JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **22**: 677–694 (2007)
Published online 13 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jae.916

FINITE SAMPLE EVIDENCE OF IV ESTIMATORS UNDER WEAK INSTRUMENTS

ALFONSO FLORES-LAGUNES*
Department of Economics, Eller College of Management, University of Arizona, Tucson, AZ, USA

SUMMARY

We present finite sample evidence on different IV estimators available for linear models under weak instruments; explore the application of the bootstrap as a bias reduction technique to attenuate their finite sample bias; and employ three empirical applications to illustrate and provide insights into the relative performance of the estimators in practice. Our evidence indicates that the random-effects quasi-maximum likelihood estimator outperforms alternative estimators in terms of median point estimates and coverage rates, followed by the bootstrap bias-corrected version of LIML and LIML. However, our results also confirm the difficulty of obtaining reliable point estimates in models with weak identification and moderate-size samples. Copyright © 2007 John Wiley & Sons, Ltd.

Received 31 August 2003; Revised 12 December 2005

1. INTRODUCTION

Linear instrumental variables (IV) estimation is commonly used in econometrics to avoid the problems caused when one or more explanatory variables are endogenous, that is, correlated with the disturbance term. A requirement of IV estimation is to have instrumental variables that are: (a) 'highly' correlated with the endogenous explanatory variables (i.e., 'relevant'); and (b) uncorrelated with the disturbance term ('exogenous'). In this paper, we focus on the finite sample problems of IV estimators when the first requirement above is not met, which has been called the problem of 'weak instruments'.

The literature on the finite sample problems of IV estimators under weak instruments has grown after Nelson and Startz (1990a, 1990b) and Maddala and Jeong (1992) pointed out that, under weak instruments, two-stage least squares (TSLS) is severely biased in finite samples towards the expectation of the inconsistent OLS estimator; which occurs even when using very large samples (Bound *et al*., 1995).[1] In addition, the standard error of TSLS is severely downward biased (Staiger and Stock, 1997; Hahn and Hausman, 2003), resulting in misleading statistical inference.

This literature focuses on several aspects. One of them proposes ways to detect weak instruments in practical situations (Hall *et al*., 1996; Shea, 1997; Hahn and Hausman, 2002b; Stock and Yogo, 2004). A second aspect proposes alternative approximations to the distributions of IV estimators under weak instruments, such as allowing the instruments to grow with the sample size (Bekker, 1994), considering the relevance of instruments to be in a neighborhood around zero (Staiger

* Correspondence to: Alfonso Flores-Lagunes, Department of Economics, Eller College of Management, University of Arizona, McClelland Hall 401, Tucson, AZ 85721, USA. E-mail: alfonso@eller.arizona.edu

[1] The fact that the bias is in the direction of the OLS estimator may incorrectly lead one to conclude that there is no endogeneity problem if a Hausman specification test comparing the OLS and IV estimators is employed.

Copyright © 2007 John Wiley & Sons, Ltd.

and Stock, 1997), and higher-order asymptotic expansions (Donald and Newey, 2001; Hahn and Hausman, 2002b). A third aspect provides alternative frameworks for undertaking robust statistical inference on parameters of interest under weak instruments (Wang and Zivot, 1998; Zivot *et al.*, 1998; Kleibergen, 2002; Moreira, 2003). A final aspect aims at finding estimators that have smaller finite sample bias than TSLS in the presence of weak instruments. A large number of these alternative IV estimators are reviewed and evaluated in this paper.[2]

We add to the understanding of the consequences and potential solutions of IV estimation under weak instruments in three main ways. First, we provide finite sample evidence through Monte Carlo simulations regarding point estimation and confidence interval construction when employing several different IV estimators under weak instruments. Second, we explore the usefulness of employing the bootstrap method as a finite-sample bias-reduction technique on some of the estimators considered: while the higher-order expansions typically used to theoretically justify the bootstrap have been found to break down when instruments are arbitrarily weak or irrelevant, it is an open question whether the method is successful in reducing bias of IV estimators in finite samples for small but fixed instrument relevance. Finally, we illustrate the methods using three different empirical applications that illustrate and provide insights into the relative performance of the different estimators in practice. Our focus is on providing the empirical researcher with guidance when only potentially weak instruments are available. The evidence we present indicates that the random-effects quasi-maximum likelihood estimator of Chamberlain and Imbens (2004) outperforms alternative estimators in terms of median point estimates and coverage rates, followed by LIML and the bootstrap bias-corrected version of LIML. However, our results also confirm that reliable point estimates are very difficult to obtain in models with weak identification.

The paper is organized as follows. Section 2 reviews the finite sample bias of the TSLS estimator. Section 3 discusses the alternative IV estimators we consider and the bootstrap bias-reduction technique we apply to some them. Section 4 delineates the design of the Monte Carlo experiment and describes the results. Section 5 presents the results of the three empirical applications we employ. Concluding remarks are provided in the last section of the paper.

2. THE FINITE SAMPLE BIAS OF THE TSLS ESTIMATOR

Consider the following model:

$$Y = X\beta + u \tag{1}$$

$$X = Z\gamma + v \tag{2}$$

where Y is a $T \times 1$ vector of a dependent variable, X is, for simplicity, a $T \times 1$ vector of an explanatory variable, β is the scalar parameter of interest, u is a $T \times 1$ disturbance with mean zero and correlated with X, and v is a $T \times 1$ vector of reduced form disturbances with mean zero.[3] Further, there exists a $T \times n$ nonrandom matrix Z to instrument for X, with $n \geq 1$ and uncorrelated with v and u, and γ is a $n \times 1$ vector of coefficients.

It is well known that, while asymptotically (as $T \to \infty$) the TSLS estimate of β is consistent, in finite samples it is a biased estimator. Rothenberg (1983) provides an expression for the

[2] Stock *et al.* (2002) and Hahn and Hausman (2003) are useful reviews of the literature on weak instruments.

[3] The model can be one in which additional included exogenous variables have been partialled out.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

approximate bias of the TSLS estimator using higher-order asymptotic expansions:[4]

$$\text{bias}_{\text{TSLS}} = \frac{(n-2)\cdot\sigma_{uv}}{T\cdot(\gamma' Z' Z\gamma)^{-1}} \tag{3}$$

Several points are worth noting in equation (3). First, the finite sample bias of TSLS is present even if instrument relevance is acceptable; however, when the instruments are weak ($\gamma \approx 0$) this bias will be exacerbated. Second, the bias is an increasing function of the degree of endogeneity of the endogenous variable (σ_{uv}) and of the number of instrumental variables used (n). Finally, the bias is a decreasing function of the sample size (T).[5] We consider situations in which instrument relevance is very small (but non-zero so that the model is identified), and study the finite-sample behavior of different IV estimators as the factors that play a role in (3) change.

3. ALTERNATIVE IV ESTIMATORS

Several alternative estimators to TSLS have been proposed in an attempt to attenuate its finite sample problems, in particular under weak instruments. We consider in this paper a large number of them, which are outlined in this section in the context of model (1)–(2). The exposition is divided into the κ-class of estimators, those that employ bias-reduction techniques, and other estimators.

