TECHNICAL WORKING PAPER SERIES

JACKKNIFE INSTRUMENTAL VARIABLES ESTIMATION

Joshua D. Angrist
Guido W. Imbens
Alan Krueger

Technical Working Paper No. 172

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 1995

We thank seminar participants at the Harvard-MIT Econometrics workshop, the Duke-UNC-SAS Econometrics workshop, Tel Aviv University, and Hebrew University for helpful comments. This paper is part of NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #172
February 1995

JACKKNIFE INSTRUMENTAL VARIABLES ESTIMATION

ABSTRACT

Two-stage-least-squares (2SLS) estimates are biased towards OLS estimates. This bias grows with the degree of over-identification and can generate highly misleading results. In this paper we propose two simple alternatives to 2SLS and limited-information-maximum-likelihood (LIML) estimators for models with more instruments than endogenous regressors. These estimators can be interpreted as instrumental variables procedures using an instrument that is independent of disturbances even in finite samples. Independence is achieved by using a "leave-one-out" jackknife-type fitted value in place of the usual first-stage equation. The new estimators are first-order equivalent to 2SLS but with finite-sample properties superior to those of 2SLS and similar to LIML when there are many instruments. Moreover, the jackknife estimators appear to be less sensitive than LIML to deviations from the linear reduced form used in classical simultaneous equations models.

Joshua D. Angrist
Department of Economics
MIT
E52-380A
Cambridge, MA 02139
and NBER

Guido W. Imbens
Department of Economics
Harvard University
Littauer 117
Cambridge, MA 02138
and NBER

Alan Krueger
Office of Chief Economist
Room S2018
US Department of Labor
200 Constitution Avenue, NW
Washington, DC 20210
and NBER

1. INTRODUCTION

This paper develops two simple alternatives to two-stage-least-squares (2SLS) and limited-information-maximum-likelihood (LIML) estimators for models with more instruments than endogenous regressors. The new estimators can be interpreted as instrumental variables estimators based on an asymptotically optimal instrument constructed in a manner that ensures that even in finite samples it is independent of the disturbance in the regression equation. One way to achieve this is by doing the first stage regression N times, once for each observation, leaving out one observation at a time. While this may seem cumbersome and computationally expensive, this estimator can be written in a way that requires only two passes through the data. The resulting computation is of the order of weighted least squares. The second version removes the dependence on the value of the endogenous regressor on the estimated instrument in a similar manner. Both estimators are simple to implement in standard packages and are first-order equivalent to 2SLS and LIML.

The finite sample properties of these estimators, which we refer to jointly as jackknife instrumental variables estimators (JIVE1 and JIVE2), are superior to those of 2SLS and similar to those of LIML in the case of many instruments which are only weakly correlated with the endogenous regressor. This case has received considerable attention in the recent literature on instrumental variables estimation (Bekker, 1994; Staiger and Stock, 1994; Angrist and Krueger, 1995; Bound, Jaeger and Baker, 1995) in reaction to applications (Angrist, 1990; Angrist and Krueger, 1991). Unlike 2SLS, JIVE1 and JIVE2 are centered around the true parameter value, even with many weak instruments. The JIVE estimators are closely related in spirit to the two sample instrumental variables (TSIV) and split sample instrumental variables (SSIV) estimators developed by Angrist and Krueger (1992, 1995). But they differ importantly from the SSIV estimators in two respects. First, they are asymptotically efficient. Second, they do not require an arbitrary sample split.

Like 2SLS, JIVE1 and JIVE2 can be interpreted as instrumental variables estimators with a constructed instrument of the same dimension as the endogenous regressor. For

2SLS, as well as for JIVE1 and JIVE2, this constructed instrument converges to the best linear predictor of the endogenous regressor given the instruments. The probability limit of the new estimators is therefore identical to that of 2SLS even under general misspecification. This is important because if the model is mis-specified, LIML and 2SLS can have very different properties. While neither dominates the other, Fisher (1966, 1967) suggests that 2SLS (and therefore JIVE1 and JIVE2) may have the edge in more cases.

Section 2 discusses the bias of 2SLS and Section 3 introduces the JIVE estimators. A Nagar-type approximation argument (Nagar 1959, Buse 1992) is used in Section 4 to explain why the JIVE estimators are likely to perform better than 2SLS when there are many instruments. In Section 5, we use a Bekker (1994) group–asymptotic parameter sequence to further characterize and compare the finite–sample properties of JIVE, 2SLS, and LIML. 2SLS is not consistent under this sequence while LIML and JIVE are. The group–asymptotic approach also provides some insight as to when JIVE might perform better than LIML. Section 6 presents a small Monte Carlo study.[1] Here we report coverage rates for confidence intervals constructed using conventional asymptotic approximations, as well as quantiles and median bias from the Monte Carlo sampling distribution. Section 7 applies the estimators to the Angrist and Krueger (1991) data and Section 8 concludes.

