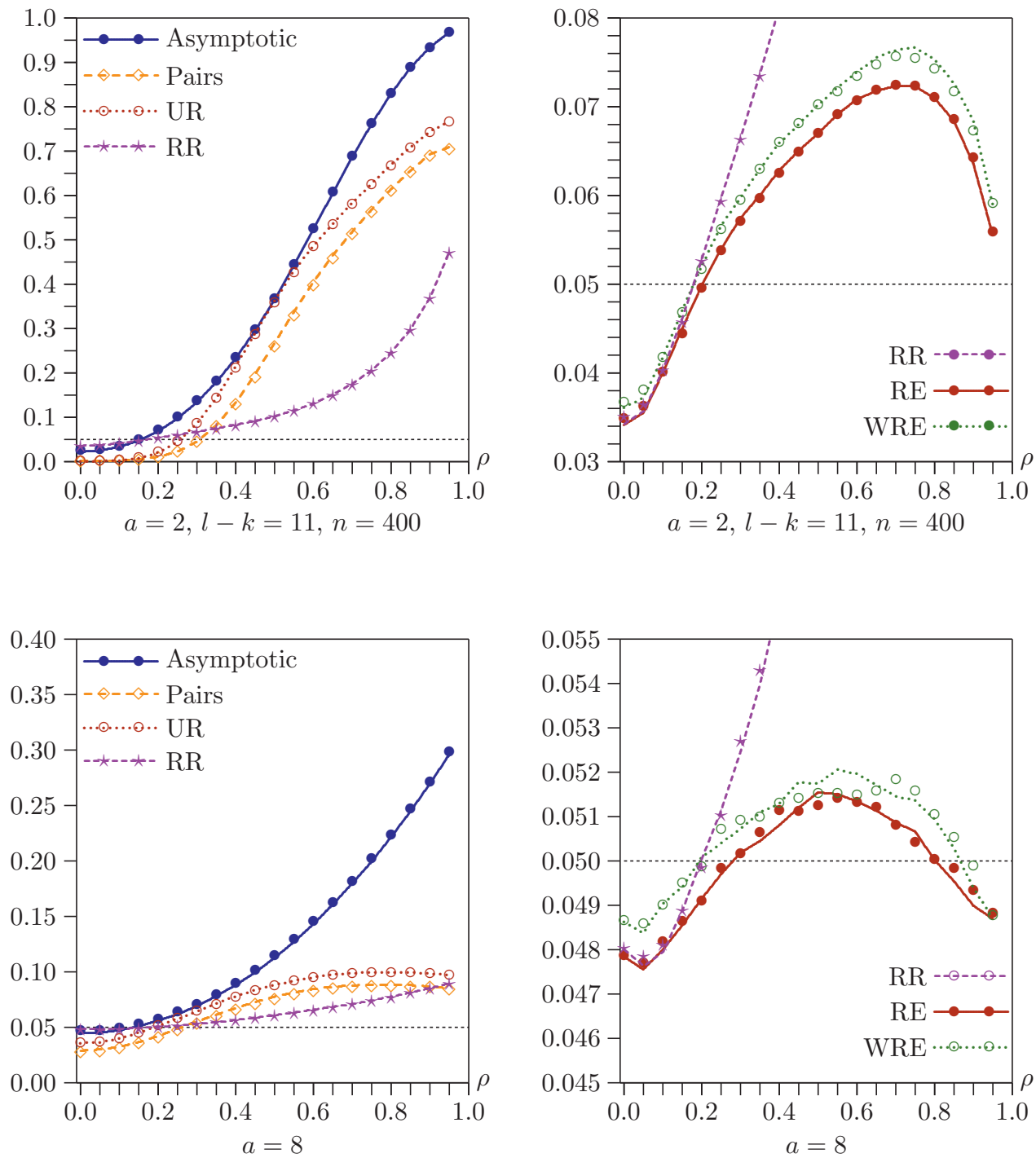


Figure 5. Rejection frequencies as functions of ρ for symmetric bootstrap tests