3.1. The κ-Class of Estimators

The κ-class of estimators (Nagar, 1959) is given by

$$\hat{\beta}_{\kappa\text{-class}} = (X'P_Z X - \kappa\cdot X'M_Z X)^{-1}(X'P_Z Y - \kappa\cdot X'M_Z Y) \tag{4}$$

where $P_Z \equiv Z(Z'Z)^{-1}Z'$ and $M_Z = (I - P_Z)$. Several estimators belong to this class; for instance, TSLS is obtained when $\kappa = 0$ and limited information maximum likelihood (LIML) when κ is chosen as the smallest eigenvalue of the matrix $W'P_Z W(W'M_Z W)^{-1}$, where $W \equiv [Y:X]$.

The LIML estimator has been generally found to perform better than TSLS in the presence of weak instruments (see, among others, Bekker, 1994; and Blomquist and Dahlberg, 1999). A potential drawback of this estimator, however, is the non-existence of any of its finite sample moments under normally distributed disturbances (see Theorem 12 of Mariano, 2001, and references therein to the large earlier literature), which is known as the 'moment problem' (Hahn *et al*., 2004).

Two other κ-class estimators have been considered for use under the presence of weak instruments: the 'modified LIML' estimator by Fuller (1977), which was expressly developed to have finite sample moments and is obtained when $\kappa = \phi - \frac{\alpha}{T-n}$ where ϕ is the same as κ for LIML, and $\alpha > 0$ is a constant to be chosen by the researcher;[6] and the bias-adjusted 2SLS (B2SLS) (Donald and Newey, 2001; Nagar, 1959), which is mean unbiased but shares the moment problem with LIML, and is obtained when $\kappa = \frac{n-2}{T-n-2}$.

[4] See also Hahn and Hausman (2002a), who obtain a similar approximate finite sample bias expression for TSLS.

[5] We also note that for the case of one endogenous explanatory variable, the expression $\gamma' Z' Z\gamma$ can be written as the product of the R^2 of the first stage regression and $X'X$.

[6] In the simulation experiment we consider $\alpha = 4$, which can be shown to minimize a MSE criterion (Fuller, 1977).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

3.2. Estimators Employing Bias-Reduction Techniques

Since the TSLS estimator is consistent but biased in finite samples, and such bias is exacerbated in the presence of weak instruments, a natural approach is to apply bias-reduction techniques to TSLS. In this paper, we propose to employ the bootstrap method as a bias-reduction technique and explore its practical usefulness. We also apply the bootstrap bias reduction to LIML and Fuller's estimator (denoted as BCLIML and BCFULL).[7] The bias-reduction approach has also been followed by Hahn *et al*. (2004), who apply a jackknife bias correction to TSLS.

Bias reduction consists of computing an estimate of the bias of an estimator and using it to obtain an approximately unbiased estimator. The bias of say, TSLS, is defined as the difference between the expectation of $\hat{\beta}_{\text{TSLS}}$ and the true value of the parameter it estimates:

$$\text{bias}_{\text{TSLS}} = E[\hat{\beta}_{\text{TSLS}}] - \beta \tag{5}$$

To estimate the bias using the bootstrap principle, B samples are drawn (with replacement) from the original data, 'recentered' (see below), and TSLS is computed B times to be used in estimating $E\ [\hat{\beta}_{\text{TSLS}}]$ by $\frac{1}{B}\ \sum_{i=1}^{B} \hat{\beta}_{\text{TSLS}}^{(i)}$. The last term in (5) is approximated by using the TSLS estimator obtained using the original data. In this way, the 'bias-corrected' TSLS estimator is obtained as

$$\hat{\beta}_{\text{BCTSLS}} = \hat{\beta}_{\text{TSLS}} - \widehat{\text{bias}}_{\text{TSLS}} = 2\hat{\beta}_{\text{TSLS}} - \frac{1}{B}\sum_{i=1}^{B} \hat{\beta}_{\text{TSLS}}^{(i)} \tag{6}$$

For consistent estimates that are smooth functions of sample moments in place of population parameters, such as the TSLS estimator, the bias-corrected estimator reduces the order of the bias by the factor T^{-1} under regularity conditions (see Hall, 1992; Shao and Tu, 1995; and references therein).[8]

Freedman (1984) first applied the bootstrap method to TSLS, which implies the following 'recentering' of the TSLS residuals $\hat{u}(T)(= Y - X\hat{\beta}_{\text{TSLS}})$:

$$\tilde{u}(T) = (I - P_Z)\hat{u}(T) \tag{7}$$

This is required since, in general, $\hat{u}(T)$ will not be exactly orthogonal to the set of instrumental variables Z, which is an assumption imposed in the model. The bootstrap method is then used to resample (with replacement) from the empirical distribution of the triplets $(X, Z, \tilde{u}(T))$ in obtaining the B samples to be used in estimating the bias. Freedman (1984) shows using conventional asymptotics ($T \to \infty$) that, under the assumptions that the vector (Y, X, Z, u) is jointly independent

[7] Hahn, Kuersteiner and Newey (2002, unpublished manuscript) analyze the higher-order properties of bootstrap and jackknife bias-corrected maximum likelihood estimators (such as LIML). They find that both methods have the same properties in terms of bias reduction and mean squared error to order $T^{-3/2}$. There are no similar higher-order results for TSLS (or Fuller's estimator), but our simulations suggest that these results may apply to TSLS as well.

[8] There are two potential issues arising in the bootstrap estimation of the bias of TSLS (and other related estimators). One is the use of the expectation in equation (5), since it is known (see, for instance, Theorem 12 in Mariano, 2001) that some IV estimators might not have first or second moments. In the simulation results presented below, we find that using either the mean or the median of the bootstrap distribution of the estimators to bias-correct them makes little difference in practice. The second issue is the possibility that sampling with replacement can result in degenerate samples (e.g., all observations identical), although this can only happen with very low probability. In the simulations and empirical applications below we did not find any such case.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

and identically distributed (iid) and the population random vector (Y, X, Z) has finite fourth moment, the TSLS estimator obtained in each of the bootstrap samples converges in distribution to the same limit as the conventional TSLS estimator.

In the case of weak instruments, however, it is an open question whether the bootstrap method is successful in achieving acceptable bias reduction. The theoretical literature on the subject finds that the higher-order expansions used to justify the bootstrap break down in the unidentified case ($\gamma = 0$) (e.g., Moreira *et al.*, 2004, unpublished manuscript). Since we analyze the situation when γ is small but different from zero, it is valuable to explore how the bootstrap bias correction performs in our simulations as instrument relevance decreases and other features of the estimation environment change.