2. The Bias of Two-Stage-Least-Squares

The model we are interested in is

$$Y_i = X_i\beta + \varepsilon_i$$

$$X_i = Z_i\pi + \eta_i.$$

The random variable Y_i is a scalar, X_i is an L dimensional row vector and the instrument Z_i is a K dimensional row vector, with $K \geq L$. The number of overidentifying restrictions is $K - L$. In matrix notation we can write this model as

[1] In a comment on an earlier version of this paper, Blomquist and Dahlberg (1994) found (in a Monte Carlo study) that JIVE is typically the minimum mean squared error estimator in the group of approximately unbiased estimators they consider.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{1}$$

$$\mathbf{X} = \mathbf{Z}\pi + \eta. \tag{2}$$

In the matrix formulation, $\mathbf{Y}$ is an N vector with typical element Y_i, $\mathbf{X}$ is an $N \times L$ dimensional matrix with typical row X_i, $\mathbf{Z}$ is an $N \times K$ dimensional matrix with ith row equal to Z_i, ϵ is an N vector and η is a matrix of dimension $N \times L$ with typical rows ϵ_i and η_i. If there are M common elements in the vector of regressors and the vector of instruments, then M columns of the $N \times L$ matrix η are identically zero.

We assume that conditional on Z_i the disturbance ϵ_i has expectation zero and variance σ^2. We also assume that $\mathrm{E}[\eta|\mathbf{Z}] = 0$ and $\mathrm{E}[\eta_i'\eta_i] = \Sigma_\eta$, with rank $L - M$. Finally, $\mathrm{E}[\epsilon_i\eta_i']$ is equal to the L dimensional column vector $\sigma_{\epsilon\eta}$ and the probability limits of $\mathbf{Z}'\mathbf{Z}/N$ and $\mathbf{X}'\mathbf{X}/N$ will be denoted by Σ_Z and Σ_X respectively. We assume that all observations are independent and identically distributed.

The standard estimators for β are, first the OLS estimator:

$$\hat{\beta}_{ols} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right).$$

OLS is not consistent if $\sigma_{\epsilon\eta}$ differs from zero. Its probability limit equals $\beta + (\pi'\Sigma_Z\pi + \Sigma_\eta)^{-1}\sigma_{\epsilon\eta}$. Second, the instrumental variables estimator using the optimal instrument of lowest dimension, $\mathbf{Z}\pi$:

$$\hat{\beta}_{opt} = \left((\mathbf{Z}\pi)'\mathbf{X}\right)^{-1}\left((\mathbf{Z}\pi)'\mathbf{Y}\right).$$

Third, the 2SLS estimator:

$$\hat{\beta}_{2sls} = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}\right)$$

It is useful to work with a characterization of $\hat{\beta}_{2sls}$ as an instrumental variables estimator using the instrument $\mathbf{Z}\hat{\pi}$ where $\hat{\pi} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$. In particular,

$$\hat{\beta}_{2sls} = \left((\mathbf{Z}\hat{\pi})'\mathbf{X}\right)^{-1}\left(\mathbf{Z}\hat{\pi})'\mathbf{Y}\right). \tag{3}$$

The limiting distribution of both $\sqrt{N}(\hat{\beta}_{2sls} - \beta)$ and $\sqrt{N}(\hat{\beta}_{opt} - \beta)$ is normal with mean zero and variance $(\pi'\Sigma_z\pi)^{-1}\sigma_\epsilon^2$.

The main idea behind our approach is that $\hat{\beta}_{opt}$ has much better small sample properties than $\hat{\beta}_{2sls}$ in the presence of many instruments, even though the two estimators have the same asymptotic normal distribution. This follows directly from the Nagar (1959) bias formula, which shows that keeping the explanatory power of the instruments constant while increasing the number of instruments increases the bias of $\hat{\beta}_{2sls}$. The intuition for this is that the first stage fitted values, $\mathbf{Z}\hat{\pi}$, can be written as $P_{\mathbf{Z}}\mathbf{X} = \mathbf{Z}\pi + P_{\mathbf{Z}}\epsilon$ where $P_{\mathbf{Z}}$ is the projection matrix $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. These fitted values are therefore correlated with ϵ. Even though this correlation vanishes in large samples, it increases with the number of instruments for fixed sample size. It should be kept in mind, however, that $\hat{\beta}_{opt}$ does not have any moments, while $\hat{\beta}_{2sls}$ can have moments up to order $K - L$ (Phillips, 1983). We therefore focus on robust measures of dispersion around the true parameter value such as median absolute error.

The previous discussion suggests that the bias of $\hat{\beta}_{2sls}$ towards $\hat{\beta}_{ols}$ is related to the difference between the estimated instrument $Z_i\hat{\pi}$ and the optimal instrument $Z_i\pi$. This leads us to develop new estimators of β based on different estimates of the optimal instrument $Z_i\pi$. The key feature of this approach is that these alternative estimates of the optimal instrument are independent of ϵ_i even in finite samples, unlike the standard estimate $Z_i\hat{\pi}$ which is only asymptotically independent of ϵ_i. While these estimated instruments may have a variance different from the variance of $Z_i\hat{\pi}$, the difference in variance goes to zero fast enough to give the resulting estimators the same first order asymptotic properties as both $\hat{\beta}_{opt}$ and $\hat{\beta}_{2sls}$. The resulting bias reduction is such that in models with many instruments the associated estimators of β are superior to 2SLS, and, in some cases, LIML.

3. Jackknife Instrumental Variables Estimation

Our approach begins with the instrumental variables interpretation of 2SLS. The ith row of the estimated instrument $\mathbf{Z}\hat{\pi}$ underlying 2SLS can be written:

$$Z_i\hat{\pi} = Z_i(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X}) \tag{4}$$

$$= Z_i\pi + Z_i(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\eta).$$

The covariance of this with ϵ_i given Z_i is generally different from zero:

$$E[\epsilon_i Z_i\hat{\pi}|\mathbf{Z}] = Z_i(\mathbf{Z}'\mathbf{Z})^{-1}Z_i' \cdot E[\epsilon_i\eta_i] = Z_i(\mathbf{Z}'\mathbf{Z})^{-1}Z_i' \cdot \sigma_{\epsilon\eta}'.$$

The reason for this is that X_i is used in the construction of $Z_i'\hat{\pi}$, and X_i is correlated with ϵ_i if $\sigma_{\epsilon\eta} \neq 0$.

