UvA ECONOMETRICS

Discussion Paper: 2011/08

Bootstrapping Subset Test Statistics in IV Regression

Noud P.A. van Giersbergen

www.ase.uva.nl/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Valckenierstraat 65-67
1018 XE AMSTERDAM
The Netherlands

UvA UNIVERSITEIT VAN AMSTERDAM

Bootstrapping Subset Test Statistics in IV Regression

Noud P.A. van Giersbergen*
Department of Quantitative Economics,
Amsterdam School of Economics
University of Amsterdam
Valckenierstraat 65-67
1018 XE Amsterdam
The Netherlands
E-mail: N.P.A.vanGiersbergen@uva.nl

February 21, 2012

Abstract

The finite-sample performance of various bootstrap procedures is studied by simulation in a linear regression model containing 2 endogenous regressors. Besides several residual-based bootstrap procedures, we also consider the GMM bootstrap. The test statistics include t-statistics based on k-class estimators and the robust subset quasi-LR (MQLR) statistic. In the simulations, the restricted fully efficient (RFE) bootstrap DGP based on Fuller estimates and the LIML t-statistic performs best of the Wald-type statistics. Unfortunately, the bootstrap only marginally reduces the conservativeness of the subset MQLR statistic. Finally, the GMM bootstrap does not seem to improve upon the asymptotic approximation. An empirical example illustrates the use of these procedures.

Keywords: Bootstrap, Subset Tests, Weak Instruments

*This paper was written while the author enjoyed the hospitality of the Department of Economics at McGill University, Montreal, Canada. I thank Russell Davidson and Frank Kleibergen for their very helpful and constructive comments on previous versions of this paper. The Swedish National Data Service (SND) is gratefully acknowledged for providing the data of the Swedish Level of Living Survey (LNU, 1981). In addition, I thank Sören Blomquist for providing the selection and transformation rules of the variables. All errors remain my responsibility.

1 INTRODUCTION

In the last decade, the econometric instrumental variables (IV) literature has focused on the so-called weak instruments problem, i.e. situations where instruments are poorly correlated with endogenous explanatory variables; see e.g. Dufour (2003) and Stock et al. (2002) for reviews. When instruments are weak, inference based on standard test statistics in structural models can be quite misleading; see e.g. Staiger and Stock (1997). To conduct accurate inference in the case of weak identification, several (asymptotic) pivotal test statistics have been proposed that are robust to weak identification including the AR statistic of Anderson and Rubin (1949), the KLM statistic of Kleibergen (2002) and the CLR statistic of Moreira (2003). When the complete parameter vector of the structural model is under test, inference based on these robust test statistics has (asymptotically) the correct size regardless of the strength of identification. In structural models that include more than one endogenous regressor, the parameter(s) of interest may only be a subset of the structural parameters. In these models, however, many of the results obtained for the single endogenous regressor case do not continue to hold for the individual structural coefficients. To solve this problem, two approaches for testing subsets of parameters have been developed. One approach is based on projection-type inference for subsets as initially proposed by Dufour and coauthors, see for instance Dufour (1997), Dufour and Jasiak (2001), and Dufour and Taamouti (2005). However, these projection-type tests can lead to quite conservative inference; see for instance Chaudhuri et al. (2010) and Chaudhuri and Zivot (2011) for two ways to reduce the conservativeness. The other approach, first suggested by Kleibergen (2004), is to conduct inference on a subset of parameters by substituting estimates for the unspecified coefficients of the structural equation. The analysis of Kleibergen and Mavroeidis (2011) shows that inference based on this approach will asymptotically have the correct size, although again inference might become conservative when instruments are weak.

The literature that has considered the use of the bootstrap in the weak instruments setting is rather limited. Flores-Lagunes (2007) investigates the usefulness of correcting the bias using the bootstrap, but he finds mixed results. Moreira et al. (2009) show that bootstrapping the LM test gives first-order correct inference in the weak instrument case, although the bootstrap does not deliver higher-order refinement. In a sequel of papers, Davidson and MacKinnon (2008, 2010 and 2011) propose a so-called restricted efficient (RE) residual bootstrap, which appears to work quite well in the single endogenous regressor case. In their 2010 paper (Section 5), a RE bootstrap procedure for IV t-statistics is suggested in case the structural equation includes multiple endogenous regressors. However, they do not investigate the performance of this bootstrap procedure. Recently, Kleibergen (2011) has suggested a GMM bootstrap for the GMM analogues of the AR, KLM and CLR statistic when the full parameter vector is under test. Using an Edgeworth approximation that exploits the independence between the score and the appropriate information matrix, he shows that the GMM bootstrap leads to higher-order refinement.

The main contribution of this paper is to investigate the finite-sample performance of several bootstrap procedures for conducting inference about a single coefficient in a structural linear regression model containing 2 endogenous regressors. In this model, various alternative estimators exist of the unspecified coefficient leading to more than one RE bootstrap procedure. Besides residual-based bootstrap procedures, the performance of the GMM bootstrap is investigated. We propose a modification to the GMM bootstrap of Kleibergen (2011) to ensure that it works well when testing only a subset of the parameters. In addition to several t-statistics, which are not robust to weak identification, the robust subset MQLR[1] statistic (Moreira's quasi-LR statistic; an extension of the LR statistic based on Kleibergen, 2007) is considered. This subset MQLR test statistic is as easy to use as the CLR test statistic, although it actually is an approximation to it in case of multiple endogenous regressors. As mentioned before, inference based on subset test statistics, which asymptotically have the correct size, can lead to low rejection probabilities in some regions of the parameter space. Hence, it is interesting to investigate if the bootstrap is able to reduce the conservativeness of the subset MQLR test. We focus on the subset MQLR test, because simulation results of Kleibergen and Mavroeidis (2011) indicate that inference based on this statistic leads

[1] Although QLR might be a better acronym for the quasi-LR test statistic, we follow the papers by Kleibergen (and Mavroeidis) in notation.

to almost the same conclusions as inference based on the LR statistic in the two endogenous regressor case. Although the MQLR statistic depends on the AR and KLM statistic, these statistics are not investigated individually because the analysis of Andrews et al. (2006) shows that the LR statistic is the most powerful among the robust test statistics, at least in the single endogenous regressor case.

The paper is organized as follows. In the next section, all the test statistics are introduced. Besides the test statistics based on k-class estimators like 2SLS, Fuller and LIML, the subset MQLR test is defined as well as its GMM counterpart. In Section 3, we discuss several bootstrap procedures. Most are based on bootstrapping residuals, but we also consider the GMM bootstrap. This latter bootstrap has to be adjusted for the fact that subsets are being tested. In Section 4, we investigate the finite-sample performance of all the bootstrap procedures by simulation. By first looking at the finite-sample properties of the original test statistics, the Monte Carlo design is chosen in such a way that all peculiarities of the test statistics are covered by it. Section 5 contains an empirical application that illustrates how to construct confidence intervals by inverting test statistics. Finally, Section 6 concludes.

2 The Model and Test statistics

Although the test statistics below are easily defined for any number of endogenous regressors, we consider the linear IV model with only two endogenous regressors denoted by

$$y = \beta x + \gamma w + \varepsilon \tag{1}$$

$$x = Z\pi_x + v \tag{2}$$

$$w = Z\pi_w + u, \tag{3}$$

where y, x and w are $n \times 1$ vectors and Z denotes a $n \times k$ dimensional matrix of instruments. The unknown coefficients are the scalar parameters β, γ, and the two $k \times 1$ vectors π_x and π_w. The i-th row of the $N \times 3$ matrix $[\varepsilon : v : u]$ is denoted by $(\varepsilon_i, v_i, u_i)$ and this zero-mean triplet is assumed to be serially uncorrelated. When they are homoskedastic, they have a contemporaneous 3×3 covariance matrix denoted by Σ. In the Monte Carlo experiments, the disturbances are assumed to be normally distributed with a constant variance. However, many of the test statistics in this section can be modified to deal with heteroskedasticity. Although the structural equation in (1) does not include additional exogenous variables, they can easily be included in the equation like

$$y = \beta x + \gamma w + X\psi + \varepsilon,$$

where X denotes a $n \times m$ matrix of exogenous variables such that the subspace spanned by the columns of X, denoted by $\mathcal{S}(X)$, is contained in $\mathcal{S}(Z)$. However, the system in equations (1)-(3) results if all the included exogenous regressors have been partialled out and modifying the degrees of freedom as needed. Therefore, we shall proceed assuming this is the case and no exogenous regressors appear in the structural equation.

Suppose we are interested in testing $H_0 : \beta = \beta_0$. A Wald-type test statistic is given by the t-statistic

$$t(\beta_0) = \frac{\hat{\beta}_\kappa - \beta_0}{SE(\hat{\beta}_\kappa)},$$

where $\hat{\beta}_\kappa$ is some k-class estimator of β and $SE(\hat{\beta}_\kappa)$ denotes its standard error. In this paper, we shall consider 2SLS (also known as GIVE or simply IV), LIML and Fuller (1977) estimators. If the test statistic is based on 2SLS, then a heteroskedasticity-consistent variance estimator could be used if heteroskedasticity is suspected. Although it is known that Wald-type test statistics are not robust against weak instruments, see inter alia Dufour (1997), the analysis in Davidson and MacKinnon (2008) shows that a bootstrapped version of the 2SLS t-statistic seems to work reasonable well even when instruments are quite weak.

A little algebra shows that the 2SLS estimators for β and γ are given by

$$\begin{aligned}\hat{\beta} &= \frac{P_{wy}P_{xw} - P_{ww}P_{xy}}{P_{wx}P_{xw} - P_{ww}P_{xx}}\\ \hat{\gamma} &= \frac{P_{xy}P_{xw} - P_{xx}P_{wy}}{P_{wx}P_{xw} - P_{ww}P_{xx}},\end{aligned}$$

where $P_{rq} = r'P_Z q$ for $r, q \in \{y, x, w\}$ and $P_Z = Z(Z'Z)^{-1}Z'$. It is easy to see that $\hat{\beta}$ is homogeneous of degree 1, -1 and 0 with respect to y, x and w, while $\hat{\gamma}$ is homogeneous of degree 1, 0, -1 with respect to y, x and w. The standard error of $\hat{\beta}$ is given by

$$SE(\hat{\beta}) = n^{-1/2}||y - \hat{\beta}x - \hat{\gamma}w|| \left(\frac{P_{ww}}{P_{ww}P_{xx} - P_{xw}^2}\right)^{1/2},$$

which is homogeneous of degree 1, -1 and 0 with respect to y, x, w. Therefore, the t-statistic is scale invariant with respect to y, x and w. Furthermore, the t-statistic is location invariant to γ. To see this, let $\gamma^+ = \gamma + \mu$, so that $y^+ = \beta x + \gamma^+ w + \varepsilon = y + \mu w$. Then we have $P_{wy^+} = P_{wy} + \mu P_{ww}$ and $P_{xy^+} = P_{xy} + \mu P_{xw}$. For $\hat{\gamma}$ based on y^+, we find

$$\begin{aligned}\hat{\gamma}^+ &= \frac{P_{xy^+}P_{xw} - P_{xx}P_{wy^+}}{P_{wx}P_{xw} - P_{ww}P_{xx}} \\ &= \frac{(P_{xy} + \mu P_{xw})P_{xw} - P_{xx}(P_{wy} + \mu P_{ww})}{P_{wx}P_{xw} - P_{ww}P_{xx}} = \hat{\gamma} + \mu.\end{aligned}$$

A similar calculation for $\hat{\beta}$ shows that $\hat{\beta}^+ = \hat{\beta}$, so the numerator of the t-statistic does not change. In addition, we have

$$||y^+ - \hat{\beta}^+ x - \hat{\gamma}^+ w|| = ||y - \hat{\beta}x - \hat{\gamma}w||,$$

so the standard error remains unchanged and therefore the t-statistic is invariant to γ.

Besides the t-statistic based on 2SLS, we also consider t-statistics based on LIML-type estimators because inference based on these estimators is often found to be more reliable than based on 2SLS estimators. A k-class estimator of $\theta = (\beta, \gamma)'$ depends on the random quantities

$$M_{rq}^Z = r'M_Z q \quad \text{and} \quad 1_{rq} = r'q \quad \text{for } r, q \in \{y, x, w\},$$

where $M_Z = I - P_Z$ and is of the form

$$\hat{\theta}_\kappa = (X'(I - \kappa M_Z)'X)^{-1}(X'(I - \kappa M_Z)y) \quad \text{with } X = [x : w].$$

Its variance can be estimated by

$$\widehat{Var}(\hat{\theta}_\kappa) = \hat{\sigma}_\kappa^2 (X'(I - \kappa M_Z)X)^{-1} \quad \text{with } \hat{\sigma}_\kappa^2 = n^{-1}||y - X\hat{\theta}_\kappa||^2.$$

When $\kappa = 1$, the 2SLS estimator results, while the LIML estimator is obtained when $\kappa = \hat{\kappa}$, where $\hat{\kappa}$ is the smallest eigenvalue of the matrix $(Y'M_ZY)^{-1}(Y'Y)$ with $Y = [y : x : w]$. It is well-known that the distribution of the LIML estimator has fat tails, although the median of LIML is typically much closer to the population values than IV. Furthermore, LIML is invariant to normalization and much less susceptible to weak instruments than 2SLS; see e.g. Stock and Yogo (2005a) that LIML and Fuller estimators are even consistent under many weak instrument asymptotics. The Fuller estimator results when $\kappa = \hat{k} - c/(n-k)$ for some constant c. Under standard asymptotics and $c = 1$, the Fuller estimator has moments to all orders (provided the sample size is large enough) and is approximately mean unbiased. In a simulation study, Hahn et al. (2004) find that inference based on Fuller's estimator outperforms inference based on LIML when the instruments are weak. Although Fuller's modification depends on the tuning parameter c, we shall only consider the case $c = 1$ in this paper.