Hahn *et al.* (2004) propose the use of the jackknife method to estimate the bias of TSLS in (5) and derive a higher order mean-squared error (MSE) for their jackknife 2SLS (J2SLS) estimator, which is shown to be equivalent to the corresponding one for the κ-class estimators derived in Donald and Newey (2001). The jackknife estimate of the bias is given by $(T - 1) \cdot (\hat{\beta}^{(\cdot)}_{\mathrm{TSLS}} - \hat{\beta}_{\mathrm{TSLS}})$, where $\hat{\beta}^{(\cdot)}_{\mathrm{TSLS}} \equiv \frac{1}{T} \sum_{t=1}^{T} \hat{\beta}^{(t)}_{\mathrm{TSLS}}$ and $\hat{\beta}^{(t)}_{\mathrm{TSLS}}$ is the TSLS estimator obtained by removing the tth observation from the data matrices Y, X, and Z. The J2SLS estimator is obtained as: $\hat{\beta}_{\mathrm{J2SLS}} = T\hat{\beta}_{\mathrm{TSLS}} - (T - 1)\hat{\beta}^{(\cdot)}_{\mathrm{TSLS}}$.

As can be seen from the current exposition, the difference between J2SLS and the BCTSLS estimators is the way in which each of them estimates the bias of the TSLS estimator. We discuss briefly some points of comparison between them. First, while the higher-order properties of the BCTSLS estimator have not been derived, we conjecture that they are similar to those of J2SLS.[9] Indeed, our simulation results and empirical applications suggest there is a close agreement in the performance of BCTSLS and J2SLS. A second point of comparison is the ease of computation and flexibility. Both estimators are computationally intensive, since TSLS has to be computed T times for J2SLS and B times for BCTSLS. For a given choice of B, BCTSLS (J2SLS) will be less computationally expensive if $T > B$ ($T < B$). Also, in a given application, T is fixed while B can be chosen, which renders BCTSLS more flexible than J2SLS. In fact, Andrews and Buchinsky (2000) have devised a three-step method for choosing the number of bootstrap repetitions to achieve a given level of accuracy, which can be easily applied in particular applications.[10]

Other estimators based on the jackknife method (called JIVE1 and JIVE2) have been proposed by Phillips and Hale (1977) and more recently by Angrist *et al.* (1999). These estimators have been recently criticized by Davidson and MacKinnon (2006) and improved by Ackerberg and Devereux (2003, unpublished manuscript). They differ from J2SLS in that the jackknife is applied to the first-stage equation to avoid the finite sample bias of IV estimators. In this paper we consider the JIVEs as developed in Angrist *et al.* (1999).

3.3. Other Estimators

There are other estimators that do not naturally fall into the two classes outlined above. Chamberlain and Imbens (2004) propose a random-effects quasi-maximum likelihood estimator

[9] For instance, there is evidence that the higher-order properties of jackknife and bootstrap methods are identical in terms of bias reduction and mean squared error to order $T^{-3/2}$ in the context of maximum likelihood estimators, such as LIML (Hahn *et al.*, 2002, unpublished manuscript).

[10] Using formulae in Andrews and Buchinsky (2000), for a typical simulated sample in the simulation experiment below, the percentage deviation of the estimated bias using $B = 500$ with respect to $B = \infty$ is ±5–6% with 95% probability. This provides support for using $B = 500$ in the simulations below.

Copyright © 2007 John Wiley & Sons, Ltd.

(REQML) for a model with many instrumental variables in the particular case of one endogenous regressor. Their main insight is to put a random coefficients structure (centered at zero) on the statistical relationship between the endogenous regressor and the instrumental variables. The estimator can be obtained from the maximization of a log-likelihood function that depends on the typical parameters of the simultaneous equations model plus the variance parameter of the random coefficients structure of the parameter relating the endogenous variable and instruments.[11] The REQML estimator seems to perform well in situations with many (potentially) weak instrumental variables compared to TSLS and LIML, but it has not been compared to other estimators as we do it here.

Finally, other estimators that are conjectured to perform better than TSLS in situations with weak instruments are generalized empirical likelihood (GEL) estimators. A recent paper by Guggenberger (2005, unpublished manuscript), however, presents finite-sample evidence indicating that this is not the case in the linear model under weak instruments. We do not simulate GEL estimators here, but instead point out below some of Guggenberger's results.[12]

4. MONTE CARLO EXPERIMENT

4.1. Experimental Design

The goal of the Monte Carlo experiment is to compare the relative performance of the estimators presented in the previous section under several model specifications. The data-generating process (DGP) follows model (1)–(2) in which

$$\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{uv} \\ \cdot & 1 \end{bmatrix} \right) \tag{8}$$

β is set to 1, and $Z \overset{\text{iid}}{\sim} N(0, I_n)$. We explore the effects on the performance of each estimator of varying the degree of endogeneity: $\sigma_{uv} \in \{0.5, 0.9\}$; the degree of overidentification: $n \in \{1, 5, 30\}$; and the sample size: $T \in \{100, 500\}$. For the degree of instrument relevance, we fix the value of the R^2 of the first-stage regression (R_f^2) and obtain the implied coefficients for the first-stage regression in (2) following the relationship proposed in Hahn and Hausman (2002b):

$$R_f^2 = \frac{\gamma' E[Z'Z]\gamma}{\gamma' E[Z'Z]\gamma + 1} \tag{9}$$

and then assigning the total explanatory power of the first-stage equation equally among the n corresponding coefficients: $\gamma_j = \sqrt{R_f^2/[n(1 - R_f^2)]}$, $j = 1, \ldots, n$. We consider the following values for R_f^2: $R_f^2 \in \{0.001, 0.01, 0.1\}$. Note that all the factors we consider impact the finite sample bias of TSLS in equation (5). We undertake 5000 replications for the models with $T = 100$ and 1000 for those with $T = 500$.

[11] Chamberlain and Imbens (2004) show that both TSLS and LIML can be obtained as special cases of REQML.

[12] A general computational problem with GEL estimators is to find the solution to the saddle point problem necessary to obtain them. This is even more pronounced under weak instruments, and thus Guggenberger (2005, unpublished) uses a time-consuming fine grid search over the whole parameter space.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Furthermore, to evaluate the possible lenient effect of employing normal variates, we also report results for models with $T = 100$, five instruments generated log-normally distributed, and the disturbances in (8) multivariate-t distributed with either 12 or 1 degrees of freedom.[13] Finally, when applying the bootstrap bias-reduction technique, we employ 500 bootstrap replications, which keeps the experiment manageable and provides an acceptable approximation (see footnote 10).