Let $\mathbf{Z}(i)$ and $\mathbf{X}(i)$ denote matrices equal to $\mathbf{Z}$ and $\mathbf{X}$ with the ith row removed. JIVE1 removes the dependence of the constructed instrument $Z_i'\hat{\pi}$ on the endogenous regressor for observation i by using

$$\tilde{\pi}(i) = (\mathbf{Z}(i)'\mathbf{Z}(i))^{-1}(\mathbf{Z}(i)'\mathbf{X}(i)),$$

as an estimate of π. The estimate of the optimal instrument is

$$Z_i\tilde{\pi}(i) = Z_i(\mathbf{Z}(i)'\mathbf{Z}(i))^{-1}(\mathbf{Z}(i)'\mathbf{X}(i)).$$

Define $\hat{\mathbf{X}}_{jive1}$ to be the $N \times L$ dimensional matrix with ith row $Z_i\tilde{\pi}(i)$. Then the associated estimator for β, denoted by JIVE1, is equal to:

$$\hat{\beta}_{jive1} = \left(\hat{\mathbf{X}}_{jive1}'\mathbf{X}\right)^{-1}\left(\hat{\mathbf{X}}_{jive1}'\mathbf{Y}\right).$$

The JIVE1 estimator appears to require separate calculation of N least squares estimates, $\tilde{\pi}(i)$ for $i = 1, \ldots, N$. In large samples this would be prohibitively expensive. All we actually need, however, is the estimated instrument $Z_i\tilde{\pi}(i)$. This can be calculated using a formula from the literature on influential observations (Cook 1979):

$$Z_i\tilde{\pi}(i) = Z_i\frac{(\mathbf{Z}'\mathbf{Z})^{-1}}{1 - Z_i'(\mathbf{Z}'\mathbf{Z})^{-1}Z_i}(\mathbf{Z}'\mathbf{X} - Z_i'X_i) = \frac{Z_i\hat{\pi} - h_iX_i}{1 - h_i}, \tag{5}$$

where $h_i = Z_i(\mathbf{Z}'\mathbf{Z})^{-1}Z_i$.[2] Given $Z_i\tilde{\pi}(i)$, calculation of $\hat{\beta}_{jive1}$ is straightforward.

An alternative version of this, denoted by JIVE2, removes the dependence on ε_i by adjusting only the $\mathbf{Z}'\mathbf{X}$ component of $\hat{\pi} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$. Define

$$\bar{\pi}(i) = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}(i)'\mathbf{X}(i)) \cdot (N/(N-1)) = (N/(N-1)) \cdot (\hat{\pi} - (\mathbf{Z}'\mathbf{Z})^{-1}Z_i'X_i)$$

as the associated first stage parameter. A formula similar to equation (5) is:

$$Z_i\bar{\pi}(i) = Z_i\frac{(\mathbf{Z}'\mathbf{Z})^{-1}}{1-1/N}(\mathbf{Z}'\mathbf{X} - Z_i'X_i) = \frac{Z_i\hat{\pi} - h_iX_i}{1-1/N}. \tag{6}$$

The resulting estimator for β is:

$$\hat{\beta}_{jive2} = \left(\hat{\mathbf{X}}_{jive2}'\mathbf{X}\right)^{-1}\left(\hat{\mathbf{X}}_{jive2}'\mathbf{Y}\right),$$

where the $N \times L$ dimensional matrix $\hat{\mathbf{X}}_{jive2}$ has ith row equal to $Z_i\bar{\pi}(i)$. Note that the only difference between JIVE1 and JIVE2 is the difference between $1 - Z_i(\mathbf{Z}'\mathbf{Z})^{-1}Z_i'$ and $1 - 1/N$ in the denominator of (5) and (6). Again, this estimator requires only two passes through the data.

4. The Bias of Jackknife Instrumental Variables Estimators

In this section we investigate the bias of the leading terms of an expansion of $\hat{\beta}_{2sls}$, $\hat{\beta}_{jive1}$ and $\hat{\beta}_{jive2}$. This expansion allows us to interpret differences between the estimators as arising from alternative estimates of $Z_i\pi$, and provides intuition for the superior performance of JIVE with many weak instruments. The bias calculation is similar to those by Nagar (1959) and Buse (1992), with the main difference that we expand the estimators of interest around the just-identified instrumental variables estimator $\hat{\beta}_{opt}$.