Since the residual-based bootstrap procedures rely on the parameter estimates under the null hypothesis, let $\tilde{\gamma}(\beta_0)$ denote the LIML estimator for γ under the restriction $\beta = \beta_0$. It depends upon the smallest eigenvalue of the matrix $(Y_0' M_Z Y_0)^{-1}(Y_0' Y_0)$ for $Y_0 = [y - \beta_0 x : w]$. Since this is a 2 × 2 matrix, the smallest eigenvalue can be calculated explicitly and equals

$$\tilde{\kappa}(\beta_0) = \frac{c_1 - \sqrt{c_1^2 - 4c_2 c_3}}{2c_3}$$

with

$$\begin{aligned} c_1 &= 1_{ww} M^Z_{y_0 y_0} - 2M^Z_{y_0 w} 1_{y_0 w} + M^Z_{ww} 1_{y_0 y_0} \\ c_2 &= 1_{ww} 1_{y_0 y_0} - (1_{y_0 w})^2 \\ c_3 &= M^Z_{ww} M^Z_{y_0 y_0} - (M^Z_{y_0 w})^2, \end{aligned}$$

where $y_0 = y - \beta_0 x$. It is easy to see that the eigenvalue $\tilde{\kappa}(\beta_0)$ is scale invariant with respect to y_0 and w.

In the weak instrument literature, several (asymptotic) pivotal test statistics are proposed that are robust to weak identification including the AR statistic, the KLM statistic and the CLR statistic. These test statistics can be used for testing joint hypotheses on β and γ, e.g. $H_0^{\beta,\gamma} : \beta = \beta_0$ and $\gamma = \gamma_0$. If we only want to test a hypothesis about one coefficient, these testing procedures can be modified resulting in so-called subset tests; see the introduction for another approach based on projection methods. The main idea behind these subset tests is to replace the nuisance parameter γ with the LIML estimator $\tilde{\gamma}(\beta_0)$. Kleibergen and

Mavroeidis (2011) investigate the asymptotic properties of the subset AR, subset KLM and approximate version of the subset LR test in linear IV models. In the single endogenous regressor case, the analysis of Mikusheva (2010) shows that confidence intervals based on the CLR test statistic seem to perform best. So, although this paper is mainly concerned with testing, it seems reasonable to concentrate on LR-like test statistics. However, the asymptotic distribution of the subset LR test depends on the smallest characteristic root of an $(m+1)$-square matrix, where m denotes the number of endogenous regressors. Hence, Kleibergen and Mavroeidis (2011) propose to use the subset MQLR test statistic, which equals the LR test when all the m singular values are restricted to the smallest one. In their simulations, they find that inference based on the subset MQLR statistic leads to almost the same conclusions as inference based on the subset LR statistic. Hence, we focus on the performance of the subset MQLR statistic applied in the linear IV setting. This test statistic can be written as a function of the subset AR test, the subset KLM test and a root of a characteristic polynomial. From Definition 1 from Kleibergen and Mavroeidis (2011), the subset AR statistic (times k) for testing $H_0 : \beta = \beta_0$ reads

$$\mathrm{AR}(\beta_0) = \frac{1}{\hat{\sigma}_{\varepsilon\varepsilon}(\beta_0)}(y - x\beta_0 - w\tilde{\gamma}(\beta_0))' P_Z (y - x\beta_0 - w\tilde{\gamma}(\beta_0)),$$

where

$$\hat{\sigma}_{\varepsilon\varepsilon}(\beta_0) = \begin{pmatrix} 1 \\ -\beta_0 \\ -\tilde{\gamma}(\beta_0) \end{pmatrix}' \hat{\Omega} \begin{pmatrix} 1 \\ -\beta_0 \\ -\tilde{\gamma}(\beta_0) \end{pmatrix} \quad \text{with } \hat{\Omega} = \frac{1}{n-k}(y : X : W)' M_Z (y : X : W).$$

The subset Kleibergen Lagrange Multiplier (KLM) statistic is given by

$$\mathrm{KLM}(\beta_0) = \frac{1}{\hat{\sigma}_{\varepsilon\varepsilon}(\beta_0)}(y - x\beta_0 - w\tilde{\gamma}(\beta_0))' P_{Z(\tilde{\pi}_x(\beta_0):\tilde{\pi}_w(\beta_0))}(y - x\beta_0 - w\tilde{\gamma}(\beta_0)), \quad (4)$$

where

$$\begin{aligned} \tilde{\pi}_x(\beta_0) &= (Z'Z)^{-1}Z'\left[x - (y - x\beta_0 - w\tilde{\gamma}(\beta_0))\frac{\hat{\sigma}_{\varepsilon v}(\beta_0)}{\hat{\sigma}_{\varepsilon\varepsilon}(\beta_0)}\right] \\ \tilde{\pi}_w(\beta_0) &= (Z'Z)^{-1}Z'\left[w - (y - x\beta_0 - w\tilde{\gamma}(\beta_0))\frac{\hat{\sigma}_{\varepsilon u}(\beta_0)}{\hat{\sigma}_{\varepsilon\varepsilon}(\beta_0)}\right] \end{aligned}$$

and

$$\hat{\sigma}_{\varepsilon v}(\beta_0) = \begin{pmatrix} 1 \\ -\beta_0 \\ -\tilde{\gamma}(\beta_0) \end{pmatrix}' \hat{\Omega} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \text{ and } \hat{\sigma}_{\varepsilon u}(\beta_0) = \begin{pmatrix} 1 \\ -\beta_0 \\ -\tilde{\gamma}(\beta_0) \end{pmatrix}' \hat{\Omega} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

A subset extension of the CLR statistic reads

$$\text{MQLR}(\beta_0) = \tfrac{1}{2}\left[\text{AR}(\beta_0) - \text{rk}(\beta_0) + \sqrt{(\text{AR}(\beta_0) + \text{rk}(\beta_0))^2 - 4(\text{AR}(\beta_0) - \text{KLM}(\beta_0))\text{rk}(\beta_0)}\right],$$

where $\text{rk}(\beta_0)$ is the smallest characteristic root of $\hat{\Sigma}_{MQLR}(\beta_0) = T(\beta_0)'T(\beta_0)$ with

$$T(\beta_0) = (Z'Z)^{1/2}\left[\tilde{\pi}_x(\beta_0) : \tilde{\pi}_w(\beta_0)\right]\hat{\Sigma}^{-1/2}_{(x:w)(x:w).\varepsilon}$$

and

$$\hat{\Sigma}^{-1/2}_{(x:w)(x:w).\varepsilon} = \begin{pmatrix} \hat{\sigma}^{-1/2}_{xx.(\varepsilon:w)} & 0 \\ -\hat{\sigma}^{-1}_{ww.\varepsilon}\hat{\sigma}_{wx.\varepsilon}\hat{\sigma}^{-1/2}_{xx.(\varepsilon:w)} & \hat{\sigma}^{-1/2}_{ww.\varepsilon} \end{pmatrix}$$

in which

$$\begin{aligned} \hat{\sigma}_{xx.(\varepsilon:w)} &= (n-k)^{-1}x'M_{(Z:w:\tilde{\varepsilon})}x \\ \hat{\sigma}_{wx.\varepsilon} &= (n-k)^{-1}w'M_{(Z:\tilde{\varepsilon})}x \\ \hat{\sigma}_{ww.\varepsilon} &= (n-k)^{-1}w'M_{(Z:\tilde{\varepsilon})}w \end{aligned}$$

and $\tilde{\varepsilon} = y - x\beta_0 - w\tilde{\gamma}(\beta_0)$. As noted before, this subset extension is an approximation to the subset LR statistic that results under i.i.d. normal disturbances. However, this subset MQLR is as easy to use as the LR statistic in the single endogenous regressor case.

Kleibergen and Mavroeidis (2011) show that the limiting distributions of all these subset tests are bounded from above (for all values of π_w) by the limiting distributions that result under the strong instrument assumption (with respect to π_w). However, when instruments are weak, these subset tests can be quite conservative. Therefore, it is interesting to see if the bootstrap improves upon the approximation based on the bounding limiting distributions. Kleibergen and Mavroeidis (2011) note that the bounding results do not apply if the 2SLS estimator of γ is used instead of the LIML estimator. Since the bounding distribution of the subset MQLR statistic is the same as the CLR statistic in the single endogenous regressor case, we can determine its p-value very fast by numerically integrating a one dimensional function as shown in the Appendix.

To apply the GMM-based bootstrap, we shall also consider the GMM version of the subset MQLR statistic. The following paragraph draws heavily from Kleibergen and Mavroeidis (2009). Inference about the 2-dimensional parameter vector $\theta = (\beta, \gamma)'$ in the GMM framework is based on the k-dimensional moment equation

$$\mathbb{E}[f_i(\theta)] = 0, \qquad i = 1, ..., n.$$

For our model, the moment vector is given by

$$f_i(\theta) = Z_i(y_i - \beta x_i - \gamma w_i),$$

where Z_i denotes the transpose of the i^{th} row of the $n \times k$ matrix Z. Let $\bar{f}_i(\theta) = f_i(\theta) - f_n(\theta)$ with $f_n(\theta) = n^{-1}\sum_{i=1}^{n} f_i(\theta)$ denote the moment vector in deviation from its mean. Besides the moment vector $f_i(\theta)$, its derivative with respect to θ plays an important role in GMM, which for our model reads

$$q_i(\theta) = vec\left(\frac{\partial f_i(\theta)}{\partial \theta'}\right) = -\left(\begin{pmatrix} x_i \\ w_i \end{pmatrix} \otimes Z_i\right) = \begin{pmatrix} q_i^x(\theta) \\ q_i^w(\theta) \end{pmatrix}.$$

Similarly, let $\bar{q}_i(\theta) = q_i(\theta) - q_n(\theta)$ with $q_n(\theta) = n^{-1}\sum_{i=1}^{n} q_i(\theta)$. Suppose the Eicker-White covariance matrix is used in the GMM approach, so that

$$\begin{aligned} \hat{V}_{ff}(\theta) &= n^{-1}\sum\nolimits_{i=1}^{n} f_i(\theta) f_i(\theta)' - f_n(\theta) f_n(\theta)', \\ \hat{V}_{qf}(\theta) &= n^{-1}\sum\nolimits_{i=1}^{n} q_i(\theta) f_i(\theta)' - q_n(\theta) f_n(\theta)', \\ \hat{V}_{qq}(\theta) &= n^{-1}\sum\nolimits_{i=1}^{n} q_i(\theta) q_i(\theta)' - q_n(\theta) q_n(\theta)'. \end{aligned}$$

The use of the Eicker-White covariance matrix makes the GMM statistics defined below asymptotically valid in the presence of heteroskedasticity of unknown form. In a time series setting, one could also use a HAC estimator which would make inference valid even in the presence of autocorrelated disturbances. The weak instrument robust test statistics use the following estimator for the derivative of the unconditional expectation of the Jacobian:

$$\hat{D}_n(\theta) = [q_n^x(\theta) - \hat{V}_{qf}^x(\theta)\hat{V}_{ff}(\theta)^{-1} f_n(\theta) : q_n^w(\theta) - \hat{V}_{qf}^w(\theta)\hat{V}_{ff}(\theta)^{-1} f_n(\theta)].$$

The projection of $q_n^r(\theta)$ for $r = \{x, w\}$ onto $f_n(\theta)$ ensures that $\hat{D}_n(\theta)$ is asymptotically uncorrelated with $f_n(\theta)$. Note that θ denotes a 2-dimensional parameter vector, so for testing the hypothesis $H_0 : \beta = \beta_0$, the subset testing approach uses $\ddot{\theta}_0 = (\beta_0, \ddot{\gamma}(\beta_0))$ where $\ddot{\gamma}(\beta_0)$ is the continuous updating estimator (CUE; see Hansen et al. (1996)) of γ given $\beta = \beta_0$, i.e.

$$\arg\min_{\gamma} \quad n f_n(\beta_0, \gamma)' \hat{V}_{ff}(\beta_0, \gamma)^{-1} f_n(\beta_0, \gamma). \tag{5}$$