Some other simulation exercises have compared different subsets of the estimators considered here. Blomquist and Dahlberg (1999) consider LIML and the JIVEs, concluding that none of the analyzed estimators performs uniformly better than TSLS. More recently, Hahn *et al.* (2004) undertake a simulation study considering a number of κ-class estimators, J2SLS, JIVE and TSLS. They find that J2SLS and the modified LIML estimator by Fuller (1977) have better performance than the rest of the estimators. A point raised by Hahn *et al.* (2004) is that estimators such as LIML and JIVEs exhibit the moment problem, evidenced in the form of high RMSE and interquartile range, and poor mean estimates. Finally, Guggenberger (2005, unpublished manuscript) presents finite-sample evidence for generalized empirical likelihood (GEL) estimators under weak instrument relevance and compares them to TSLS, LIML, and Fuller's (1977) estimator, finding that the latter κ-class estimators perform typically better than the GEL estimators.[14] Given the similarity of Guggenberger's experimental design to some of our model specifications, his results pertaining to the performance of GEL estimators can be compared to the performance of the alternative estimators presented here.[15]

4.2. Measures to Evaluate the Performance of the Estimators

One important issue concerns the choice of measures to compare the different estimators. Angrist *et al.* (1999), for instance, point out that median point estimates and MAE might be more useful measures than mean point estimates and RMSE, since many of the IV estimators do not necessarily have first or second finite sample moments. On the other hand, Hahn *et al.* (2004) argue that the absence of finite sample moments is of concern, which is reflected in higher-order risk properties for those estimators. In fact, it is common to obtain extremely high mean estimates and/or RMSE for certain estimators in simulations, which most probably results from the absence of finite sample moments. Under this latter view, measures such as mean estimates, RMSE, and interdecile range (IDR) may become important in evaluating the finite sample performance of the estimators. To save space, we discuss only the results on the median point estimates, MAE, IDR, and the coverage rates of the 95% confidence interval yielded by the estimators.[16]

4.3. Results

The results from the Monte Carlo experiment are contained in four tables that correspond to each one of the measures of performance considered and in figures plotting the empirical density of

[13] Both sets of variates are standardized to have mean zero and variance one. The other features of the DGP take on values as before.

[14] The GEL estimators considered are the continuous updating, empirical likelihood and exponential tilting estimators. See Guggenberger (2005, unpublished manuscript) for details.

[15] Guggenberger (2005, unpublished manuscript) considers $R_f^2 \in \{0.002, 0.02, 0.2\}$, $\sigma_{uv} \in \{0, 0.5\}$, $n \in \{1, 3, 5, 20\}$, and $T = 100$, with normally distributed variates.

[16] The results pertaining to the mean point estimates and RMSE are available upon request. We do point out, however, that the REQML estimator, which is found to perform well in the results presented here, typically has high mean point estimates and RMSE.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

selected estimators. We summarize below the simulation results based on our tables and figures; however, in order to save space, we only present here four figures to illustrate the main points of our findings and make available the tables and the rest of the figures on the JAE Data Archive at www.econ.queensu.ca/jae.

Median Point Estimates

In terms of median point estimates, REQML outperforms all other estimators across model specifications, followed by LIML and BCLIML. Importantly, REQML performs substantially better than the other estimators when the model has the weakest instrument relevance and highest degree of endogeneity. Nevertheless, the remaining bias is sometimes quite considerable. This point illustrates the difficulty associated with obtaining reliable point estimates when instrument relevance is considerably poor. It is also interesting to note that the bootstrap bias reduction is successful on the estimators, except in the models with weakest instrument relevance; and that TSLS is typically outperformed by most of the other estimators in terms of median point estimates.

Median Absolute Error (MAE)

In this respect, FULLER and BCFULL outperform the other estimators, while REQML has minimum MAE for some of the model specifications where instrument relevance is weakest and the degree of endogeneity highest. At the other end, the estimators with highest MAE across model specifications are both JIVEs and to a lesser extent BCLIML.

Interdecile Range (IDR)[17]

FULLER has the smallest IDR across model specifications, except when the models contain 30 instruments, where TSLS has smallest IDR. However, the empirical densities of these two estimators (to be discussed below) are concentrated away from the true parameter value. The estimators with relatively high IDR are JIVE2, JIVE1, REQML, J2SLS and BCLIML. In particular, REQML typically has highest IDR in the model specifications with weakest instrument relevance and highest degree of endogeneity.

Coverage Rates[18]

The estimator that performs best in this regard is the REQML estimator. In particular, it is worth noting that, even though REQML does not have best coverage rate in all model specifications, it is the only estimator that yields coverage rates that are consistently closer to the nominal 95% across all model specifications. In contrast, other estimators that perform well in some model specifications (e.g., JIVE1, BCLIML, JIVE2 and BCTSLS) show considerable under-coverage rates in other model specifications, sometimes on the order of 0.5 or worse. Finally, it is interesting to note that the confidence intervals for the bootstrap bias-corrected estimators that use the bootstrap estimate of the standard error typically result in better coverage rates compared to the uncorrected estimators, especially for BCLIML, which has best coverage rate in many model specifications.

[17] The IDR is the width of the interval defined by the difference between the ninth and first deciles of the values of the estimator obtained in the replications.

[18] The confidence intervals are computed using the asymptotic distribution of the estimators, that is, the estimate plus or minus 1.96 times the asymptotic standard error. For the bootstrap bias-corrected estimators (BCTSLS, BCLIML, and BCFULL), the confidence intervals are constructed using the bootstrap estimate of the standard error, which results in better coverage rates. For the REQML confidence interval, we follow Chamberlain and Imbens (2004) and find upper and lower values such that the concentrated log-likelihood function differs from its maximum value by $G^{-1}(0.95)/2$, where G is a chi-squared distribution with one degree of freedom.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Performance Across DGP Features

As expected, all measures of performance for the estimators improve as the degree of instrument relevance increases. In fact, in most of the cases with $R_f^2 = 0.1$ the majority of the estimators are median unbiased and have acceptable coverage rates, with the exception of TSLS, FULLER and BCFULL when $n = 30$.

As the degree of overidentification increases, the performance of the estimators typically deteriorates, especially of those closely related to TSLS. This is particularly true for median point estimates and coverage rates; and in model specifications with high degree of endogeneity and values of R_f^2 of 0.01 or 0.1. Nevertheless, the following estimators are more robust to a high degree of overidentification: LIML, FULLER, BCLIML, BCFULL, and REQML. This evidence is consistent with the recent theoretical results of Hansen *et al*. (2005, unpublished manuscript), who argue that using a large number of instruments improves efficiency but results in inaccurate statistical inference. They suggest the use of FULLER or LIML and propose standard errors based on Bekker's (1994) asymptotics to improve inference in cases of high degree of overidentification.[19]

The performance (except in terms of IDR) of most of the estimators deteriorates as the degree of endogeneity increases across model specifications, although it typically does not make much of a difference in the specifications with highest instrument relevance. When the sample size increases from 100 to 500 there is a marked improvement in the performance of the estimators, but this trend does not totally hold for specifications with the weakest instrument relevance, in which clearly the sample size does not completely offset the weak relevance.

Finally, the models that depart from normality yield further insights. When the disturbances are generated multivariate t with 12 degrees of freedom the results outlined above hold: REQML and BCLIML dominate the other estimators in terms of median point estimates, FULLER dominates in terms of MAE and IDR, and REQML yields good coverage rates across specifications. As expected, however, when the disturbances depart further from normality (multivariate t with 1 degree of freedom) all measures of performance for all estimators deteriorate substantially. Importantly, REQML disproportionately deteriorates under these models, suggesting its good performance may not hold under strong departures from normality.