For any $N \times L$ matrix $\hat{\mathbf{X}}$, we can define the following estimator for β:

$$\hat{\beta}(\hat{\mathbf{X}}) = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{Y}.$$

This is the just-identified estimator for β based on the single instrument $\hat{\mathbf{X}}$. In this notation, the four estimators of interest are:

[2]The term h_i is sometimes called the observation leverage and is computed by many regression packages.

$$\hat{\beta}_{opt} = \hat{\beta}(\mathbf{Z}\pi) = (\pi'\mathbf{Z}'\mathbf{X})^{-1}(\pi'\mathbf{Z}'\mathbf{Y}),$$

$$\hat{\beta}_{2sls} = \hat{\beta}(\hat{\mathbf{X}}_{2sls}) = \hat{\beta}(P_{\mathbf{Z}}\mathbf{X}) = (\mathbf{X}'P_{\mathbf{Z}}\mathbf{X})^{-1}(\mathbf{X}'P_{\mathbf{Z}}\mathbf{Y}),$$

$$\hat{\beta}_{jive1} = \hat{\beta}(\hat{\mathbf{X}}_{jive1}) = (\hat{\mathbf{X}}'_{jive1}\mathbf{X})^{-1}(\hat{\mathbf{X}}'_{jive1}\mathbf{Y}),$$

and

$$\hat{\beta}_{jive2} = \hat{\beta}(\hat{\mathbf{X}}_{jive2}) = (\hat{\mathbf{X}}'_{jive2}\mathbf{X})^{-1}(\hat{\mathbf{X}}'_{jive2}\mathbf{Y}),$$

where the ith row of $\hat{\mathbf{X}}_{jive1}$ and $\hat{\mathbf{X}}_{jive2}$ is defined as before.

The Nagar (1959) and Buse (1992) approximate bias of $\hat{\beta}_{opt}$ equals $-(\pi'\Sigma_z\pi)^{-1}\sigma_{\epsilon\eta}/N$. This is proportional to the covariance of ϵ and η and therefore to the bias of OLS, as is the approximate bias of 2SLS. But unlike the Nagar bias of 2SLS, which equals $(K - L - 1) \cdot (\pi'\Sigma_z\pi)^{-1}\sigma_{\epsilon\eta}/N$, the approximate bias of $\hat{\beta}_{opt}$ is not a function of the number of instruments. The Nagar bias of $\hat{\beta}_{opt}$ therefore provides a natural standard when comparing the finite-sample properties of alternative feasible estimators. We begin our discussion of the bias of JIVE and 2SLS with this comparison.

Note that for 2SLS, JIVE1 and JIVE2 we can write

$$\hat{\mathbf{X}} = \mathbf{Z}\pi + \mathbf{C}\eta,$$

where $\hat{\mathbf{X}}$ is an estimator of $\mathbf{Z}\pi$ with C a $N \times N$ matrix such that the elements of $\mathbf{C}\eta$ are of stochastic order $O_p(1/\sqrt{N})$. The (i,j)th element of $\mathbf{C}_{2sls}$ is:

$$C_{2sls,(i,j)} = Z_i(\mathbf{Z}'\mathbf{Z})^{-1}Z'_j, \tag{7}$$

the (i,j)th element of $\mathbf{C}_{jive1}$ is:

$$C_{jive1,(i,j)} = \begin{cases} Z_i(\mathbf{Z}(i)'\mathbf{Z}(i))^{-1}Z'_j & \text{if } i \neq j \\ 0 & \text{if } i = j, \end{cases} \tag{8}$$

and the (i,j)th element of $\mathbf{C}_{jive2}$ is:

$$C_{jive2,(i,j)} = \begin{cases} Z_i \frac{N}{N-1}(\mathbf{Z}'\mathbf{Z})^{-1} Z_j' & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases} \tag{9}$$

The next step is to derive the (order $1/N$) bias of $\hat{\beta}(\hat{\mathbf{X}})$ relative to the bias of $\hat{\beta}(\mathbf{Z}\pi)$, for any $\hat{\mathbf{X}}$:

$$\hat{\beta}(\hat{\mathbf{X}}) - \hat{\beta}(\mathbf{Z}\pi) = (\hat{\mathbf{X}}'\mathbf{X})^{-1}(\hat{\mathbf{X}}'\mathbf{Y}) - \hat{\beta}(\mathbf{Z}\pi)$$

$$= (\pi'\mathbf{Z}'\mathbf{X} + \eta'\mathbf{C}'\mathbf{X})^{-1}(\pi'\mathbf{Z}'\mathbf{Y} + \eta'\mathbf{C}'\mathbf{Y}) - \hat{\beta}(\mathbf{Z}\pi).$$

Defining $R = (\pi'\mathbf{Z}'X)^{-1}$, we can write this as

$$\hat{\beta}(\hat{\mathbf{X}}) - \hat{\beta}(\mathbf{Z}\pi) = (R^{-1}(\mathcal{I} + R\eta'\mathbf{C}'\mathbf{X}))^{-1}(\pi'\mathbf{Z}'\mathbf{Y} + \eta'\mathbf{C}'\mathbf{Y}) - \hat{\beta}(\mathbf{Z}\pi)$$

$$= (\mathcal{I} + R\eta'\mathbf{C}'\mathbf{X})^{-1}(R\pi'\mathbf{Z}'\mathbf{Y} + R\eta'\mathbf{C}'\mathbf{Y}) - \hat{\beta}(\mathbf{Z}\pi).$$