The minimum value of the objective function shown in (5) is actually the subset S-statistic of Stock and Wright (2000), which can be seen as the GMM analogue of the AR statistic. To connect the AR, KLM and MQLR statistics to their GMM counterparts, we note that $\hat{D}_n(\ddot{\theta}_0)$ is related to $\tilde{\Pi}(\beta_0) = (\tilde{\pi}_x(\beta_0) : \tilde{\pi}_w(\beta_0))$ in the following way

$$\hat{D}_n(\ddot{\theta}_0) \approx -(n^{-1}Z'Z)\tilde{\Pi}(\beta_0).$$

The relationship is only approximate because GMM is based on the CUE estimator and the Eicker-White covariance matrix, while $\tilde{\Pi}(\beta_0)$ is based on the LIML estimator and the covariance matrix that exploits the homoskedasticity assumption. In addition, we have

$$f_n(\ddot{\theta}_0) \approx n^{-1}Z'\tilde{\varepsilon}(\beta_0) \quad \text{and} \quad \hat{V}_{ff}(\ddot{\theta}_0) \approx \tilde{\sigma}_{\varepsilon\varepsilon} n^{-1} Z'Z,$$

where

$$\begin{aligned} f_n(\ddot{\theta}_0) &= n^{-1} \textstyle\sum_{i=1}^n f_i(\ddot{\theta}_0) = n^{-1} Z'\ddot{\varepsilon}(\beta_0) \\ \hat{V}_{ff}(\ddot{\theta}_0) &= n^{-1} \textstyle\sum_{i=1}^n \ddot{\varepsilon}(\beta_0)_i^2 Z_i Z_i' - n^{-2} Z'\ddot{\varepsilon}(\beta_0)\ddot{\varepsilon}(\beta_0)'Z \end{aligned}$$

with $\ddot{\varepsilon}(\beta_0) = (y_i - \beta_0 x_i - \ddot{\gamma}(\beta_0) w_i)$. The GMM analogues of the AR and KLM statistic for testing $H_0 : \beta = \beta_0$ are given by

$$\begin{aligned} S_{\perp}(\beta_0) &= n f_n(\ddot{\theta}_0)' \hat{V}_{ff}(\ddot{\theta}_0)^{-1} f_n(\ddot{\theta}_0) \\ &\approx n(n^{-1}\tilde{\varepsilon}(\beta_0)'Z)(\tilde{\sigma}_{\varepsilon\varepsilon} n^{-1} Z'Z)^{-1}(n^{-1}Z'\tilde{\varepsilon}(\beta_0)) \\ &= \frac{1}{\tilde{\sigma}_{\varepsilon\varepsilon}}(y - x\beta_0 - w\tilde{\gamma}(\beta_0))' P_Z (y - x\beta_0 - w\tilde{\gamma}(\beta_0)) = \text{AR}(\beta_0) \end{aligned}$$

and

$$\begin{aligned}\mathrm{KLM}_{\perp}(\beta_0) &= n f_n(\ddot{\theta}_0)'\hat{V}_{ff}(\ddot{\theta}_0)^{-1}\hat{D}_n(\ddot{\theta}_0)\left[\hat{D}_n(\ddot{\theta}_0)'\hat{V}_{ff}(\ddot{\theta}_0)^{-1}\hat{D}_n(\ddot{\theta}_0)\right]^{-1} \\ &\quad \hat{D}_n(\ddot{\theta}_0)'\hat{V}_{ff}(\ddot{\theta}_0)^{-1} f_n(\ddot{\theta}_0) \\ &\approx \frac{n}{\tilde{\sigma}_{\varepsilon\varepsilon}}(n^{-1}\tilde{\varepsilon}(\beta_0)'Z)\tilde{\Pi}(\beta_0)\left[\tilde{\Pi}(\beta_0)'(Z'Z/n)\tilde{\Pi}(\beta_0)\right]^{-1}\tilde{\Pi}(\beta_0)'(n^{-1}Z'\tilde{\varepsilon}(\beta_0)) \\ &= \frac{1}{\tilde{\sigma}_{\varepsilon\varepsilon}}(y - x\beta_0 - w\tilde{\gamma}(\beta_0))'P_{Z\tilde{\Pi}(\beta_0)}(y - x\beta_0 - w\tilde{\gamma}(\beta_0)) = \mathrm{KLM}(\beta_0).\end{aligned}$$

Finally, the subset extension of the LR statistic in a GMM setting is given by

$$\begin{aligned}\mathrm{MQLR}_{\perp}(\beta_0) &= \tfrac{1}{2}[\mathrm{AR}_{\perp}(\beta_0) - \mathrm{rk}_{\perp}(\beta_0) + \\ &\quad \sqrt{(\mathrm{AR}_{\perp}(\beta_0) + \mathrm{rk}_{\perp}(\beta_0))^2 - 4(\mathrm{AR}_{\perp}(\beta_0) - \mathrm{KLM}_{\perp}(\beta_0))\mathrm{rk}_{\perp}(\beta_0)}],\end{aligned}$$

where $\mathrm{rk}_{\perp}(\beta_0)$ is a statistic that tests for a lower rank value of the expected value of the Jacobian. Following Kleibergen and Mavroeidis (2009), we use the Cragg and Donald rank statistic; see equation (22) and the Appendix of their paper.

3 Bootstrap Methods for IV Regression

Monte Carlo simulations reported by Davidson and MacKinnon (2008) show that the so-called residual bootstrap outperforms the pairs bootstrap in the single endogenous regressor case. The pairs bootstrap was suggested by Freedman (1984), who analyzed its properties under strong instrument asymptotics. One of the reasons why the residual bootstrap works so well is because its DGP is based on estimates under the null hypothesis, which among other things eliminates the estimation error about the coefficient under test. Davidson and MacKinnon (2008) consider two residual bootstrap methods and they find that the so-called restricted efficient (RE) bootstrap delivers the most accurate inference in the single endogenous regressor case. In Davidson and MacKinnon (2010), the following RE bootstrap is suggested for more than one endogenous regressors. For testing $H_0 : \beta = \beta_0$, the parameters in the structural equation (1) are estimated by 2SLS under the null hypothesis, e.g.

$$y - \beta_0 x = \gamma w + \varepsilon.$$

Let $\hat{\gamma}(\beta_0)$ denote this estimate, which lead to the 2SLS residuals $\hat{\varepsilon}(\beta_0)$. Then the reduced form equations (2) and (3) are estimated to obtain efficient estimates, i.e. estimates that use all available information which are asymptotically equivalent to 3SLS. This is done by adding the restricted 2SLS residuals $\hat{\varepsilon}(\beta_0)$ to the reduced form equations, i.e.

$$\begin{aligned} x &= Z\pi_x + \phi_x\hat{\varepsilon}(\beta_0) + \text{errors} \\ w &= Z\pi_w + \phi_w\hat{\varepsilon}(\beta_0) + \text{errors.} \end{aligned}$$

The OLS estimators are denoted by $\hat{\pi}_x(\beta_0)$ and $\hat{\pi}_w(\beta_0)$. If $\hat{v}(\beta_0) = x - Z\hat{\pi}_x(\beta_0)$ and $\hat{u}(\beta_0) = w - Z\hat{\pi}_w(\beta_0)$ denote the residuals (disregarding $\hat{\phi}_x\hat{\varepsilon}(\beta_0)$), then the RE bootstrap DGP is given by

$$\begin{aligned} y_i^* - \beta_0 x_i^* &= \hat{\gamma}(\beta_0)w_i^* + \varepsilon_i^* \qquad (6) \\ x_i^* &= Z_i\hat{\pi}_x(\beta_0) + v_i^* \\ w_i^* &= Z_i\hat{\pi}_w(\beta_0) + u_i^*, \end{aligned}$$

with

$$\begin{bmatrix} \varepsilon_i^* \\ v_i^* \\ u_i^* \end{bmatrix} \sim \hat{F}\begin{pmatrix} \hat{\varepsilon}(\beta_0)_i \\ (n/(n-k))^{1/2}\hat{v}(\beta_0)_i \\ (n/(n-k))^{1/2}\hat{u}(\beta_0)_i \end{pmatrix}.$$

Here, $\hat{F}$ denotes an estimator of the distribution of the triplet $(\varepsilon_i, v_i, u_i)$. This can either be a parametric estimator, e.g. a multivariate normal distribution, or a non-parameter estimator, e.g. the empirical distribution function (EDF). Of course, the wild bootstrap can also be used if heteroskedasticity is suspected; see Davidson and MacKinnon (2010). In this paper, only the i.i.d. bootstrap with respect to the residuals is considered, i.e. $\hat{F}$ is the EDF. Bootstrapped test statistics are obtained by evaluating the test statistics in the bootstrap sample $[y^* : x^* : w^*]$. In Section 2, we have seen that the 2SLS t-statistic for testing the restriction $\beta = \beta_0$ is invariant to γ. This implies that the bootstrap 2SLS t-statistic is invariant to $\hat{\gamma}(\beta_0)$, so for this test statistic w_i^* could be left out of the structural equation given in (6).

Note that the RE bootstrap is especially suited for the t-statistic based on 2SLS estimation. However, as noticed earlier, LIML and Fuller estimators are partially robust to weak instruments, so exploiting these estimators might lead to better bootstrap DGPs. The next bootstrap procedure is called the restricted fully efficient (RFE) bootstrap, since it not only incorporates efficient estimators of the parameters in the reduced form equations, but also uses an efficient estimator in the structural equation. In effect, the RFE bootstrap DGP is the same as the RE bootstrap DGP, but uses the LIML estimator $\tilde{\gamma}(\beta_0)$ instead of the 2SLS estimator $\hat{\gamma}(\beta_0)$. Since all three equations in the bootstrap DGP are influenced by this, we shall describe the bootstrap DGP explicitly. After the efficient estimate $\tilde{\gamma}(\beta_0)$ is obtained by LIML, calculate the efficient restricted residuals $\tilde{\varepsilon}(\beta_0) = (y - x\beta_0 - w\tilde{\gamma}(\beta_0))$. Then estimate the reduced form equations, although now $\tilde{\varepsilon}(\beta_0)$ is added to the equations, i.e.

$$x = Z\pi_x + \phi_x\tilde{\varepsilon}(\beta_0) + \text{errors} \qquad (7)$$

$$w = Z\pi_w + \phi_w\tilde{\varepsilon}(\beta_0) + \text{errors}, \qquad (8)$$

leading to the OLS estimators $\tilde{\pi}_x(\beta_0)$ and $\tilde{\pi}_w(\beta_0)$. If $\tilde{v}(\beta_0) = x - Z\tilde{\pi}_x(\beta_0)$ and $\tilde{u}(\beta_0) = w - Z\tilde{\pi}_w(\beta_0)$ denote the residuals of the reduced equations, then the restricted fully efficient (RFE) bootstrap DGP is given by

$$\begin{aligned} y_i^* - \beta_0 x_i^* &= \tilde{\gamma}(\beta_0)w_i^* + \tilde{\varepsilon}_i^* \\ x_i^* &= Z_i\tilde{\pi}_x(\beta_0) + \tilde{v}_i^* \\ w_i^* &= Z_i\tilde{\pi}_w(\beta_0) + \tilde{u}_i^*, \end{aligned} \qquad (9)$$

with

$$\begin{bmatrix} \tilde{\varepsilon}_i^* \\ \tilde{v}_i^* \\ \tilde{u}_i^* \end{bmatrix} \sim \hat{F}\begin{pmatrix} \tilde{\varepsilon}(\beta_0)_i \\ (n/(n-k))^{1/2}\tilde{v}(\beta_0)_i \\ (n/(n-k))^{1/2}\tilde{u}(\beta_0)_i \end{pmatrix}.$$

Note that the RFE bootstrap reduces to the RE bootstrap in case there is only one endogenous regressor.

Due to the non-existing moments of the LIML estimator in finite samples, it seems logical to also consider the Fuller estimator in the structural equation. In the Monte Carlo experiments, the 2SLS based t-statistic is considered under both bootstrap DGPs, while the Fuller and LIML based t-statistics are investigated under only the RFE bootstrap. Moreover, the KLM statistic is also based on efficient estimators of the reduced form parameters, viz. $\tilde{\Pi}(\beta_0) = (\tilde{\pi}_x(\beta_0) : \tilde{\pi}_w(\beta_0))$. In fact, $\tilde{\Pi}(\beta_0)$ is numerically identical to the OLS estimators of π_x and π_w in equations (7)-(8) when $\tilde{\varepsilon}(\beta_0)$ is based on LIML. Therefore, we also consider bootstrap DGPs that exploit $\tilde{\Pi}(\beta_0)$ in the reduced equations. The RE and RFE bootstrap DGPs are different from Moreira et al. (2009) in two ways: (i) they use restricted estimators of the structural parameters in their bootstrap DGP and (ii) they incorporate efficient estimators in the reduced equations.

Next, we turn to the GMM bootstrap as suggested by Kleibergen (2011). The advantage of the GMM bootstrap is that only the moments under the null hypothesis are resampled, so the reduced form equations are not involved in this bootstrap. The GMM bootstrap is just the classical i.i.d. bootstrap applied to the centered moment vectors $\bar{f}_i(\ddot{\theta}_0)$. This centering ensures that the expectation under the bootstrap DGP of $f_n^*(\cdot)$ equals 0; see e.g. Hall and Horowitz (1996). So, the bootstrap sample $\{f_i^*(\theta_0^*), i = 1, ..., n\}$ is defined as

$$P[f_i^*(\theta_0^*) = \bar{f}_j(\ddot{\theta}_0)] = \frac{1}{n}, \qquad 1 \leq i, j \leq n.$$

Using $\{f_i^*(\theta_0^*), i = 1, ..., n\}$, we define the bootstrap quantities

$$\begin{aligned} f_n^*(\theta_0^*) &= n^{-1} \sum\nolimits_{i=1}^n f_i^*(\theta_0^*) \\ \hat{V}_{ff}^*(\theta_0^*) &= n^{-1} \sum\nolimits_{i=1}^n f_i^*(\theta_0^*) f_i^*(\theta_0^*)' - f_n^*(\theta_0^*) f_n^*(\theta_0^*)'. \end{aligned}$$

Note, however, that in the original sample, the f_i's result from a FOC for a minimum of the subset S-statistic as shown in (5). To incorporate this FOC in the bootstrap, let $\{I_i, i = 1, ..., B\}$ denote the indices used to define the bootstrap sample, so that I_i is randomly distributed over $\{1, ..., n\}$. Recall that $\bar{f}_i(\theta) = Z_i(y_i - \beta x_i - \gamma w_i) - f_n(\beta_0)$, so we can write

$$f_i^*(\beta_0, \gamma^*) = Z_{I_i}(y_{I_i} - \beta_0 x_{I_i} - \gamma^* w_{I_i}) - f_n(\beta_0).$$