Lessons from the Empirical Distribution of the Estimators

Figures 1–4 plot the density of selected estimators in four different model specifications, which are illustrative of the full set of plots for all model specifications.[20] Figure 1 corresponds to a model with the following specifications: $R_f^2 = 0.001$, $\sigma_{uv} = 0.9$, $n = 30$, $T = 100$; which is one of the most difficult estimation scenarios considered. Consequently, the density of all the estimators is centered away and to the right of the true value ($\beta_0 = 1$), which illustrates the difficulty of obtaining reliable point estimates and inference under these circumstances. However, a few estimators clearly perform better in relative terms: REQML is the estimator that puts highest mass on the true value of the parameter, followed by LIML and BCLIML. All other estimators put essentially zero mass on the true value. Figure 2 corresponds to a model similar to that in Figure 1 except that instrument relevance is now higher at $R_f^2 = 0.01$. Despite this higher instrument relevance, the performance

[19] We thank an anonymous referee for suggesting this point.

[20] The density estimates are based on a Gaussian kernel using Silverman's rule of thumb bandwidth.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

of the estimators is very similar. This illustrates the importance of other factors besides instrument relevance, such as the high degree of overidentification and the degree of endogeneity, which affect the properties of IV estimators.

Figure 3 corresponds to a model with the same specification as before but now highest instrument

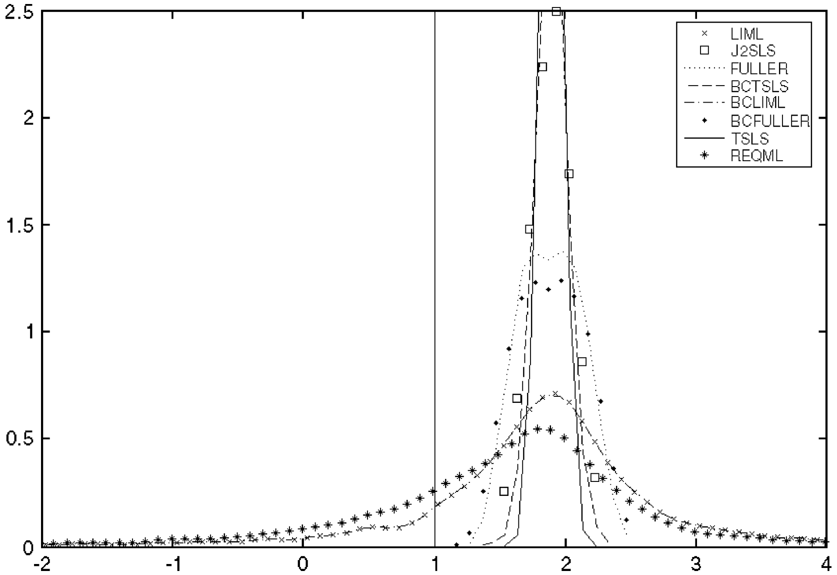

Figure 1. Empirical density of selected estimators for model with $R_f{}^2 = 0.001$, $\sigma_{uv} = 0.9$, $n = 30$, $T = 100$

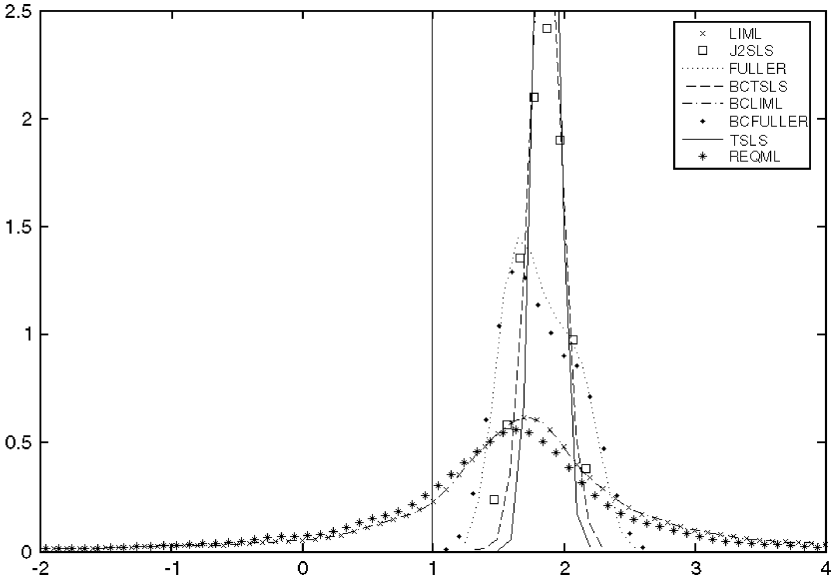

Figure 2. Empirical density of selected estimators for model with $R_f{}^2 = 0.01$, $\sigma_{uv} = 0.9$, $n = 30$, $T = 100$

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

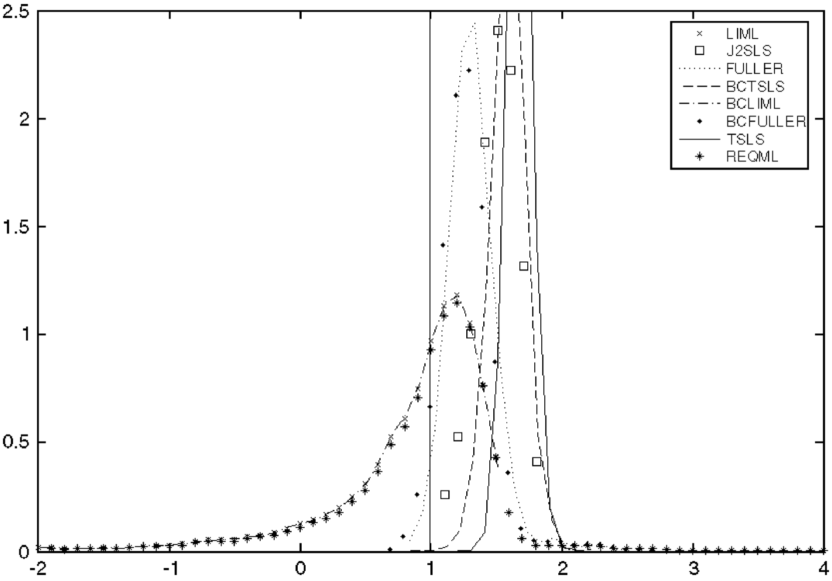

Figure 3. Empirical density of selected estimators for model with ${R_f}^2 = 0.1$, $\sigma_{uv} = 0.9$, $n = 30$, $T = 100$