Expanding $(\mathcal{I} + R\eta'\mathbf{C}'\mathbf{X})^{-1}$ around $R\eta'\mathbf{C}'\mathbf{X} = 0$ and ignoring terms of order less than $1/N$ in the multiplication with $(R\pi'\mathbf{Z}'\mathbf{Y} + R\eta'\mathbf{C}'\mathbf{Y})$ leaves only terms involving $\mathcal{I} - R\eta'\mathbf{C}'\mathbf{X}$. In particular, we have

$$\hat{\beta}(\hat{\mathbf{X}}) - \hat{\beta}(\mathbf{Z}\pi) = (\mathcal{I} - R\eta'\mathbf{C}'\mathbf{X}) \cdot (R\pi'\mathbf{Z}'\mathbf{Y} + R\eta'\mathbf{C}'\mathbf{Y}) - \hat{\beta}(\mathbf{Z}\pi) + o_p(1/N)$$

$$= R\pi'\mathbf{Z}'\mathbf{Y} + R\eta'\mathbf{C}'\mathbf{Y} - R\eta'\mathbf{C}'\mathbf{X}R\pi'\mathbf{Z}'\mathbf{Y} - R\eta'\mathbf{C}'\mathbf{X}R\eta'\mathbf{C}'\mathbf{Y} - \hat{\beta}(\mathbf{Z}\pi) + o_p(1/N).$$

Using the fact that $\hat{\beta}(\mathbf{Z}\pi) = (\pi'\mathbf{Z}'\mathbf{X})^{-1}(\pi'\mathbf{Z}'\mathbf{Y}) = R\pi\mathbf{Z}'\mathbf{Y}$, this equals

$$R\eta'\mathbf{C}'\mathbf{Y} - R\eta'\mathbf{C}'\mathbf{X}\hat{\beta}(\mathbf{Z}\pi) - R\eta'\mathbf{C}'\mathbf{X}R\eta'\mathbf{C}'\mathbf{Y} + o_p(1/N)$$

$$= R\eta'\mathbf{C}'\varepsilon - R\eta'\mathbf{C}'\mathbf{X}(\hat{\beta}(\mathbf{Z}\pi) - \beta) - R\eta'\mathbf{C}'\mathbf{X}R\eta'\mathbf{C}'\mathbf{Y} + o_p(1/N).$$

The third term in this representation, $R\eta'\mathbf{C}'\mathbf{X}R\eta'\mathbf{C}'\mathbf{Y}$, is of lower order than $1/N$ and is ignored in the approximate bias calculation. We therefore focus on the bias coming from the terms

$$R\eta'\mathbf{C}'\varepsilon - R\eta'\mathbf{C}'\mathbf{X}(\hat{\beta}(\mathbf{Z}\pi) - \beta).$$

Expanding the ith row of $\mathbf{X}(\hat{\beta}(\mathbf{Z}\pi) - \beta)$, we have

$$X_i(\hat{\beta}(\mathbf{Z}\pi) - \beta) = X_i(\pi'\mathbf{Z}'\mathbf{X})^{-1}(\pi'\mathbf{Z}'\varepsilon) = Z_i\pi(\pi'\mathbf{Z}'\mathbf{Z}\pi)^{-1}(\pi'\mathbf{Z}'\varepsilon) + o_p(1/\sqrt{N}).$$

Finally, expanding $N \cdot R$ around $R_0 = \text{plim}(\pi'\mathbf{Z}'\mathbf{Z}\pi/N)^{-1} = (\pi\Sigma_{\mathbf{Z}}\pi)^{-1}$, we can write

$$\begin{aligned} & R\eta'\mathbf{C}'\varepsilon - R\eta'\mathbf{C}'\mathbf{X}(\hat{\beta}(\mathbf{Z}\pi) - \beta) \\ & \quad = \frac{1}{N}\left(R_0\eta'\mathbf{C}'\varepsilon - R_0\eta'\mathbf{C}'P_{\mathbf{Z}\pi}\varepsilon\right) + o_p(1/N). \end{aligned} \tag{10}$$

The expectation of the first term is the approximate bias of $\hat{\beta}(\hat{\mathbf{X}})$ relative to $\hat{\beta}(\mathbf{Z}\pi)$, up to order (1/N). The fact that this bias is nonzero stems from the nonvanishing covariance between $\mathbf{C}\eta$ and ε.

The final step is to evaluate the expected value of (10) for 2SLS, JIVE1 and JIVE2. For $\hat{\beta}(\hat{\mathbf{X}}_{2sls}) - \hat{\beta}(\mathbf{Z}\pi)$, we have:

$$\begin{aligned} \text{bias}_{2sls} &= E[(R_0/N)\eta'\mathbf{C}'_{2sls}\varepsilon - (R_0/N)\eta'\mathbf{C}'_{2sls}P_{\mathbf{Z}\pi}\varepsilon] \\ &= E[(R_0/N)\eta'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}\varepsilon - (R_0/N)\eta'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}P_{\mathbf{Z}\pi}\varepsilon] \\ &= (R_0/N)E[\eta'(P_Z - P_Z P_{Z\pi} P_Z)\varepsilon]. \end{aligned} \tag{11}$$

In the last equality we use the fact that $P_{Z\pi} = P_{Z\pi}P_Z$. The ith element of the bias vector is equal to $(R_0/N) \cdot \text{trace}(P_{\mathbf{Z}} - P_{\mathbf{Z}}P_{\mathbf{Z}\pi}P_{\mathbf{Z}})$ times the ith element of $\sigma'_{\varepsilon\eta}$. Note that $(P_Z - P_Z P_{Z\pi} P_Z)$ is an idempotent matrix with rank and trace equal to $K - L$, the number of overidentifying restrictions. This implies that the expectation of (11) equals

$$\text{bias}_{2sls} = \frac{1}{N}(K - L)R_0\sigma_{\varepsilon\eta}, \tag{12}$$

which is the difference between the bias of $\hat{\beta}_{2sls}$ and the bias of $\hat{\beta}_{opt}$. This is the same result one would obtain by direct application of the Nagar formula to $\hat{\beta}_{2sls}$ and $\hat{\beta}_{opt}$.