The average of the moment vector equals $f_n^*(\beta_0, \gamma^*) = n^{-1} \sum f_i^*(\beta_0, \gamma^*) - f_n(\beta_0)$ and can be viewed as a function of β_0 and γ^*. If $\ddot{\gamma}^*(\beta_0)$ is defined as the value of γ^* that solves

$$\arg\min_{\gamma^*} \quad n f_n^*(\beta_0, \gamma^*) \hat{V}_{ff}^*(\beta_0, \gamma^*)^{-1} f_n^*(\beta_0, \gamma^*), \tag{10}$$

then we obtain $\ddot{\theta}_0^* = (\beta_0, \ddot{\gamma}^*(\beta_0))$, which can be considered the bootstrap analogue of $\ddot{\theta}_0 = (\beta_0, \ddot{\gamma}(\beta_0))$. The GMM bootstrap than proceeds along the normal way using

$$f_i^*(\ddot{\theta}_0^*) = Z_{I_i}(y_{I_i} - \beta_0 x_{I_i} - \ddot{\gamma}^*(\beta_0) w_{I_i}) - f_n(\beta_0),$$

leading to the following bootstrapped GMM subset statistics

$$\begin{aligned} \mathrm{S}_{\perp}^*(\beta_0) &= n f_n^*(\ddot{\theta}_0^*)' \hat{V}_{ff}^*(\ddot{\theta}_0^*)^{-1} f_n(\ddot{\theta}_0^*) \\ \mathrm{KLM}_{\perp}^*(\beta_0) &= n f_n^*(\ddot{\theta}_0^*)' \hat{V}_{ff}(\ddot{\theta}_0^*)^{-1} \hat{D}_n(\ddot{\theta}_0) \left[\hat{D}_n(\ddot{\theta}_0)' \hat{V}_{ff}(\ddot{\theta}_0^*)^{-1} \hat{D}_n(\ddot{\theta}_0) \right]^{-1} \\ & \quad \hat{D}_n(\ddot{\theta}_0)' \hat{V}_{ff}(\ddot{\theta}_0^*)^{-1} f_n(\ddot{\theta}_0^*). \end{aligned}$$

Note that $\hat{D}_n(\ddot{\theta}_0)$ is kept fixed in $\mathrm{KLM}_{\perp}^*(\beta_0)$. The subset $\mathrm{MQLR}_{\perp}^*(\beta_0)$ follows from the values of $\mathrm{S}_{\perp}^*(\beta_0)$, $\mathrm{KLM}_{\perp}^*(\beta_0)$ and the value of $\mathrm{rk}_{\perp}(\beta_0)$ that was obtained from the original sample. Kleibergen (2011) also suggests a second GMM bootstrap, which resamples both the moment vectors and their derivatives, i.e. $(\bar{f}_j(\theta_0)', \bar{q}_i(\theta_0)')'$. In simulations, however, Kleibergen (2011) found that resampling the $q_i(\theta)$'s is of lesser importance and does not lead to any further size improvements. So, we shall not consider this kind of GMM bootstrap in the Monte Carlo simulations.

For all the test statistics, say $\hat{\tau}$, B bootstrap replications, say $\{\hat{\tau}_j^*, j = 1, ...B\}$, are obtained under the various bootstrap DGPs. For a t-statistic, the following estimator of the equal-tail bootstrap p-value is used:

$$\hat{p}_{et}^*(\hat{\tau}) = 2 \min \left(\frac{1}{B} \sum\nolimits_{j=1}^{B} I(\hat{\tau}_j^* \leq \hat{\tau}), \frac{1}{B} \sum\nolimits_{j=1}^{B} I(\hat{\tau}_j^* > \hat{\tau}) \right).$$

For the MQLR statistic, the right-tail bootstrap p-value is used, i.e.

$$\hat{p}^*(\hat{\tau}) = \frac{1}{B} \sum\nolimits_{j=1}^{B} I(\hat{\tau}_j^* \geq \hat{\tau}).$$

The null hypothesis is rejected if the appropriate bootstrap p-value is at most the nominal significance level α.

4 Monte Carlo Simulations

In Section 2, it was shown that the 2SLS t-statistic is scale and location invariant. Under the Gaussian assumption, the joint distribution for $Y = [y : x : w]$ is of the form

$$Y \sim N(Z\Pi[\theta : I_2], I_n \otimes \Omega),$$

with the $k \times 2$ matrix $\Pi = (\pi_x : \pi_w)$ and 2×1 vector $\theta = (\beta, \gamma)'$. Hence, the Fuller and LIML estimators are also scale invariant and it is sufficient to consider the covariance (=correlation) matrix

$$\Sigma = \begin{pmatrix} 1 & \rho_{\varepsilon v} & \rho_{\varepsilon u} \\ \rho_{\varepsilon v} & 1 & \rho \\ \rho_{\varepsilon u} & \rho & 1 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{\varepsilon v} : \rho_{\varepsilon u} \\ (\rho_{\varepsilon v} : \rho_{\varepsilon u})' & \Sigma_{vu} \end{pmatrix},$$

where the correlations $\rho_{\varepsilon v}$ and $\rho_{\varepsilon u}$ parameterize the degree of endogeneity of x and w. Since there are two endogenous regressors, the concentration parameter is a 2×2 matrix, see e.g. Stock and Yogo (2005b), and is given by

$$\Sigma_{vu}^{-1/2}(\pi_x : \pi_w)'Z'Z(\pi_x : \pi_w)\Sigma_{vu}^{-1/2}.$$

Note that the matrix concentration parameter depends on Z and Π only through $(Z\pi_x : Z\pi_w)$, so we only have to consider two linear combinations of Z. By reparameterization, this leads to the reduced form parameter vectors $\pi_x = (\pi_{x1}, \pi_{x2}, 0, ..., 0)'$ and $\pi_w = (\pi_{w1}, \pi_{w2}, 0, ..., 0)'$.

In the weak-instrument asymptotics of Staiger and Stock (1997), the concentration parameter is kept fixed when the sample size goes to infinity. This can be achieved in two ways. Either the parameters $(\pi_x : \pi_w)$ are in a $1/\sqrt{n}$ neighborhood of 0 or the matrix $Z'Z$ is somehow kept constant; see Davidson and MacKinnon (2008) for the latter approach in the single endogenous regressors case. Suppose Z consist of k orthogonal vectors $z_1, ..., z_k$ with a length of one, so that $||z_i||^2 = 1$ and $z_i'z_j = 0$ for $i \neq j$. Then, we have

$$(\pi_x : \pi_w)'Z'Z(\pi_x : \pi_w) = \begin{pmatrix} \pi_x'\pi_x & \pi_x'\pi_w \\ \pi_w'\pi_x & \pi_w'\pi_w \end{pmatrix}. \tag{11}$$

If the matrix in (11) is positive definite, we can find a Cholesky decomposition which suggests that we can parameterize $Z(\pi_x : \pi_w)$ as an upper triangular matrix. Taking $\pi_x = (a, 0, ..., 0)'$ and $\pi_w = (b, c, 0, ..., 0)'$, we can keep the matrix in (11) constant while varying the number of instruments k. Using a singular value decomposition, we find

$$\Sigma_{vu}^{-1/2} = \frac{1}{2} \begin{pmatrix} \frac{1}{\sqrt{1-\rho}} + \frac{1}{\sqrt{1+\rho}} & -\frac{1}{\sqrt{1-\rho}} + \frac{1}{\sqrt{1+\rho}} \\ -\frac{1}{\sqrt{1-\rho}} + \frac{1}{\sqrt{1+\rho}} & \frac{1}{\sqrt{1-\rho}} + \frac{1}{\sqrt{1+\rho}} \end{pmatrix}.$$

Some algebra shows that the eigenvalues of the concentration matrix are equal to

$$\lambda_{\pm} = \frac{a^2 + b^2 + c^2 - 2\rho ab \pm \sqrt{(a^2 + b^2 + c^2 - 2\rho ab)^2 - 4a^2c^2(1-\rho^2)}}{2(1-\rho^2)}.$$

To keep the number of parameters manageable and spill-over effects from one endogenous regressor to the other as small as possible, we set $b = 0$. When ρ is positive (and $b = 0$), the smallest eigenvalue is an increasing function of ρ. Since both parameters β and γ can be either weakly, moderately or strongly identified, we can distinguish a number of cases; see also Stock and Wright (2000) and Chaudhuri and Zivot (2011). Here we focus on the following four cases, Case I: β and γ weakly identified, Case II: β weakly and γ moderately identified, Case III: β moderately and γ weak and Case IV: β and γ moderately identified. To keep the number of tables limited, we do not consider situations in which the parameters are strongly identified since they are of lesser importance. For the weakly identified case, we set a^2 and/or c^2 to 2 and for the moderately identified case we set a^2 and/or c^2 equal to 25. In the simulations, both size and power properties are investigated. In choosing the structural parameters, we follow Dufour and Taamouti (2007) and take $\beta = 0.5$ and $\gamma = 1$.

First, we investigate how the rejection frequencies vary over the degree of simultaneity $(\rho_{\varepsilon v}, \rho_{\varepsilon u})$ and the correlation between v and u given by ρ. We take $\rho_{\varepsilon v}$, $\rho_{\varepsilon v}$, $\rho \in \{0.3, 0.6, 0.9\}$, where the values 0.3/0.6/0.9 are referred to as low, medium and high. However, not all combinations lead to a valid covariance matrix Σ. For instance, if the endogeneity increases, then ρ has to increase as well. For the $27 (= 3^3)$ combinations, only 21 are permissible. Figures 1 and 2 show the rejection frequencies of $H_0 : \beta = \beta_0$ based on 20,000 replications, $n = 200$ and $k = 10$ for the 2SLS t-statistic (Figure 1) and the subset MQLR statistic (Figure 2). From Figure 1, we conclude that Case I leads to the largest variation in the rejection frequencies. Furthermore, we observe that the rejection frequencies of the 2SLS t-statistic increase as the endogeneity becomes stronger. The greatest variation is seen when $\rho_{\varepsilon v}$ is varied, the parameter that measures the endogeneity of the regressor whose coefficient is under test. When $\rho_{\varepsilon v}$ is medium to high, the 2SLS t-statistic exhibits huge size distortions and rejection frequencies as high as 95% can be observed. Moreover, the rejection frequencies do not seem to depend heavily on $\rho_{\varepsilon u}$ and ρ (variation within a group of 9 consecutive bars is less than the variation between the three groups). When β is moderately identified, i.e. Case III and IV, the overrejection is only halve that of Case I and II, although large size distortions are still present. Figure 2 shows the same information but now for the subset MQLR statistic. Notice the significant change in scale of the vertical axis, which represents the rejection frequencies. Only in Case II, the subset MQLR statistic has a little tendency to overreject when endogeneity is small to medium. As expected, rejection frequencies are very near the 5% level for Case IV. When γ is weakly identified, the subset MQLR statistic can lead to very conservative inference. This is especially so, when ρ is low. Overall, it seems that all extreme outcomes in the rejection frequencies are included when the Monte Carlo design is limited to the equal correlation case, i.e. $\rho_{\varepsilon v} = \rho_{\varepsilon u} = \rho \in \{0.3, 0.6, 0.9\}$.

Table 1 shows among other things the eigenvalues that result for the chosen Monte Carlo design. When one (or both) of the parameters is weakly identified, the smallest eigenvalue ranges from $1.1 - 2.0$. The rejection frequencies reported in this table are obtained under the null hypothesis and they are based on 5,000 Monte Carlo replications, $n = 200$ and $k = 10$. Besides the results for the 2SLS t-statistic and the subset MQLR statistic, the table shows the rejection frequencies for the t-statistics based on Fuller and LIML. Although their rejection frequencies are highly correlated with the ones obtained by 2SLS, the overrejection is much smaller for Fuller and the smallest for LIML t-statistics. Although the reduction of the error in rejection probability (ERP) can be substantial, huge ERPs can still be observed when correlation is high.

Table 2 shows the rejection frequencies under the null for inference based on the 2SLS t-statistic when using the asymptotic critical value and bootstrap critical values obtained by different bootstrap DGPs when $k = 5, 10$ and $n = 200$. The use of the RE bootstrap seems to control the size when the correlation is low or medium, i.e. $\rho \in \{0.3, 0.6\}$. When the correlation is low, the test even becomes conservative. However, when the correlation is high, the rejection frequencies are sometimes as high as 3 times the nominal level especially when β is weakly identified, i.e. Case I and II. The use of the RFE Fuller bootstrap is able to bring down the high rejection frequencies that are observed in the high correlation case. Overall, the RFE bootstrap seems to control the size, although many rejection frequencies are now significantly lower than the significance level. The last column shows the rejection frequencies based on the RFE LIML bootstrap. Again the size seems to be correct, but inference based on the bootstrapped 2SLS t-statistic becomes very conservative. In some cases, the rejection frequencies are only halve the values that are obtained by the RFE Fuller bootstrap. The use of the efficient KLM estimates $\tilde{\Pi}(\beta_0)$ in the bootstrap DGP seems to have little impact on the rejection frequencies. However, a significant increase in the frequencies is observed in Case I when correlation is high leading to significant overrejection.

Table 3 shows the rejection frequencies under the null for inference based on the Fuller t-statistic. It seems reasonable to only consider RFE bootstrap DGPs, since 2SLS is less efficient than Fuller-based estimation. The overrejection as seen when using the asymptotic critical value is completely eliminated if inference is based on the classic RFE Fuller bootstrap. None of the rejection frequencies are larger than the nominal significance level. Again, the use of $\tilde{\Pi}(\beta_0)$ in the bootstrap DGP has only a small effect on the rejection frequencies, although again the rejection frequency becomes significantly too high in Case I when $\rho = 0.9$. Looking at the results of the RFE LIML bootstrap, we see that the test can become quite conservative similar to the 2SLS t-statistic case. In fact, there is little difference between the t-statistic based on 2SLS or Fuller under the RFE LIML bootstrap DGP. Interestingly, the Fuller t-statistic under the RFE Fuller bootstrap leads to smaller ERPs than the 2SLS t-statistic under the RFE Fuller bootstrap in case II and $k = 10$. Apparently, the robustness of the Fuller estimator to many weak instruments shows up in the simulation.