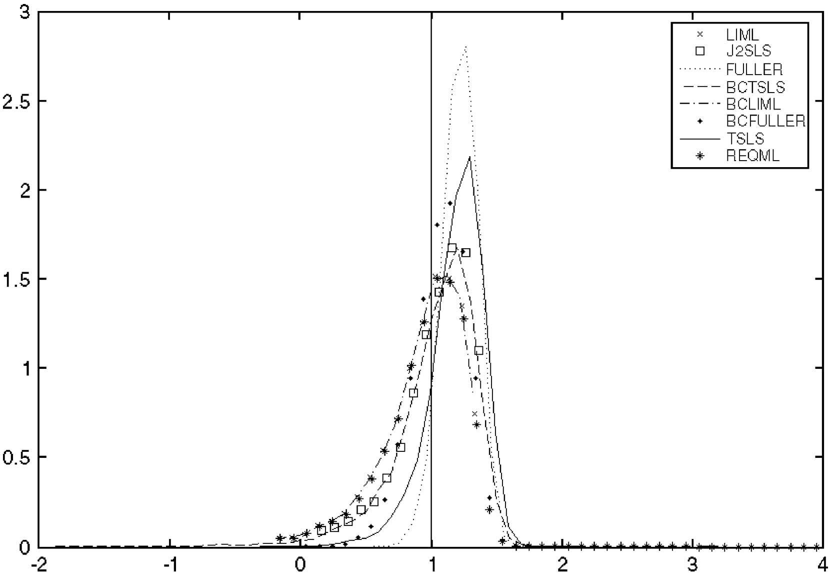

Figure 4. Empirical density of selected estimators for model with ${R_f}^2 = 0.1$, $\sigma_{uv} = 0.9$, $n = 5$, $T = 100$

relevance ($R_f^2 = 0.1$), while Figure 4 is based on a model as Figure 3 but now with $n = 5$ (smaller degree of overidentification). A number of points are worth mentioning about these two figures. First, in Figure 3 the densities of REQML, LIML and BCLIML are very similar and almost centered at the true value. Second, in both figures the rest of the estimators improve with respect

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

to the previous two models in Figures 1 and 2, but are still centered far from the true value. Third, FULLER and BCFULL perform relatively better in Figure 3 than the rest of the estimators based on TSLS, which is consistent with the results in Hansen *et al.* (2005, unpublished manuscript) since this advantage is not present in Figure 4, where the degree of overidentification is much lower. Fourth, the bootstrap bias correction for TSLS and FULLER recenters the density of the estimators towards the true value, but not nearly enough. This property of the bootstrap bias correction is present in most of the model specifications, and we point out the amount of correction decreases as the instrument relevance decreases, which is similar to the findings in Moreira *et al.*'s (2004, unpublished manuscript) simulations for bootstrap size adjustments of t-statistics under weak instruments.

Summary

The results from the simulation experiment illustrate the properties of alternative IV estimators under several model specifications. First of all, our results make evident the difficulty of obtaining reliable point estimates when the model is weakly identified, in particular if that is coupled with high degree of overidentification and/or endogeneity. Nevertheless, our results point towards REQML as an estimator with less biased median point estimates, satisfactory MAE and IDR, and acceptable coverage rates across model specifications, relative to alternative IV estimators. Other estimators worth considering are the LIML and BCLIML with bootstrapped standard errors in the construction of confidence intervals.

5. EMPIRICAL APPLICATIONS

5.1. Returns to Schooling

In an influential paper, Angrist and Krueger (1991) estimate the returns to schooling using quarter of birth as an instrument for schooling in a log-wage equation. This study has become a benchmark for testing methodologies concerning IV estimation in the presence of weak instrumental variables. The data consists of 329,509 men born in the 1930s, taken from the 1980 census. Among the various specifications estimated by Angrist and Krueger (1991), we will concentrate on two of them. The first uses 30 instruments: 3 quarter of birth dummies and 27 dummies resulting from interacting quarter of birth with 9 year of birth dummies. The second specification uses 180 instruments: the same instruments as in the 30 instrument case plus 150 dummies obtained by interacting quarter of birth with 50 state of birth dummy variables.[21] Estimates of the returns to schooling coefficient using OLS, TSLS, LIML, JIVE1, JIVE2 and B2SLS are reproduced from Angrist *et al.* (1999) and Donald and Newey (2001). We further estimate both specifications using FULLER, BCTSLS, BCLIML, BCFULL and J2SLS, while the REQML estimate is taken from Chamberlain and Imbens (2004). The estimates are presented in Table I.[22]

On both specifications, the OLS estimator is notably smaller in magnitude than all IV estimators, while TSLS is smallest in magnitude among all IV estimators and closest to the OLS estimator,

[21] The first secification includes 9 year of birth dummy variables as included exogenous variables, while the second also includes 50 state of birth dummy variables.

[22] Chamberlain and Imbens (2004) employ a different specification of the AK data which includes 504 instruments and only individuals born in either the first or fourth quarters for a $T = 162, 515$. They find, as we do, that TSLS, LIML and REQML yield very similar point estimates and confidence intervals.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Table I. Comparison of IV estimators in three empirical applications

	Returns to schooling		Elasticity of intertemporal substitution		Labor supply functions		Computation time on labor supply functions application
	30 instruments	180 instruments	$1/\psi$	ψ	Wage (β)	Non-labor income (γ)	
OLS	0.071	0.067	0.429	0.161	−0.019	0.113	0.03
	[0.070, 0.072]	[0.066, 0.068]	[0.213, 0.645]	[0.080, 0.241]	[−0.023, −0.014]	[0.035, 0.192]	
TSLS	0.089	0.093	0.683	0.059	−0.018	0.353	0.07
	[0.058, 0.120]	[0.075, 0.111]	[−0.250, 1.616]	[−0.110, 0.228]	[−0.035, −0.0005]	[0.154, 0.552]	
LIML	0.093	0.106	34.11	0.029	−0.023	0.427	0.3
	[0.058, 0.128]	[0.082, 0.130]	[−185.8, 254]	[−0.159, 0.217]	[−0.047, 0.001]	[0.164, 0.691]	
B2SLS	0.093	0.12	0.739	0.055	−0.028	0.484	0.09
	[0.058, 0.128]	[0.093, 0.147]	[−0.29, 1.768]	[−0.117, 0.227]	[−0.058, 0.002]	[0.170, 0.800]	
JIVE1	0.096	0.121	0.647	0.035	−0.003	0.230	0.95
	[0.053, 0.139]	[0.082, 0.160]	[−1.168, 2.462]	[−0.159, 0.229]	[−0.02, 0.015]	[0.030, 0.40]	
JIVE2	0.096	0.121	0.651	0.036	−0.005	0.270	0.95
	[0.053, 0.139]	[0.082, 0.160]	[−1.058, 2.360]	[−0.156, 0.228]	[−0.022, 0.013]	[0.070, 0.50]	
FULLER	0.092	0.106	1.189	0.041	−0.021	0.402	0.3
	[0.059, 0.125]	[0.086, 0.130]	[−0.542, 2.920]	[−0.139, 0.221]	[−0.043, 0.0009]	[0.161, 0.643]	
J2SLS	0.101	—	0.711	0.051	−0.023	0.450	47.45
	[0.070, 0.132]	—	[−0.226, 1.648]	[−0.118, 0.220]	[−0.041, −0.006]	[0.245, 0.655]	
BCTSLS	0.094	0.101	0.706	0.049	−0.019	0.403	31
	[0.063, 0.125]	[0.083, 0.119]	[−0.275, 1.687]	[−0.151, 0.235]	[−0.039, 0.0002]	[0.198, 0.607]	
BCLIML	0.093	0.106	50.53	−0.054	−0.054	0.505	147.52
	[0.059, 0.127]	[0.081, 0.130]	[15.18, 85.8]	[−0.278, 0.152]	[−0.146, 0.038]	[−0.437, 1.45]	
BCFULL	0.093	0.106	1.374	−0.034	−0.047	0.434	131.19
	[0.060, 0.126]	[0.085, 0.130]	[0.516, 2.231]	[−0.235, 0.156]	[−0.077, −0.017]	[0.132, 0.736]	
REQML	0.096		27.39	0.071	—	—	0.49
	[0.056, 0.139]	—	[2.385, ∞]	[−0.829, 0.771]	—	—	
(R_f^2, n, T)	(0.00043, 30, 329 509)	(0.0013, 180, 329 509)	(0.055, 4, 206)	(0.236, 4, 206)	(0.135, 29, 602)	(0.279, 29, 602)	