Next, we evaluate (10) for $\hat{\beta}_{jive2}$,

$$\text{bias}_{jive2} = \frac{1}{N} E\left(R_0 \eta' \mathbf{C}'_{jive2} \varepsilon - R_0 \eta' \mathbf{C}'_{jive2} P_{\mathbf{Z}\pi} \varepsilon\right). \tag{13}$$

Because $\mathbf{C}_{jive2}$ has diagonal elements equal to zero, the ith element of $\mathbf{C}_{jive2}\eta$ is independent of η_i and therefore of ϵ_i. Hence, the expectation of the first term in (13) is zero. To evaluate the second term,

$$-E R_0 \eta' \mathbf{C}'_{jive2} P_{\mathbf{Z}\pi} \varepsilon,$$

we use the fact that

$$\mathbf{C}_{jive2} = \frac{N}{N-1} \cdot (\mathbf{C}_{2sls} - D_{jive2}),$$

where D_{jive2} is a diagonal matrix with ith element $Z_i(\mathbf{Z}'\mathbf{Z})^{-1} Z_i'$, equal to the ith diagonal element of $\mathbf{C}_{2sls}$. Therefore

$$-E\eta' \mathbf{C}'_{jive2} P_{\mathbf{Z}\pi} \varepsilon = -\frac{N}{N-1} \cdot E\eta' \mathbf{C}'_{2sls} P_{\mathbf{Z}\pi} \varepsilon + \frac{N}{N-1} \cdot E\eta' D'_{jive2} P_{\mathbf{Z}\pi} \varepsilon$$

$$= \frac{N}{N-1} \cdot \left(-L\sigma_{\varepsilon\eta} + \text{trace}(D_{jive2} P_{\mathbf{Z}\pi})\sigma_{\varepsilon\eta}\right) = -L\sigma_{\varepsilon\eta} + o_p(1).$$

The last equality follows because order of the trace of $D_{jive2} P_{\mathbf{Z}\pi}$ is smaller than that of the leading term. The bias of JIVE2 relative to $\hat{\beta}_{opt}$ is therefore $-(R_0/N) \cdot L\sigma_{\varepsilon\eta}$ up to order $1/N$. Finally, we can use the same argument to show that

$$\text{bias}_{jive1} = \frac{1}{N} E\left(R_0 \eta' \mathbf{C}'_{jive1} \varepsilon - R_0 \eta' \mathbf{C}'_{jive1} P_{\mathbf{Z}\pi} \varepsilon\right) = -(R_0/N) \cdot L\sigma_{\varepsilon\eta}. \tag{14}$$

Thus, the approximate bias of JIVE1 is the same as the approximate bias of JIVE2.

This argument shows that in contrast to the bias of 2SLS, the approximate bias of JIVE1 and JIVE2 does not increase with the number of overidentifying restrictions. When the number of instruments K is much larger than the number of regressors L, the bias of JIVE1 and JIVE2 is therefore likely to be smaller than that of 2SLS. In the special case where $L = 1$, the difference in the approximate bias of the JIVE estimators and $\hat{\beta}(\mathbf{Z}\pi)$ is equal

to minus the approximate bias of $\hat{\beta}(\mathbf{Z}\pi)$, so that the JIVE estimators are approximately unbiased in this case.

5. Asymptotics Based on an Increasing Number of Instruments

In a recent paper, Bekker (1994) compares a number of traditional (single-equation) simultaneous equations estimators using an alternative asymptotic approximation where the number of instruments increases with sample size while keeping the explanatory power of the instruments constant. Bekker (1994) shows that LIML is consistent under this parameter sequence but 2SLS is not. The Bekker parameter sequence is justified by a range of Monte Carlo evidence showing that in practice it can give a good account of finite-sample properties.

In this section we show that JIVE1 and JIVE2 are consistent under the Bekker parameter sequence. Our argument mirrors the version of this sequence that Angrist and Krueger (1995) refer to as "group–asymptotics." Group–asymptotics amounts to drawing new i.i.d. replications of $\mathbf{X}$, $\mathbf{Y}$, and new instrumental variables $\mathbf{Z}$ so that K/N is fixed at a number, k. The new instruments, however, are uncorrelated with $\mathbf{X}$. There are no other restrictions on the distribution of the additional instruments except that $E[Z_i'Z_i]$ is finite for all N. The sequence of first stage coefficients is $\pi_N = (\pi_0', \tilde{\pi}_N')'$, with π_0 a vector of fixed length l, and $\tilde{\pi}_N$ a vector of zeros of length $k \cdot N - l = K - l$, because the additional instruments are uncorrelated with the endogenous regressor. The first-stage normalized population sum of squares, $\pi'\mathbf{Z}'\mathbf{Z}\pi/N$, (sometimes called the concentration parameter) is therefore also fixed.