Table 4 shows the rejection frequencies under the null for inference based on the LIML t-statistic. Looking at the rejection frequencies based on the asymptotic critical value, the test is conservative for all correlation values considered in Case III, i.e. β/γ is moderately/weakly identified. Furthermore, the ERPs are smaller than the ERPs of the 2SLS and Fuller t-statistic. In this sense, the LIML t-statistic is the most robust statistic of the Wald-type test statistics considered. Bootstrap inference based on the RFE Fuller bootstrap leads to a correct size, although there is little difference between the bootstrapped Fuller t-statistic and LIML t-statistic. The use of $\tilde{\Pi}(\beta_0)$ in the bootstrap DGP seems to be helpful in pushing the rejection frequencies to the nominal significance level in almost all cases. The results of the RFE LIML bootstrap for the LIML t-statistic are almost identical to the results based on the Fuller t-statistic under the RFE LIML bootstrap. Furthermore, the use of the RFE LIML bootstrap leads to the lowest rejection frequencies of all bootstrap DGPs.

Table 5 shows the rejection frequencies under the null for inference based on the subset MQLR statistic. As we have seen in Figure 2, the subset MQLR statistic can be very conservative when γ is weakly identified, i.e. Case I and III. The RFE bootstrap is able to make the subset MQLR statistic less conservative, although the reduction in ERP is much better for $k = 5$ than for $k = 10$. Interestingly, the use of $\tilde{\Pi}(\beta_0)$ in the bootstrap DGP reduces the ERP for both $k = 5$ and $k = 10$, except in case I when $\rho = 0.9$. There is almost no difference between inference based on the RFE LIML and the RFE Fuller bootstrap. Due to the fat tail problems associate with RFE LIML, we prefer the RFE Fuller bootstrap.

Table 6 contains the results for the GMM subset test statistic $\text{MQLR}_{\perp}(\beta_0)$ using the asymptotic p-value and its bootstrap analogue $\text{MQLR}^*_{\perp}(\beta_0)$ based on the GMM bootstrap. Comparing the rejection frequencies of $\text{MQLR}(\beta_0)$ to $\text{MQLR}_{\perp}(\beta_0)$, we see that the performance of the GMM-based $\text{MQLR}_{\perp}(\beta_0)$ is almost the same as $\text{MQLR}(\beta_0)$. Apparently, the effect of using an Eicker-White covariance matrix when no heteroskedasticity is present on the size is very small. However, the performance of inference based on the GMM bootstrap is quite disappointing. For $n = 200$, most of the rejection frequencies of $\text{MQLR}^*_{\perp}(\beta_0)$ can be substantial lower than $\text{MQLR}_{\perp}(\beta_0)$. To some extent, this is due to the fact that n is rather small. When the sample size is enlarged to 500, we see some improvement for Case II and IV. However, inference based on the GMM bootstrap remains extremely conservative even for $n = 1000$. This somewhat unexpected result clearly shows that the residual-based bootstrap DGPs are clearly superior to the GMM bootstrap procedure. Hence, the GMM bootstrap will not be considered in the remainder of the paper.

Next, we look at the power properties of some of the test statistics as shown in Figures 3 and 4. Since power comparisons are only meaningful if the size is correct, only the RFE Fuller bootstrap DGP is considered. To keep the number of figures manageable, the results are shown when the correlation is low or high, i.e. $\rho \in \{0.3, 0.9\}$. In Case I with $\rho = 0.3$, when both β and γ are weakly identified, all the considered test statistics have virtually no power since their rejection frequencies stay well below the 5% level. This is not totally unexpected since any test statistic will have low power due to the poorly identified parameters. In Case I with $\rho = 0.9$, all test statistics display at least some power, although the Fuller and LIML t-statistics also show biased behavior, i.e. the rejection frequencies reach their minimum value for β_0 different from the true value of β. Irrespective of the correlation, however, the estimated power curve of the bootstrapped MQLR statistic lies above the estimated power curve of the MQLR statistic using the asymptotic p-value. In Case IV, when both β and γ are moderately identified, all estimated power curves are nearly identical, although the Fuller and LIML t-statistics exhibit a drop in their curves for values far from the true value when $\rho = 0.9$. In Case II, when β/γ is weakly/moderately identified, we again see that the use of the Fuller and LIML t-statistics result in biased tests, especially when $\rho = 0.9$. The MQLR and its bootstrapped version lead to almost identical inference. Surprisingly, in Case III, when β/γ is moderately/weakly identified, the Fuller and LIML based t-statistics are hugely more powerful than the MQLR and MQLR* statistics. Overall, we conclude that tests based on the Fuller and LIML t-statistics can be biased. Tests based on the MQLR and MQLR* statistics are size controlled, unbiased, and their power curves depends on the identification strength of the parameters and the correlation structure of the disturbances. It is unfortunate that the bootstrap is able to reduce the conservativeness of the MQLR statistic in the case when it is least helpful, i.e. when the power is low.

5 Empirical Example

In this section, we shall compare various 95% confidence intervals by inverting several asymptotic and bootstrap tests in an empirical example. Blomquist and Dahlberg (1999) investigate inference based on IV methods in estimating linearized labor supply functions when the budget constraints are non-linear. As noted by Blomquist (1996), a non-linear tax/transfer system causes individuals' budget constraints to be non-linear inducing endogeneity of net wage rates and non-labor income. He used actual data to generate the hours of work in the presence of non-linear budget constraints of 602 Swedish married men between 25 and 55 years of age; see also Johansson et al. (2010). For individual i, the model reads (using the notation of this paper)

$$\begin{aligned} y_i &= \beta x_i + \gamma w_i + X_i \psi_y + \varepsilon_i \\ x_i &= Z_i \pi_x + X_i \psi_x + v_i \\ w_i &= Z_i \pi_w + X_i \psi_w + u_{i,} \end{aligned}$$

where y_i are hours worked, x_i is the hourly wage rate and w_i is non-labor income. The vector X_i contains the included exogenous variables, while the vector Z_i contains the excluded exogenous variables. There are three exogenous variables: (i) a constant (ii) a dummy for age and (iii) the number of children, and there are 26 socio-demographic variables that are used as instruments like dummies for the educational level of the individual, his wife, father and mother, a dummy indicating home ownership, dummies for the region where the individual lives, and the number of children in three different age groups. This equation was also considered in Flores-Lagunes (2007) and we follow him by dividing the net wage rate by 100 to scale its coefficient.

First, we assess the quality of the instruments. The partial R^2 of Shea (1997) is 0.0790 (0.0356) for the hourly wage rate x_i and 0.1656 (0.1264) for the non-labor income w_i (adjusted R^2 in parentheses). The F-version of the Cragg-Donald Wald statistic, $(n-k)/k_2 g_{\min}$, where k denotes the total number of instruments and k_2 is the number of excluded instruments is only 1.362. The Stock and Yogo (2005b) critical values for the Cragg-Donald F

statistic for LIML t-statistics in case of two endogenous regressors are 1.96 (for 20% max. LIML size) and 1.78 (for 25% max. LIML size), so the observed test statistic is smaller than the critical value for a maximum rejection probability of 25% when the significance level is 5%. Based on this test procedure, the instruments can be classified as weak. The Sargan statistic for overidentication is 18.067 with an asymptotic p-value of 0.80. Assuming homoskedasticity, Figure 5 shows 1 minus the p-value based on asymptotic and RFE Fuller bootstrap distributions for relevant values of (β_0, γ_0) when testing $H_0 : \beta = \beta_0$ or $H_0 : \gamma = \gamma_0$. Interestingly, there is a large difference between the p-values of the MQLR tests and the Wald-type tests. Since none of the p-values based on the asymptotic or bootstrap distribution for the MQLR statistic are below 2.5%, the 95% confidence intervals are unbounded for β as well as γ. This is just indicating that the data is uninformative about these two parameters. The p-values for the bootstrapped 2SLS t-statistic behave rather oddly, but lead to the same conclusion as the MQLR statistic. In effect, its p-value does not decrease since the bootstrap distribution is not centered at zero due to the severe bias in the 2SLS estimates. Applying Fuller or LIML leads to nearly identical results, so only the results based on LIML t-statistics are shown. The confidence intervals based on inverting the bootstrapped LIML t-statistics are finite and shown in the upper panel of Table 7. From this table, we see that the bootstrapped confidence intervals based on the LIML t-statistics are much wider than the intervals based on the asymptotic approximation.

In this example, we obtain conflicting results, which might be due to the inclusion of too many instruments. Hence, we looked into the specification of the four sets of education dummy variables. It appears that there are eight classification levels for the education of the husband and wife. The reference class (not represented by a dummy) refers to ‘not completed primary school’, which account for less than 1% of the sample. Hence, we combine level 1 and level 2 (‘completed primary school’) to act as the reference class. Next, we combine level 3-5, i.e. take the sum of the mutually exclusive dummy variables, which is mainly concerned with occupational training, into a dummy indicating low level education. Medium level education is used to indicate educational level 6 and 7, while high level education is used to indicate the highest educational level 8 (‘completed university’). Next, we have looked into the five classification levels for the education of the father and mother. It seems reasonable to assume that the educational levels of the parents are correlated with the education level of their son. Since we already included the educational level of the son, the contribution of the parents' educational level might be limited. Moreover, there is much less variation in the parents' educational level since 76% of the fathers and 83% of the mothers have only completed primary school. Looking at the significance of the coefficients of the education dummies, the only relevant educational level seems to be whether or not the father has obtained a university degree or other similar qualifications. In all, the number of education dummies is reduced from 21 to 7, which leaves us with 15 instrumental variables.

Based on this smaller set of instruments, the partial R^2 is 0.066 (0.0460) for the hourly wage rate x_i and 0.1476 (0.1287) for the non-labor income w_i (adjusted R^2 in parentheses). The F-version of the Cragg-Donald Wald statistic equals 2.524, which is almost equal to the critical value for a maximum rejection probability of 10% when the significance level is 5%. Sargan's test statistic for overidentification becomes 9.683 with an asymptotic p-value of 0.47. Figure 6 shows 1 minus the p-value for the various test statistics after the new education classification. All confidence intervals become finite and are shown in the bottom panel of Table 7. The intervals based on the bootstrapped 2SLS t-statistics are again erratic due to the high p-values when testing values far from the center. The bootstrap intervals based on the FIML t-statistic are again wider compared to the intervals based on the asymptotic approximation. Interestingly, for the intervals based on the MQLR statistic, we see that the bootstrapped intervals are somewhat shorter than their asymptotic counterparts. Overall, the empirical example shows that inference can heavily depend on the number of included (weak) instruments. Especially the strange behavior of the bootstrapped 2SLS t-statistic is evidence against using this test statistics.

6 CONCLUSION

In this paper, we have looked at bootstrapping Wald-type and weak instrument robust subset test statistics in the linear regression model containing two endogenous regressors. We propose the RFE (restricted fully efficient) bootstrap that incorporates also efficient estimates of the structural equation as a modification of the RE (restricted efficient) bootstrap suggested by Davidson and MacKinnon (2010). Besides the residual-based RFE bootstrap, we have also adapted the GMM bootstrap proposed by Kleibergen (2011) to the subset testing setting. In addition to the commonly used Wald-type test statistics, we included the robust subset MQLR test statistic as proposed and investigated by Kleibergen and Mavroeidis (2011). The important features of this test statistic are that it uses efficient estimates, is robust against weak identification and is as easy to use that the CLR statistic in the single endogenous regressor case. Furthermore, its GMM counterpart allows for heteroskedasticity of unknown form.

In the simulation study, huge size distortions are observed for t-statistics based on 2SLS, LIML and Fuller estimators. We find that the RFE bootstrap performs significantly better than the RE bootstrap and is able to control the size for the LIML and Fuller based t-statistics when the instruments are weak. With respect to the size, the RFE bootstrap DGP based on the Fuller estimator performs best, even when the LIML t-statistic is considered. Of course, the RFE bootstrap cannot save the Wald-type test statistics when the parameter values are taken to their extremes, e.g. the case when all instruments become irrelevant. This, however, is not the case for the robust subset MQLR statistic, which remains valid even if the parameters are not identified. When the parameter of the endogenous regressor that is not being tested is weakly identified, inference based on the subset MQLR becomes conservative. We see that the RFE bootstrap is able to reduce this conservativeness although sometimes only marginally. In the simulations, the bootstrapped subset MQLR statistic is a little bit more powerful, however, the difference is usually small. Although the GMM counterpart of the subset MQLR statistic has comparable size properties as the ordinary MQLR statistic, the results of its GMM bootstrapped version are disappointing: no improvement over the asymptotic approximation is observed.

In the empirical example about estimating linearized labor supply functions when the budget constraints are non-linear, the finite confidence intervals based on inverting the RFE bootstrapped LIML t-test are in conflict with the infinite confidence intervals based on inverting the subset MQLR test. Apparently, the situation in our empirical example is such that the bootstrapped LIML t-statistic is more powerful than the subset MQLR statistic. However, when a number of insignificant instruments is dropped, the intervals based on both approaches become finite and are more in line with each other.