Notes:
(1) 95% asymptotic confidence intervals in brackets. They are obtained by using the standard error of the estimator (the bootstrap estimate of the standard error for the bootstrap bias-corrected estimators) and the asymptotically normal approximation.
(2) 500 replications are used for the bootstrap bias-corrected estimators.
(3) J2SLS is not estimated for the 180-instrument specification of the returns to schooling application due to computational burden. For the same reason the REQML estimate of the returns to schooling corresponds to a different specification reported in Chamberlain and Imbens (2004), see footnote 22 in the text.
(4) The confidence intervals for REQML are computed as in Chamberlain and Imbens (2004): finding upper and lower values such that the concentrated log-likelihood function differs from its maximum value by $G^{-1}(0.95)/2$ where G is a chi-squared distribution with one degree of freedom.
(5) The coefficient of non-labor income is scaled by 100.
(6) Computation times are in seconds on a Pentium M processor at 1.6 GHz with 496 MB RAM. For REQML, the computation time assumed only one endogenous variable (wage).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

which is consistent with the direction of TSLS's finite sample bias (under the assumption of orthogonal instruments) and with the sensitivity of TSLS to the degree of overidentification, since such difference is larger in the 180-instrument case. Interestingly, most other IV estimators are very similar in magnitude (within one standard deviation of each other), and also the confidence intervals of the estimators are all of roughly the same length. In summary, it seems that in this empirical application there are no substantial finite-sample bias for the IV estimators (except perhaps TSLS), which may be due to the large sample sizes offsetting the effects of the low R_f^2.

5.2. Elasticity of Intertemporal Substitution

In a recent paper, Yogo (2004) analyzes the problem of estimating the elasticity of intertemporal substitution (EIS) using the linearized Euler equation. In this literature, weak instruments have been blamed for the empirical puzzle that using conventional IV methods the estimated EIS is statistically significantly less than 1, while its reciprocal, obtained using a 'reverse regression', is not statistically different from 1. Yogo (2004) illustrates how undertaking inference with methods robust to weak instruments helps solve this empirical puzzle for 11 countries used in his analysis. Besides employing tests for weak instruments and robust methods for undertaking statistical inference on the parameter of interest, he presents LIML and FULLER as alternative estimators to TSLS.

In this subsection, we follow one of the specifications in Yogo (2004) using quarterly data from 1947.3 to 1998.4 for the United States and compare all the estimators considered in the present paper. The estimated models correspond to the following equations:[23]

$$\Delta c_{t+1} = \tau + \psi r_{f,t+1} + \xi_{t+1} \tag{10}$$

and the 'reverse regression':

$$r_{f,t+1} = \mu + (1/\psi)\Delta c_{t+1} + \eta_{t+1} \tag{11}$$

where ψ is the EIS, Δc_{t+1} is consumption growth at time $t+1$, $r_{f,t+1}$ is the real return on a risk-free asset, τ and μ are constants, and ξ_{t+1} and η_{t+1} are the innovations to consumption growth and asset return, respectively. Since the explanatory variable in each of the equations is correlated with the corresponding innovation, the following instruments are used when employing IV methods: the twice-lagged nominal interest rate, inflation, consumption growth, and log dividend–price ratio.

Yogo (2004) documents, using the test for weak instruments in Stock and Yogo (2004), that these are weak instruments for Δc_{t+1} but apparently not for $r_{f,t+1}$. This appears to be the case given the parameter estimates for each equation presented in columns 3 and 4 of Table I: while all the IV estimates for ψ are similar in magnitude (within one standard deviation from each other), those for $1/\psi$ are very different, especially LIML, BCLIML, and REQML. In particular, it is interesting to note the behavior of their respective confidence intervals: for most estimators, they are misleadingly tight, except again for LIML, BCLIML and REQML, while notably the confidence intervals for REQML and BCLIML do not include 1, which is consistent with the

[23] See Yogo (2004) and references therein for the derivation of these two equations from the linearized Euler equation, the importance that the EIS has in macroeconomics and finance, and a description of the data used.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

conclusion in Yogo (2004) that weak instruments explain the empirical puzzle in the estimation of EIS.

5.3. Labor Supply Functions

Blomquist and Dahlberg (1999) examine the performance of some IV methods in estimating linearized labor supply functions when the budget constraints are nonlinear, following Blomquist (1996). Hours of work are generated with actual data on nonlinear budget constraints of 602 Swedish males.[24] The estimated equation is

$$h_i = \alpha + \beta w_i + \gamma y_i + \theta_1 A_i + \theta_2 NC_i + \varepsilon_i$$

where h_i are hours worked, w_i is the wage, y_i is non-labor income, A_i is age and NC_i is the number of children. The corresponding literature has recognized that both w_i and y_i are endogenous. Commonly used instruments in this framework are socio-demographic variables, which appear to be weak instruments (Blomquist and Dahlberg, 1999): dummies for the educational level of the individual, his wife, father and mother, and dummies for the type of region where the individual lives. In total, 29 instruments are used.

The results for this application are presented in columns 5 (for β) and 6 (for γ) of Table I.[25] The results again show some of the patterns evident in the simulations and the previous applications, which we summarize here. First, TSLS is typically closer to OLS (which is the direction of the finite-sample bias) than most of the other estimators. Second, among the bootstrap bias-corrected estimators, BCTSLS is very close to TSLS, while BCLIML and BCFULL are further away from OLS than TSLS (which holds across the three empirical applications). Finally, the width of the confidence interval of TSLS is typically smaller than those of the other estimators, in particular LIML, BCLIML and BCFULL.

The last column in Table I shows computation times for the alternative IV estimators in this application.[26] For most IV estimators the computation time is not an issue, except for those employing bias-reduction techniques, for which the computation time is close to that of their corresponding uncorrected counterpart (TSLS, LIML or FULLER) multiplied by the number of bootstrap replications (B) or the sample size (T) in the case of the jackknife. The computation time for REQML in this application is about 60% higher than LIML.[27]

5.4. Summary of Applications

Consistent with the simulation results, when appropriate relevance of the instruments is doubtful, some of the alternative IV estimators (in particular REQML, LIML and BCLIML) yield substantially different point estimates and wider confidence intervals than the other estimators.

[24] The data and method to construct hours of work are described in Blomquist and Dahlberg (1999) and more thoroughly in Blomquist (1996).