Let $gplim(\hat{\beta})$ denote the group–asymptotic probability limit of the estimator $\hat{\beta}$. Under this sequence, it is easy to show that the group–asymptotic probability limit of all estimators of the type $\hat{\beta}(\hat{\mathbf{X}})$ can be written as

$$gplim(\hat{\beta}(\hat{\mathbf{X}})) = E(\hat{\mathbf{X}}'\mathbf{X}/N)^{-1} \cdot E(\hat{\mathbf{X}}'\mathbf{Y}/N). \tag{15}$$

In fact, group–asymptotics can be thought of as a way to rationalize passing expectations through a ratio.[3]

[3]Stoker (1995) uses a similar approach to characterize the finite-sample bias of non-parametric regression

Recall that $\hat{\mathbf{X}}$ is a generic fitted value appropriate for either 2SLS, JIVE1, or JIVE by choice of the matrix C. For any C, we have

$$\begin{aligned}\mathrm{E}(\hat{\mathbf{X}}'\mathbf{X}/N) &= \mathrm{E}((\mathbf{Z}\pi + \mathrm{C}\eta)'(\mathbf{Z}\pi + \eta)/N) \\ &= \pi'\Sigma_{\mathbf{Z}}\pi + \mathrm{E}(\pi'\mathbf{Z}'\eta/N) + \mathrm{E}(\eta'\mathrm{C}'\mathbf{Z}\pi/N) + \mathrm{E}(\eta'\mathrm{C}'\eta/N).\end{aligned}$$

The second and third terms are zero for C_{jive1}, C_{jive2} and C_{2sls} because η and $\mathbf{Z}$ are mean independent. The last term is zero for C_{jive1} and C_{jive2} because C_{jive1} and C_{jive2} have zeros on the diagonal, and hence, a trace of zero. For 2SLS, the last term is not zero but rather $k\Sigma_\eta$.

Similarly, for the second part of $\hat{\beta}(\hat{\mathbf{X}})$, we have

$$\begin{aligned}\mathrm{E}(\hat{\mathbf{X}}'\mathbf{Y}/N) &= \mathrm{E}((\mathbf{Z}\pi + \mathrm{C}\eta)'(\mathbf{Z}\pi\beta + \eta\beta + \varepsilon)/N) \\ &= (\pi'\Sigma_{\mathbf{Z}}\pi)\beta + \mathrm{E}(\pi'\mathbf{Z}'\eta\beta/N) + \mathrm{E}(\eta'\mathrm{C}'\mathbf{Z}\pi\beta/N) \\ &\quad + \mathrm{E}(\eta'\mathrm{C}'\eta\beta/N) + \mathrm{E}(\eta'\mathrm{C}'\varepsilon/N).\end{aligned}$$

The same argument as before implies that for C_{jive1} and C_{jive2} all terms other than the first are equal to zero. This establishes the group–asymptotic consistency of the JIVE estimators.

For C_{2sls}, the last two terms in the above expression differ from zero. In particular, $\mathrm{E}(\eta'\mathrm{C}\varepsilon/N) = k\sigma_{\varepsilon\eta}$ so that

$$gplim(\hat{\beta}_{2sls}) = \beta + (\pi'\Sigma_z\pi + k\Sigma_\eta)^{-1}k\sigma_{\varepsilon\eta}.$$

Finally, note that a similar argument can be used to show that the group–asymptotic probability limit of $\hat{\beta}_{ols}$ is

$$gplim(\hat{\beta}_{ols}) = \beta + (\pi'\Sigma_z\pi + \Sigma_\eta)^{-1}\sigma_{\varepsilon\eta}.$$

estimators.

5.1 Combination estimators and LIML

The argument in the previous section shows that the bias of 2SLS is proportional to the bias of OLS under group-asymptotics, as well as under the approximation argument developed in Section 4. This suggests that a linear combination of 2SLS and OLS can have less bias than either estimator alone. Interest in such "combination estimators" has a long history in the literature on finite-sample properties of simultaneous equations estimators (see, e.g., Sawa 1973, and more recently Staiger and Stock 1994).

Our interest in combination estimators stems from the link between these estimators and LIML. Like the JIVE estimators, LIML does not share the many-instruments bias of 2SLS towards OLS. For example, Bekker (1994) shows that LIML is consistent under group-asymptotics. As a practical matter, our simulations show that LIML is approximately median-unbiased.[4] This section provides an alternative proof of the group-asymptotic consistency of LIML and an intuitive explanation for this result.

Consider the following theoretical combination estimator,

$$\begin{aligned}\tilde{\beta} &= \left(\pi'\Sigma_z\pi + k\Sigma_\eta - k\Sigma_x\right)^{-1} \cdot \left((\mathbf{X}'P_\mathbf{Z}\mathbf{X})\hat{\beta}_{2sls} - k\cdot(\mathbf{X}'\mathbf{X})\hat{\beta}_{ols}\right) \\ &= \left(\pi'\Sigma_z\pi\cdot(1-k)\right)^{-1} \cdot \left((\mathbf{X}'P_\mathbf{Z}\mathbf{X})\hat{\beta}_{2sls} - k\cdot(\mathbf{X}'\mathbf{X})\hat{\beta}_{ols}\right). \qquad (16)\end{aligned}$$

This estimator can be motivated from the group-asymptotic probability limits of $\hat{\beta}_{2sls}$ and $\hat{\beta}_{ols}$, or by direct calculation, which shows that $\tilde{\beta}$ is actually unbiased. It is clearly not a feasible estimator, however, because it requires knowledge of population parameters. A feasible version of $\tilde{\beta}$ is obtained by replacing Σ_x with $\mathbf{X}'\mathbf{X}$ and $(\pi'\Sigma_z\pi + k\Sigma_\eta)$ with $\mathbf{X}'P_\mathbf{Z}\mathbf{X}$, to give:

$$\tilde{\beta} = \left[\mathbf{X}'P_\mathbf{Z}\mathbf{X} - k\mathbf{X}'\mathbf{X}\right]^{-1}\left[\mathbf{X}P_\mathbf{Z}\mathbf{X}\hat{\beta}_{2sls} - k\mathbf{X}'\mathbf{X}\hat{\beta}_{ols}\right].$$

[4]The approximate median-unbiasedness of LIML has been noted by many authors. See, e.g., Anderson, Kunitomo, and Sawa (1982).