7 Appendix: Asymptotic p-value of MQLR

When π_w is a fixed non-trivial value, Theorem 1 of Kleibergen and Mavroeidis (2011) shows that

$$\begin{aligned}\text{MQLR}(\beta_0)|\tilde{\Pi}(\beta_0) \overset{d}{\to} \; & \tfrac{1}{2}[\psi_1 + \psi_{k-2} - \text{rk}(\beta_0) + \\ & \sqrt{(\psi_1 + \psi_{k-2} + \text{rk}(\beta_0))^2 - 4\psi_{k-2}\text{rk}(\beta_0)}],\end{aligned}$$

where $\tilde{\Pi}(\beta_0) = (\tilde{\pi}_x(\beta_0) \,\vdots\, \tilde{\pi}_w(\beta_0))$ and ψ_1 and ψ_{k-2} are independent χ^2_1 and χ^2_{k-2} distributed random variables. To derive an asymptotic approximation of the p-value for MQLR(β_0), we use the results obtained by Davidson and MacKinnon (2011). Using the identity

$$(\psi_1 + \psi_{k-2} + \text{rk}(\beta_0))^2 - 4\psi_{k-2}\text{rk}(\beta_0) = (\psi_1 + \psi_{k-2} - \text{rk}(\beta_0))^2 + 4\psi_1\text{rk}(\beta_0),$$

we can reformulate the asymptotic distribution as

$$\tfrac{1}{2}\left[\psi_1 + \psi_{k-2} - \text{rk}(\beta_0) + \sqrt{(\psi_1 + \psi_{k-2} - \text{rk}(\beta_0))^2 + 4\psi_1\text{rk}(\beta_0)}\right],$$

which is the asymptotic equivalence of equation (A1) of Davidson and MacKinnon (2011) for $Z^2 \overset{a}{\sim} \psi_1$ and $Y \overset{a}{\sim} \psi_{k-2}$. Making use of equation (A3) derived in Davidson and MacKinnon (2011), we obtain the following expression for the asymptotic CDF of MQLR(β_0)

$$F_{as}(x, \text{rk}(\beta_0)|\tilde{\Pi}(\beta_0)) = \sqrt{\frac{2}{\pi}} \int_0^{\sqrt{x}} F_{\chi^2_{k-2}}\left((x + \text{rk}(\beta_0))(1 - z^2/x)\right) e^{-z^2/2} dz,$$

where $F_{\chi^2_{k-2}}(\cdot)$ denotes the cumulative distribution function of a (central) χ^2-distribution with $k-2$ degrees of freedom. The integral can be approximated easily by numerical integration methods for given values of x and rk(β_0). Of course, the asymptotic p-value conditional on $\tilde{\Pi}(\beta_0)$ is given by $1 - F_{as}(\text{MQLR}(\beta_0),\text{rk}(\beta_0)|\tilde{\Pi}(\beta_0))$. Theorem 6 of Kleibergen and Mavroeidis (2011) shows that this value is an upper bound for the actual p-value without imposing any restrictions on π_w.

References

Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics 20*, 46–63.

Andrews, D. W. K., M. Moreira, and J. H. Stock (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica 74*, 715–752.

Blomquist, S. (1996). Estimation methods for male labor supply functions: how to take account of nonlinear taxes. *Journal of Econometrics 70*, 383–405.

Blomquist, S. and M. Dahlberg (1999). Small sample properties of LIML and jackknife IV estimators: experiments with weak instruments. *Journal of Applied Econometrics 14*, 69–88.

Chaudhuri, S., T. Richardson, J. Robins, and E. Zivot (2010). A new projection-type split-sample score test in linear instrumental variables regression. *Econometric Theory 26*, 1820–1837.

Chaudhuri, S. and E. Zivot (2011). A new method of projection-based inference in GMM with weakly identified nuisance parameters. *Journal of Econometrics 164*, 239–251.

Davidson, R. and J. G. MacKinnon (2008). Bootstrap inference in a linear equation estimated by instrumental variables. *Econometrics Journal 11*, 443–477.

Davidson, R. and J. G. MacKinnon (2010). Wild bootstrap tests for IV regression. *Journal of Business and Economic Statistics 28*, 128–144.

Davidson, R. and J. G. MacKinnon (2011). Bootstrap confidence sets with weak instruments. Discussion paper, McGill-Queen's University.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica 65*, 1365–1387.

Dufour, J.-M. (2003). Identification, weak instruments and statistical inference in econometrics. *Canadian Journal of Economics 36*, 767–808.

Dufour, J.-M. and J. Jasiak (2001). Finite sample limited information inference methods for structural equations and models with generated regressors. *International Economic Review 42*, 815–843.

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). Projection-based statistical inference in linear structural models with possibly weak 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. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *The Annals of Statistics 12*, 827–842.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica 45*, 939–953.

Hahn, J., J. Hausman, and G. Kuersteiner (2004). Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. *Econometrics Journal 7*, 272–306.

Hall, P. and J. L. Horowitz (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica 4*, 891–916.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics 14*, 262–280.

Johansson, S., R. Erikson, J. Jonsson, and M. Tåhlin (2010). Survey of levels of living in sweden [file: LNU81]. Stockholm University, Swedish Institute for Social Research, Sweden, Swedish National Data Service (SND) [distributor].

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica 70*, 1781–1803.

Kleibergen, F. (2004). Testing subsets of structural parameters in the instrumental variables regression model. *Review of Economic Studies 86*, 418–423.

Kleibergen, F. (2011). Improved accuracy of weak instrument robust GMM statistics through bootstrap and edgeworth approximations. Discussion paper, Brown University.

Kleibergen, F. and S. Mavroeidis (2009). Weak instrument robust tests in GMM and the new keynesian phillips curve. *Journal of Business & Economic Statistics 27*, 293–311.

Kleibergen, F. and S. Mavroeidis (2011). Inference on subsets of parameters in linear IV without assuming identification. Discussion paper, Brown University.

Mikusheva, A. (2010). Robust confidence sets in the presence of weak instruments. *Journal of Econometrics 157*, 236–247.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica 71*, 1027–1048.

Moreira, M. J., J. Porter, and G. Suarez (2009). Bootstrap validity of the score test when instruments may be weak. *Journal of Econometrics 149*, 52–64.

Shea, J. (1997). Instrument relevance in multivariate linear models: a simple measure. *Review of Economic Studies 79*, 348–352.

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

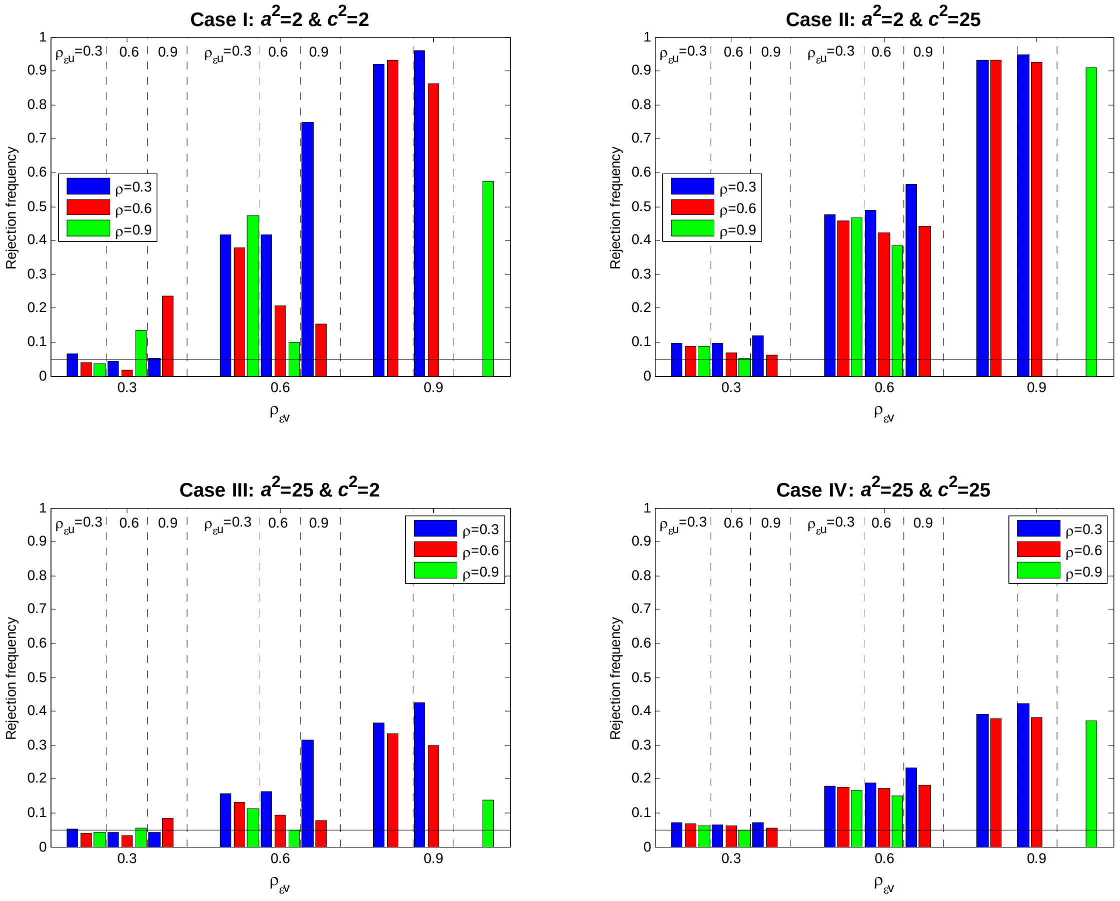
Stock, J. H. and J. H. Wright (2000). GMM with weak identification. *Econometrica 68*, 1055–1096.

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 & Economic Statistics 20*, 518–529.

Stock, J. H. and M. Yogo (2005a). Asymptotic distributions of instrumental variables statistics with many instruments. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 109–120. Cambridge: Cambridge University Press.

Stock, J. H. and M. Yogo (2005b). Testing for weak instruments in linear IV regression. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 80–108. Cambridge: Cambridge University Press.

Figure 1: Rejection frequencies of $H_0 : \beta = \beta_0$ based on the 2SLS t-statistic for $n = 200, k = 10$ and 20,000 replications. Horizontal axis refers to the endogeneity of x ($\rho_{\varepsilon v}$). First/second/third set of 3 bars refers to $\rho_{\varepsilon u} = 0.3/0.6/0.9$.



Case I: a^2=2 & c^2=2

Case II: a^2=2 & c^2=25

Case III: a^2=25 & c^2=2

Case IV: a^2=25 & c^2=25

Figure 2: Rejection frequencies of $H_0 : \beta = \beta_0$ based on the subset MQLR statistic for $n = 200, k = 10$ and 20,000 replications. Horizontal axis refers to the endogeneity of x $(\rho_{\varepsilon v})$. First/second/third set of 3 bars refers to $\rho_{\varepsilon u} = 0.3/0.6/0.9$.

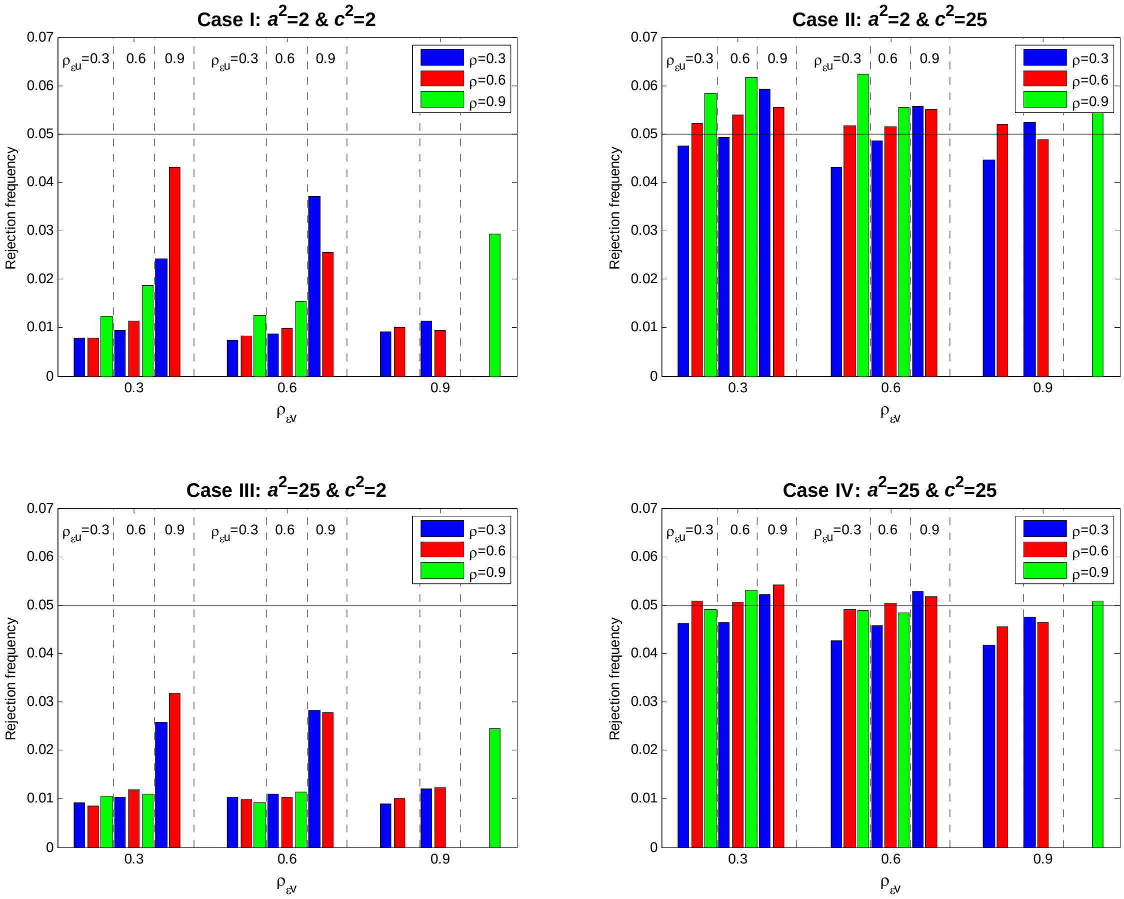


Figure 3: Power curves when testing $H_0 : \beta = \beta_0$ for $\beta = 0.5$, $\beta_0 \in \{-0.5, ..., 1.5\}$, $n = 200$ and $k = 10$ based on 5,000 Monte Carlo replications and 999 bootstrap replications.