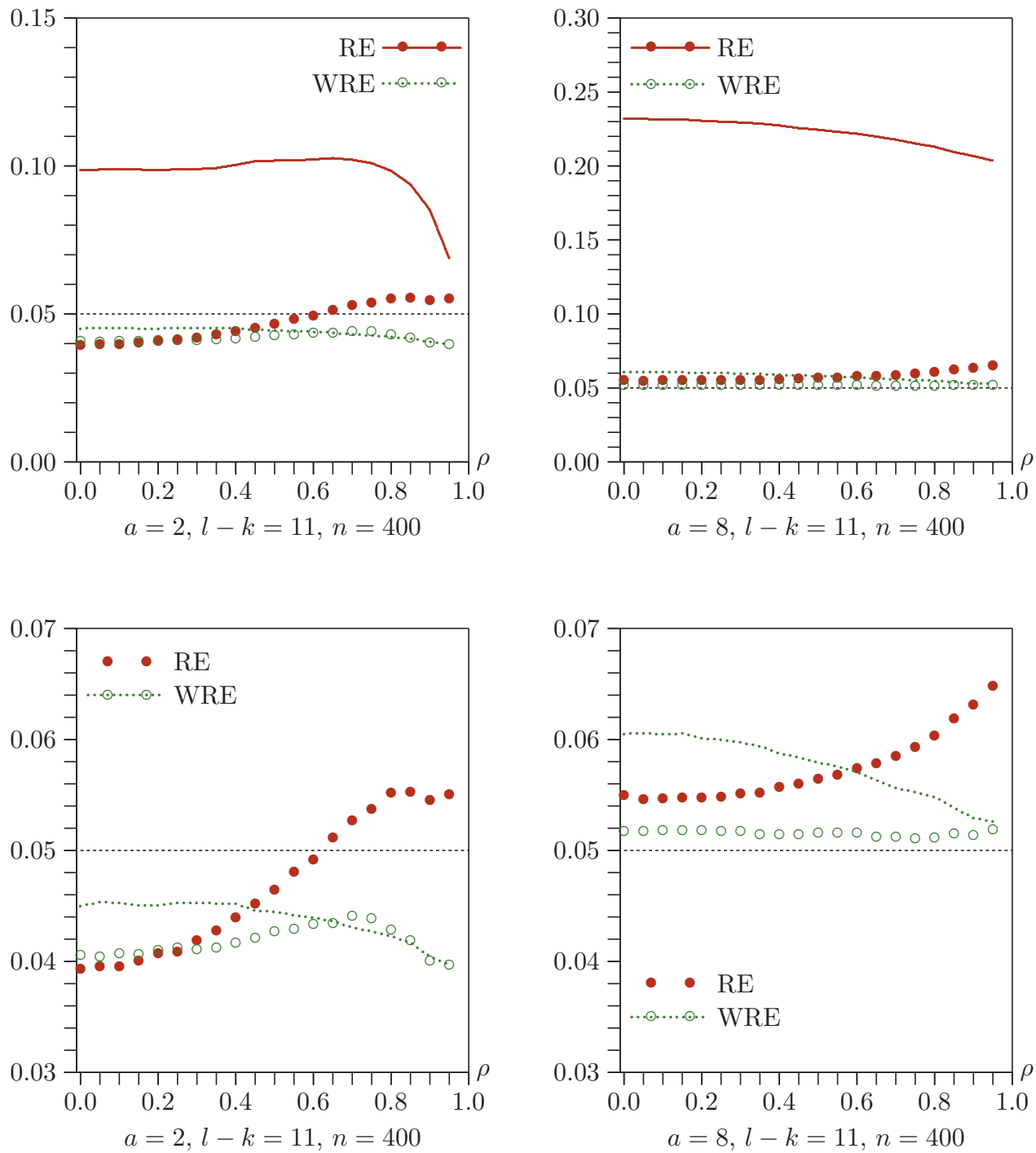


Figure 6. Rejection frequencies as functions of ρ when error terms are heteroskedastic