[25] The REQML estimator was developed for the case of one endogenous variable, and thus cannot be computed for this application, which contains two endogenous regressors.

[26] Computation times are reported in seconds on a Pentium M processor at 1.6 GHz with 496 MB RAM. For REQML we assume only one endogenous variable (wage) since the method does not allow more than one endogenous variable.

[27] We also note that, in our experience, a computational burden in REQML arises when obtaining the confidence intervals in the way proposed by Chamberlain and Imbens (2004) (see footnote 18), especially when the instruments are very weak.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Conversely, when the relevance of the instruments appears to be appropriate (especially relative to the sample size), alternative IV point estimates are within one standard deviation of each other and their confidence intervals are similar.

6. CONCLUSIONS

This paper presents finite sample evidence (using simulations and three empirical applications) of a number of IV estimators to estimate linear models, with special emphasis on weak identification; and explores the application of the bootstrap bias-reduction technique to some of these IV estimators.

The finite sample evidence we present indicates that the random-effects quasi-maximum likelihood (REQML) estimator of Chamberlain and Imbens (2004) outperforms alternative estimators in terms of median point estimates and coverage rates, followed by LIML and the bootstrap bias-corrected version of LIML. However, our results also confirm that reliable point estimates are very difficult to obtain in models with weak identification, especially when that is coupled with large overidentification and high degree of endogeneity.

In terms of the application of the bootstrap bias reduction, the correction is successful for TSLS and FULLER in recentering the density of the estimators towards the true value. However, this correction is not typically enough to obtain estimators with better overall finite sample properties, and, importantly, the extent of the correction decreases as the instrument relevance decreases. For LIML, the bootstrap bias correction does not make much of a difference; however, using bootstrapped standard errors in the construction of confidence intervals improves the coverage rates of BCLIML or LIML.

In addition, we present three different empirical applications that illustrate the relative performance of the different methods in practice. Consistent with the simulation results, when appropriate relevance of the instruments is doubtful, some of the alternative IV estimators (in particular REQML, LIML and BCLIML) yield substantially different point estimates and wider confidence intervals than the other estimators. Conversely, when the relevance of the instruments appears to be appropriate, alternative IV point estimates are within one standard deviation of each other and their confidence intervals are similar.

ACKNOWLEDGEMENTS

Detailed comments and suggestions by the referees and one co-editor are gratefully acknowledged. Comments by Stephen R. Cosslett, Carlos A. Flores, John Ham, Kei Hirano, Stefan Krause, Audrey Light, Lawrence Marsh, Nathan Porter, James Stock, Bruce Weinberg, and participants at the 2001 Econometric Society Winter Meetings are also gratefully acknowledged. Thanks also to Matz Dahlberg, Alan Krueger, and Motohiro Yogo for making their data available. Very competent research assistance was provided by Alex Shcherbakov. Remaining errors are my own.

REFERENCES

Andrews DWK, Buchinsky M. 2000. A three-step method for choosing the number of bootstrap repetitions. *Econometrica* **68**: 23–51.

Copyright © 2007 John Wiley & Sons, Ltd.
DOI: 10.1002/jae

Angrist JD, Krueger AB. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* **106**: 979–1014.

Angrist JD, Imbens G, Krueger AB. 1999. Jackknife instrumental variables estimation. *Journal of Applied Econometrics* **14**: 57–67.

Bekker PA. 1994. Alternative approximations to the distributions of instrumental variables estimators. *Econometrica* **62**: 657–682.

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

Bound J, Jaeger DA, Baker RM. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* **90**: 443–450.

Chamberlain G, Imbens G. 2004. Random effects estimators with many instrumental variables. *Econometrica* **72**: 295–306.

Davidson R, MacKinnon JG. 2006. The case against JIVE (with discussion). *Journal of Applied Econometrics* **21**: 827–833.

Donald SG, Newey W. 2001. Choosing the number of instruments. *Econometrica* **69**: 1161–1191.

Freedman DA. 1984. On bootstrapping two-stage least-squares estimates in stationary linear models. *Annals of Statistics* **12**: 827–842.

Fuller WA. 1977. Some properties of a modification of the limited information estimator. *Econometrica* **45**: 939–954.

Hahn J, Hausman J. 2002a. Notes on bias in estimators for simultaneous equations models. *Economics Letters* **75**: 237–241.

Hahn J, Hausman J. 2002b. A new specification test for the validity of instrumental variables. *Econometrica* **70**: 163–189.

Hahn J, Hausman J. 2003. Weak instruments: diagnosis and cures in empirical econometrics. *American Economic Review Papers and Proceedings* **93**: 118–125.

Hahn J, Hausman J, Kuersteiner G. 2004. Estimation with weak instruments: accuracy of higher order bias and MSE approximations. *Econometrics Journal* **7**: 272–306.

Hall P. 1992. *The Bootstrap and Edgeworth Expansion*. Springer: New York.

Hall A, Rudebusch GD, Wilcox DW. 1996. Judging instrument relevance in instrumental variables estimation. *International Economic Review* **37**: 283–298.

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

Maddala GS, Jeong J. 1992. On the exact small sample distribution of the instrumental variable estimator. *Econometrica* **60**: 181–183.

Mariano RS. 2001. Simultaneous equation model estimators: statistical properties and practical implications. In *A Companion to Theoretical Econometrics*, Baltagi BH (ed.). Blackwell: Malden, MA.

Moreira M. 2003. A conditional likelihood ratio test for structural models. *Econometrica* **71**: 1027–1048.

Nagar AL. 1959. The bias and moment matrix of the general *k*-class estimators of the parameters in simultaneous equations. *Econometrica* **27**: 575–595.

Nelson CR, Startz R. 1990a. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* **58**: 967–976.

Nelson CR, Startz R. 1990b. The Distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business* **63**: 125–140.

Phillips GDA, Hale C. 1977. The bias of instrumental variable estimators of simultaneous equation systems. *International Economic Review* **18**: 219–228.

Rothenberg TJ. 1983. Asymptotic properties of some estimators in structural models. In *Studies in Econometrics Time Series, and Multivariate Statistics*, Karlin S, Amemiya T, Goodman LA (eds). Academic Press: New York.

Shao J, Tu D. 1995. *The Jackknife and Bootstrap*. Springer: New York.

Shea J. 1997. Instrument Relevance in Multivariate Linear Models: A Simple Measure. *Review of Economics and Statistics* **79**: 348–352.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae

Staiger D, Stock JH. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica* **65**: 557–586.

Stock JH, Wright JH, Yogo M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* **20**: 518–529.

Stock JH, Yogo M. 2004. Testing for weak instruments in linear IV regression. In *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, Andrews DWK, Stock JH (eds). Cambridge University Press: Cambridge, UK.

Wang J, Zivot E. 1998. Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica* **66**: 1389–1404.

Yogo M. 2004. Estimating the elasticity of intertemporal substitution when instruments are weak. *Review of Economics and Statistics* **86**: 797–810.

Zivot E, Startz R, Nelson CR. 1998. Valid confidence intervals and inference in the presence of weak instruments. *International Economic Review* **39**: 1119–1144.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/jae