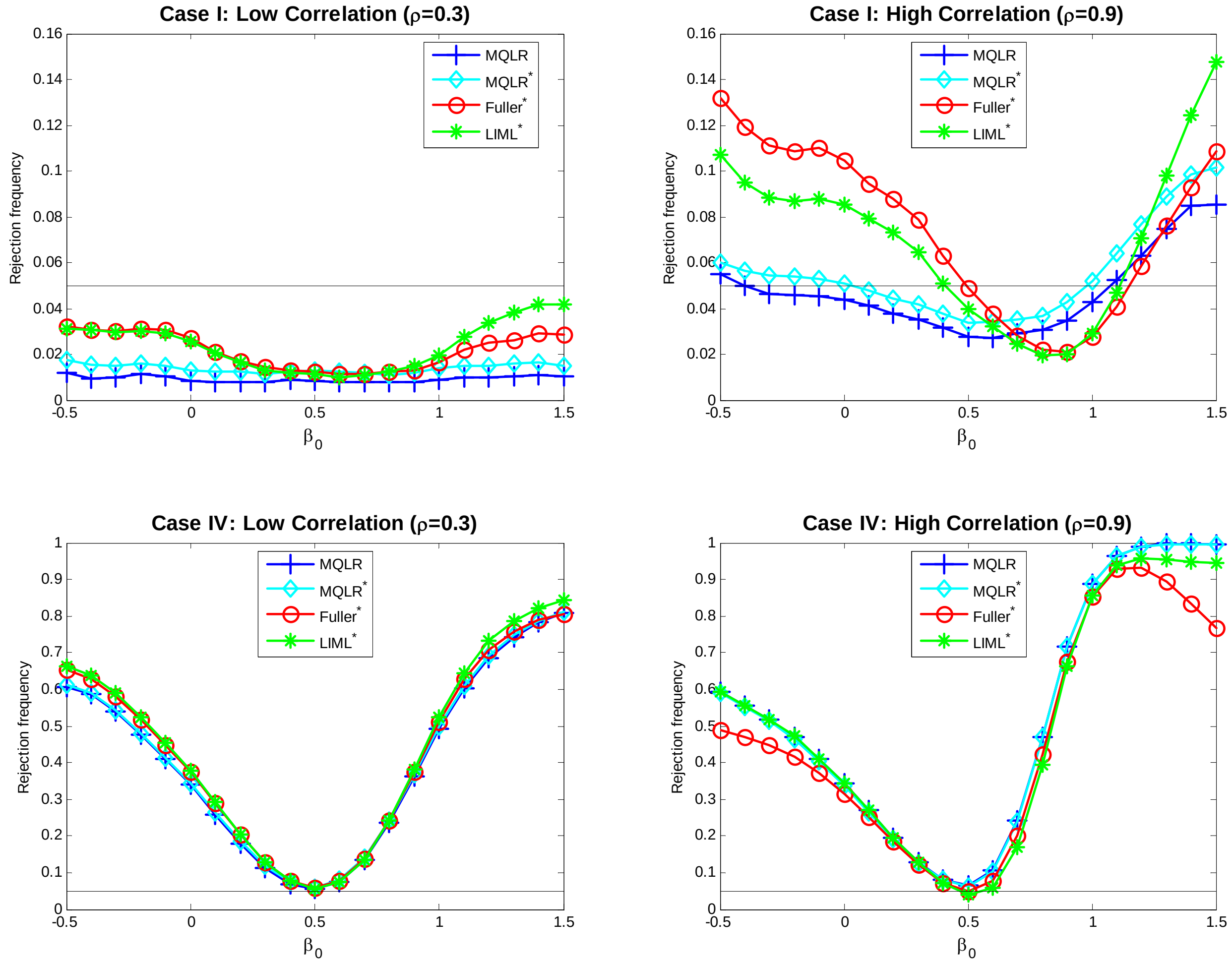
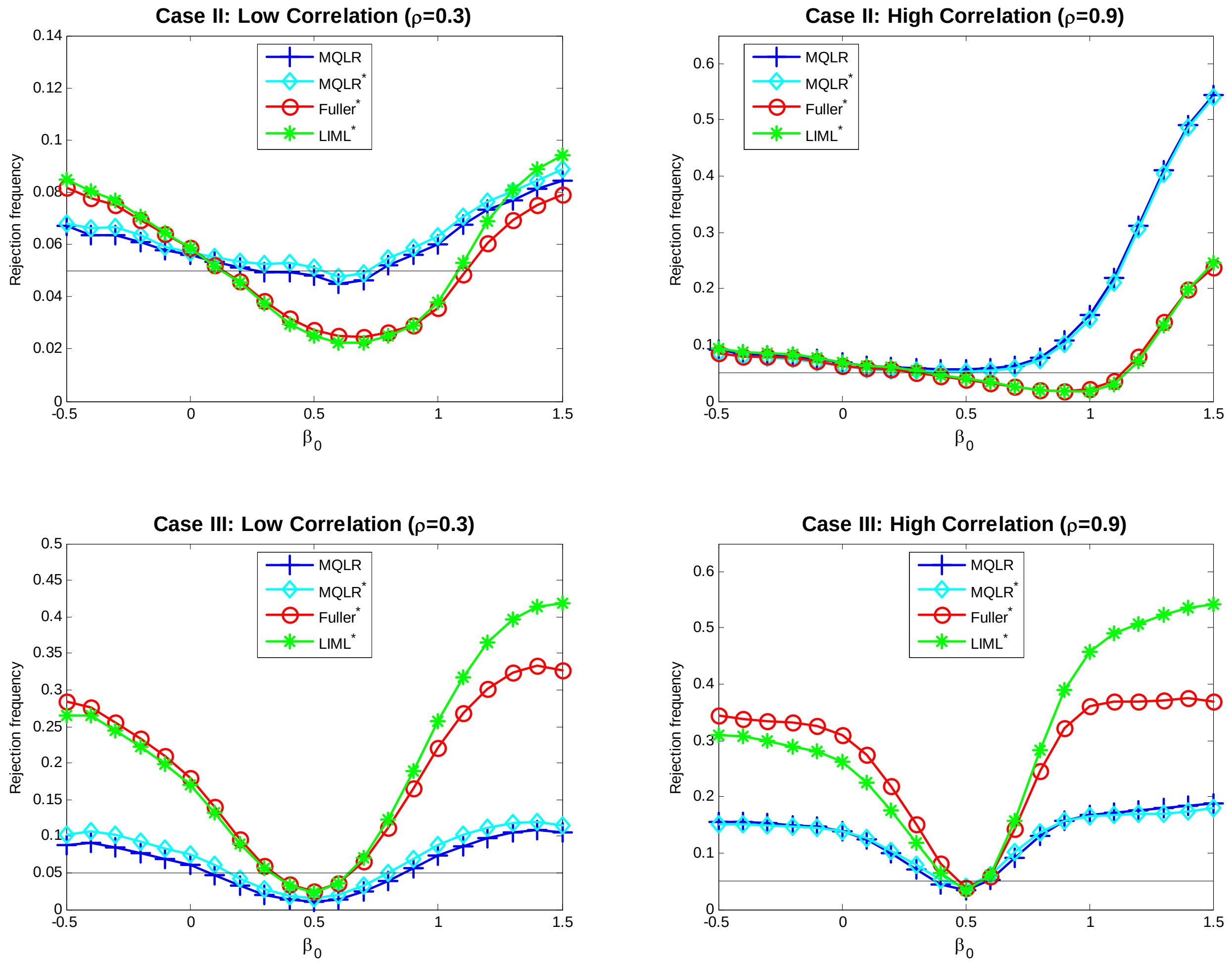


Figure 4: Power curves when testing $H_0 : \beta = \beta_0$ for $\beta = 0.5$, $\beta_0 \in \{-0.5, ..., 1.5\}$, $n = 200$ and $k = 10$ based on 5,000 Monte Carlo replications and 999 bootstrap replications.



Case II: Low Correlation (ρ=0.3)

Case II: High Correlation (ρ=0.9)

Case III: Low Correlation (ρ=0.3)

Case III: High Correlation (ρ=0.9)

Figure 5: 1 minus p-value for $H_0 : \beta = \beta_0$ (upper graph) and $H_0 : \gamma = \gamma_0$ (lower graph) based on asymptotic and the RFE bootstrap (99,999 bootstrap replications).

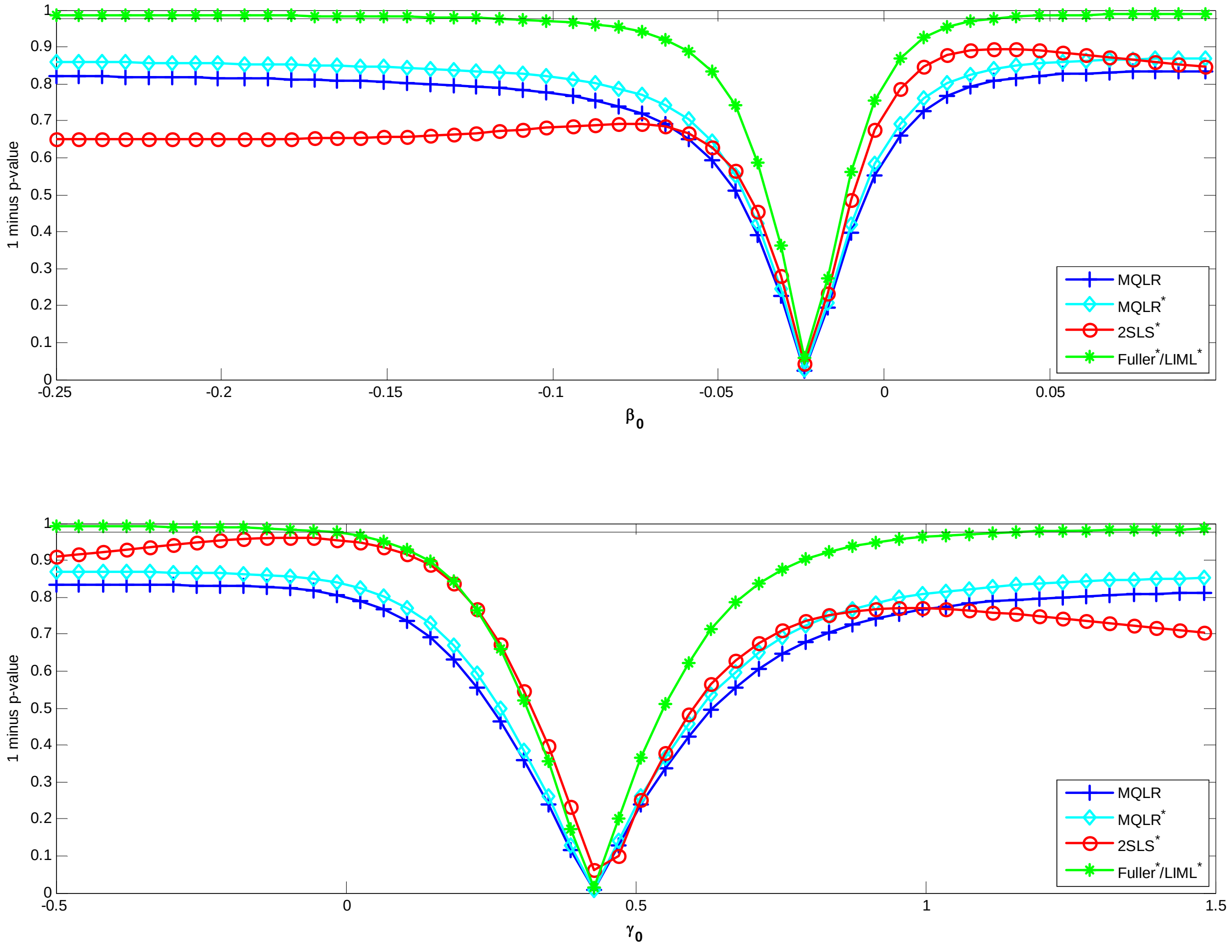
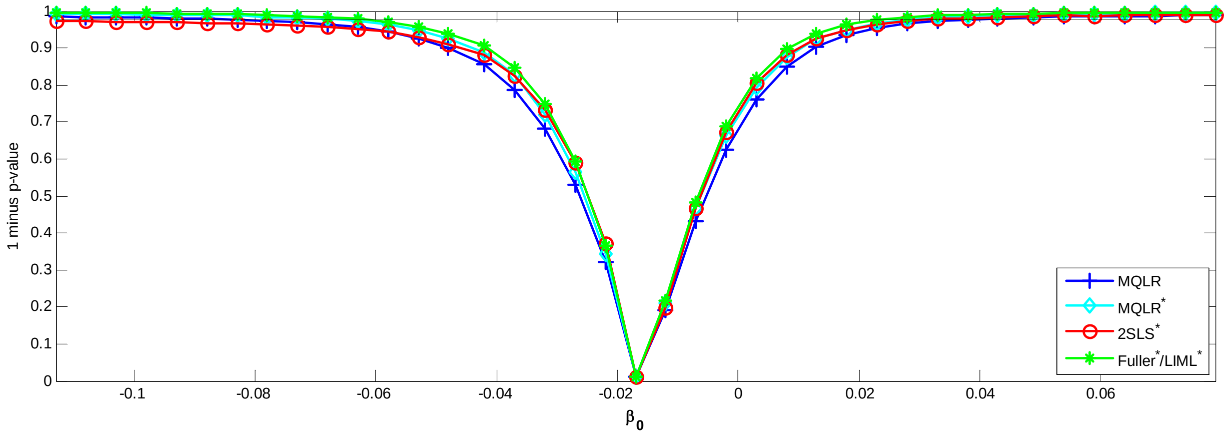


Figure 6: 1 minus p-value for $H_0 : \beta = \beta_0$ (upper graph) and $H_0 : \gamma = \gamma_0$ (lower graph) based on asymptotic and the RFE bootstrap (99,999 bootstrap replications).



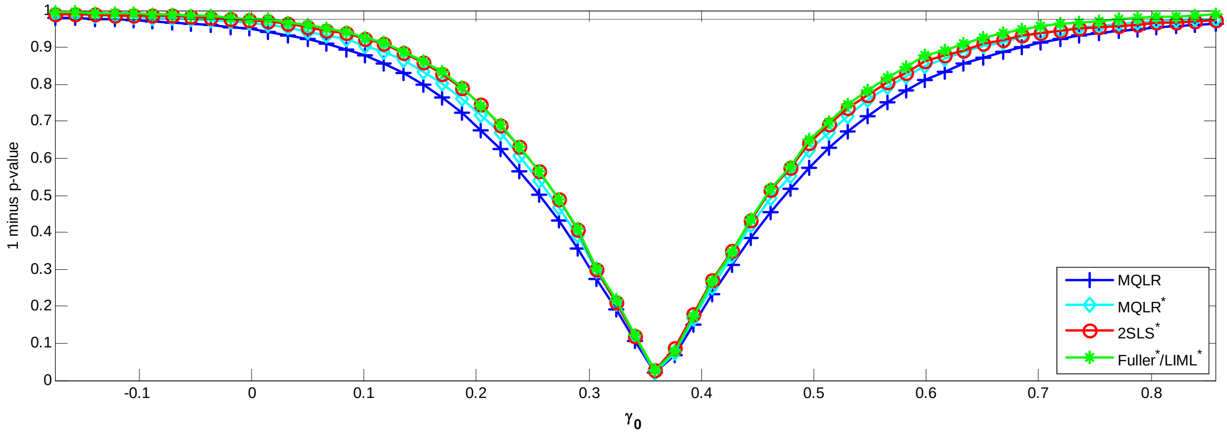


Table 1: Rejection frequencies (in percentage points) at $\alpha = 5\%$ under the null hypothesis $H_0 : \beta = \beta_0$ based on 5,000 Monte Carlo replications, $k = 10$ and sample size $n = 200$. The λ's denote the eigenvalues of the concentration matrix.

a^2	c^2	λ_-	λ_+	Case	ρ	2SLS	Fuller	LIML	MQLR
2	2	1.5	2.9	I	0.30	5.8	3.6	2.8	0.8
2	2	1.3	5.0		0.60	21.9	11.7	8.4	1.1
2	2	1.1	20.0		0.90	58.7	27.8	15.2	2.7
2	25	2.0	27.7	II	0.30	9.1	7.0	6.1	4.5
2	25	1.9	40.2		0.60	43.3	22.0	18.0	5.0
2	25	1.9	140.2		0.90	90.9	43.1	25.4	5.7
25	2	2.0	27.7	III	0.30	5.7	3.7	2.8	1.0
25	2	1.9	40.2		0.60	9.0	4.5	3.4	1.0
25	2	1.9	140.2		0.90	14.0	4.5	2.2	2.6
25	25	19.2	35.7	IV	0.30	6.3	5.2	5.0	3.9
25	25	15.6	62.5		0.60	16.1	6.8	6.1	4.9
25	25	13.2	250.0		0.90	36.4	7.8	6.1	4.7

Table 2: Rejection frequencies (in percentage points) at $\alpha = 5\%$ for the 2SLS t-statistic under the null hypothesis $H_0 : \beta = \beta_0$ based on 5,000 Monte Carlo replications, 399 bootstrap replications and sample size $n = 200$.

				RE 2SLS		RFE Fuller		RFE LIML
k	Case	ρ	Asymp.	Classic	$\tilde{\Pi}(\beta_0)$	Classic	$\tilde{\Pi}(\beta_0)$	
5	I	0.30	1.8	1.7	2.3	1.1	1.5	0.5
		0.60	8.2	3.2	3.7	1.6	2.4	0.6
		0.90	33.3	11.9	15.5	4.4	9.0	0.7
	II	0.30	3.5	3.1	2.9	2.6	2.6	2.5
		0.60	19.1	4.1	3.8	2.5	2.4	2.2
		0.90	57.8	7.4	6.5	2.3	2.2	2.1
	III	0.30	3.1	3.5	4.0	2.2	2.7	1.1
		0.60	4.1	3.0	3.3	1.7	2.2	0.9
		0.90	6.4	4.1	4.7	2.1	3.1	0.8
	IV	0.30	5.0	4.5	4.4	4.3	4.3	4.3
		0.60	8.9	5.0	4.8	4.8	4.7	4.8
		0.90	14.7	5.3	4.7	4.0	4.1	4.1
10	I	0.30	5.8	2.8	3.0	1.3	1.7	0.7
		0.60	21.9	4.1	5.2	1.5	2.5	0.7
		0.90	58.7	17.1	21.9	4.1	10.2	0.9
	II	0.30	9.1	3.6	3.4	3.0	3.0	2.9
		0.60	43.3	7.0	6.5	2.9	2.9	2.6
		0.90	90.9	18.5	18.1	1.5	1.5	1.3
	III	0.30	5.7	4.1	4.1	2.1	2.6	1.3
		0.60	9.0	4.1	4.2	2.1	2.7	1.2
		0.90	14.0	6.2	6.8	2.3	3.7	0.9
	IV	0.30	6.3	4.6	4.3	4.0	4.0	3.9
		0.60	16.1	5.6	5.1	4.0	3.8	3.8
		0.90	36.4	8.3	5.8	3.1	3.2	3.1

Table 3: Rejection frequencies (in percentage points) at $\alpha = 5\%$ for the Fuller t-statistic under the null hypothesis $H_0 : \beta = \beta_0$ based on 5,000 Monte Carlo replications, 399 bootstrap replications and sample size $n = 200$.

				RFE Fuller		RFE LIML
k	Case	ρ	Asymp.	Classic	$\tilde{\Pi}(\beta_0)$	
5	I	0.30	1.5	1.0	1.3	0.5
		0.60	6.7	1.6	2.3	0.7
		0.90	23.5	4.1	7.1	0.8
	II	0.30	3.0	2.4	2.4	2.4
		0.60	13.9	2.7	2.5	2.5
		0.90	34.7	2.0	1.7	1.7
	III	0.30	2.8	2.4	2.8	1.1
		0.60	3.3	2.1	2.5	1.2
		0.90	3.5	2.1	2.4	1.1
	IV	0.30	4.4	4.3	4.4	4.3
		0.60	6.5	4.6	4.6	4.5
		0.90	7.9	4.3	4.5	4.4
10	I	0.30	3.6	1.1	1.6	0.7
		0.60	11.7	1.6	2.5	0.8
		0.90	27.8	4.8	7.9	1.6
	II	0.30	7.0	2.5	2.4	2.4
		0.60	22.0	3.6	3.4	3.1
		0.90	43.1	4.1	3.8	3.7
	III	0.30	3.7	2.1	2.7	1.3
		0.60	4.5	2.4	2.7	1.5
		0.90	4.5	2.2	2.9	1.4
	IV	0.30	5.2	4.1	4.0	4.0
		0.60	6.8	4.3	4.3	4.3
		0.90	7.8	3.4	3.4	3.4

Table 4: Rejection frequencies (in percentage points) at $\alpha = 5\%$ for the LIML t-statistic under the null hypothesis $H_0 : \beta = \beta_0$ based on 5,000 Monte Carlo replications, 399 bootstrap replications and sample size $n = 200$.

				RFE Fuller		RFE LIML
k	Case	ρ	Asymp.	Classic	$\tilde{\Pi}(\beta_0)$	
5	I	0.30	1.0	1.0	1.3	0.5
		0.60	4.1	1.6	2.4	0.7
		0.90	10.5	3.1	4.2	1.3
	II	0.30	2.7	2.2	2.2	2.2
		0.60	10.7	2.8	2.6	2.5
		0.90	17.7	2.3	2.0	2.0
	III	0.30	2.0	2.5	2.8	1.3
		0.60	2.2	2.1	2.5	1.4
		0.90	1.7	1.7	1.7	1.3
	IV	0.30	4.1	4.5	4.5	4.5
		0.60	5.6	4.6	4.6	4.7
		0.90	6.5	4.1	4.3	4.3
10	I	0.30	2.8	1.0	1.2	0.6
		0.60	8.4	1.3	1.9	0.8
		0.90	15.2	4.2	5.2	2.0
	II	0.30	6.1	2.3	2.3	2.3
		0.60	18.0	3.6	3.6	3.3
		0.90	25.4	4.3	4.0	4.0
	III	0.30	2.8	2.1	2.4	1.4
		0.60	3.4	2.2	2.4	1.6
		0.90	2.2	1.7	1.8	1.6
	IV	0.30	5.0	4.1	4.0	4.0
		0.60	6.1	4.2	4.2	4.2
		0.90	6.1	3.3	3.4	3.4

Table 5: Rejection frequencies (in percentage points) at $\alpha = 5\%$ for the subset MQLR statistic under the null hypothesis $H_0 : \beta = \beta_0$ based on 5,000 Monte Carlo replications, 399 bootstrap replications and sample size $n = 200$.

				RFE Fuller		RFE LIML
k	Case	ρ	Asymp.	Classic	$\tilde{\Pi}(\beta_0)$	
5	I	0.30	0.9	1.5	2.8	1.6
		0.60	1.0	2.1	3.2	2.0
		0.90	3.4	4.6	7.1	4.5
	II	0.30	5.7	6.0	6.1	5.9
		0.60	4.9	5.0	5.1	5.0
		0.90	5.0	4.9	5.0	4.9
	III	0.30	1.1	1.7	2.7	1.7
		0.60	1.4	2.2	3.5	2.3
		0.90	2.6	3.8	5.8	3.7
	IV	0.30	4.2	4.7	4.8	4.8
		0.60	4.9	5.2	5.2	5.2
		0.90	5.1	5.2	5.2	5.2
10	I	0.30	0.8	1.3	2.1	1.3
		0.60	1.1	1.5	2.9	1.5
		0.90	2.7	3.3	5.1	3.3
	II	0.30	4.5	4.7	4.8	4.7
		0.60	5.0	5.0	5.0	4.9
		0.90	5.7	5.4	5.3	5.4
	III	0.30	1.0	1.3	2.0	1.4
		0.60	1.0	1.5	2.5	1.5
		0.90	2.6	3.0	4.7	3.1
	IV	0.30	3.9	4.0	4.1	4.0
		0.60	4.9	5.0	5.1	5.1
		0.90	4.7	4.4	4.6	4.5

Table 6: Rejection frequencies (in percentage points) at $\alpha = 5\%$ for the subset GMM statistic $\text{MQLR}_{\perp}$ under the null hypothesis $H_0 : \beta = \beta_0$ based on the GMM bootstrap, 5,000 Monte Carlo replications, 399 bootstrap replications.

			$n = 200$			$n = 500$		$n = 1000$	
k	Case	ρ	MQLR	$\text{MQLR}_{\perp}$	$\text{MQLR}^*_{\perp}$	$\text{MQLR}_{\perp}$	$\text{MQLR}^*_{\perp}$	$\text{MQLR}_{\perp}$	$\text{MQLR}^*_{\perp}$
5	I	0.3	1.0	0.7	0.7	0.6	0.5	0.8	0.6
		0.6	1.4	1.2	0.9	0.9	0.5	1.0	0.9
		0.9	2.9	2.5	3.2	2.8	1.5	2.9	1.4
	II	0.3	4.1	3.9	3.4	4.5	4.2	4.8	4.9
		0.6	4.7	4.5	3.7	4.6	4.3	5.1	4.9
		0.9	5.4	5.1	0.6	4.8	4.3	4.7	4.3
	III	0.3	1.0	0.9	0.5	0.7	0.5	0.9	0.8
		0.6	1.3	1.2	1.1	0.8	0.5	1.1	0.8
		0.9	3.3	2.9	3.1	2.8	1.3	2.1	1.1
	IV	0.3	4.2	4.3	3.2	4.6	4.1	3.9	3.8
		0.6	4.8	4.3	3.7	4.7	4.3	4.7	4.3
		0.9	4.9	4.4	0.1	4.9	4.3	4.8	4.5
10	I	0.3	0.6	0.5	0.1	0.7	0.4	0.6	0.4
		0.6	1.0	0.6	0.4	0.7	0.3	0.9	0.4
		0.9	2.9	2.0	1.8	2.9	0.8	2.3	0.9
	II	0.3	5.1	4.7	1.9	4.1	3.3	4.5	3.9
		0.6	5.0	4.6	1.9	4.6	3.4	4.1	3.4
		0.9	5.4	5.1	0.2	4.9	3.6	4.5	3.5
	III	0.3	1.1	1.1	0.2	0.7	0.4	0.8	0.4
		0.6	1.1	1.0	0.4	1.0	0.4	1.0	0.5
		0.9	2.8	1.9	1.0	2.0	0.6	2.2	0.9
	IV	0.3	4.0	3.5	1.6	3.6	2.4	3.7	2.9
		0.6	5.3	4.6	1.6	4.1	2.6	4.3	3.3
		0.9	4.7	4.4	0.0	5.1	3.7	4.3	3.3

Table 7: 95% confidence intervals for β and γ (LCL/UCL=lower/upper confidence limit).

Test Statistic	Method	LCL β	UCL β	LCL γ	UCL γ
Specification I: 29 instrumental variables					
2SLS t-statistic	Asymptotic	-0.035	-0.001	0.155	0.551
	RFE Bootstrap	$-\infty$	$+\infty$	$-\infty$	$+\infty$
LIML t-statistc	Asymptotic	-0.047	0.001	0.165	0.689
	RFE Bootstrap	-0.116	0.028	-0.035	1.157
MQLR	Asymptotic	$-\infty$	$+\infty$	$-\infty$	$+\infty$
	RFE Bootstrap	$-\infty$	$+\infty$	$-\infty$	$+\infty$
Specification II: 15 instrumental variables					
2SLS t-statistic	Asymptotic	-0.033	0.004	0.129	0.548
	RFE Bootstrap	-0.131	0.030	-0.017	0.932
LIML t-statistc	Asymptotic	-0.038	0.006	0.120	0.605
	RFE Bootstrap	-0.061	0.023	0.016	0.767
MQLR	Asymptotic	-0.080	0.035	-0.118	0.951
	RFE Bootstrap	-0.065	0.027	-0.040	0.837