We now show that $\hat{\beta}_{liml}$ corresponds to a version of $\bar{\beta}$ with k replaced by a random variable that estimates k. This provides some intuition for the superior performance of LIML relative to 2SLS in situations where 2SLS is likely to badly biased, and establishes that $gplim(\hat{\beta}_{liml})$ equals β.

Define $\lambda(\beta)$ as

$$\lambda(\beta) = \frac{\varepsilon' P_{\mathbf{Z}} \varepsilon}{\varepsilon'\varepsilon} = \frac{(\mathbf{Y} - \mathbf{X}\beta)' P_{\mathbf{Z}} (\mathbf{Y} - \mathbf{X}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)},$$

where $\varepsilon = \mathbf{Y} - \mathbf{X}\beta$. Note that the group–asymptotic probability limit of $\lambda(\beta)$ evaluated at the true value of β is $\mathrm{E}(\varepsilon' P_{\mathbf{Z}} \varepsilon)/\mathrm{E}(\varepsilon'\varepsilon) = k$.

The LIML estimator can be defined as the solution to

$$\min_{\beta} \lambda(\beta),$$

which has the same solution as

$$\min_{\beta} \Big[\ln\big((\mathbf{Y} - \mathbf{X}\beta)' P_{\mathbf{Z}} (\mathbf{Y} - \mathbf{X}\beta)\big) - \ln\big((\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)\big) \Big].$$

The first order conditions for this minimization problem can be written

$$\mathbf{X}' P_{\mathbf{Z}} \mathbf{Y} - \lambda(\beta) \mathbf{X}'\mathbf{Y} = (\mathbf{X}' P_{\mathbf{Z}} \mathbf{X} - \lambda(\beta) \mathbf{X}'\mathbf{X})\beta,$$

which implies

$$\begin{aligned} \hat{\beta}_{liml} &= \left[\mathbf{X}' P_{\mathbf{Z}} \mathbf{X} - \lambda(\hat{\beta}_{liml}) \mathbf{X}'\mathbf{X}\right]^{-1} \left[\mathbf{X} P_{\mathbf{Z}} \mathbf{Y} - \lambda(\hat{\beta}_{liml}) \mathbf{X}'\mathbf{Y}\right] \\ &= \left[\mathbf{X}' P_{\mathbf{Z}} \mathbf{X} - \lambda(\hat{\beta}_{liml}) \mathbf{X}'\mathbf{X}\right]^{-1} \left[\mathbf{X} P_{\mathbf{Z}} \mathbf{X} \hat{\beta}_{2sls} - \lambda(\hat{\beta}_{liml}) \mathbf{X}'\mathbf{X} \hat{\beta}_{ols}\right]. \end{aligned} \tag{17}$$

This is $\bar{\beta}$ with k replaced by $\lambda(\hat{\beta}_{liml})$.

To complete this section, note that writing $\hat{\beta}_{liml}$ as in the previous equation leads to a straightforward argument for the group–asymptotic consistency of $\hat{\beta}_{liml}$. Because matrix inversion and $\lambda(\hat{\beta}_{liml})$ are continuous functions, we have

$$gplim(\hat{\beta}_{liml}) = E\left[\mathbf{X}'P_{\mathbf{Z}}\mathbf{X} - gplim(\lambda(gplim(\hat{\beta}_{liml}))\mathbf{X}'\mathbf{X}\right]^{-1}$$

$$\cdot E\left[\mathbf{X}P_{\mathbf{Z}}\mathbf{X}\hat{\beta}_{2sls} - gplim(\lambda(gplim(\hat{\beta}_{liml})))\mathbf{X}'\mathbf{X}\hat{\beta}_{ols}\right]$$

Because $gplim(\lambda(\beta))$ at the true value of β is k, $\text{gplim}(\beta_{liml}) = \beta$ is a solution to this equation. If the model is identified, this solution will be unique.

One reason the interpretation of $\hat{\beta}_{liml}$ as the sample analog of an unbiased combination estimator is useful is that this interpretation suggests when LIML is likely to be an unattractive estimation technique. In particular, unlike 2SLS or the JIVE estimators, LIML estimation requires that $\lambda(\hat{\beta}_{liml})$ be close to k. $\hat{\beta}_{liml}$ is therefore also likely to be more sensitive than either 2SLS or JIVE to deviations from perfect instrument-error orthogonality because such deviations tend to increase $\lambda(\hat{\beta})$.

Another example when 2SLS or JIVE might be preferred to LIML is when some instruments are erroneously excluded from the second stage when they should included as covariates (for example, when lagged dependent variables are used as instruments when they should be used as controls.) If the erroneously excluded instruments are actually uncorrelated with the endogenous regressor, then 2SLS, JIVE1 and JIVE2 will still be consistent. But because of the functional relationship between $\hat{\beta}_{liml}$ and $\lambda(\beta)$, LIML is not consistent under this sort of mis-specification.

6. Monte Carlo Study

6.1 Study Design and Monte Carlo Statistics

In this section, we report evidence on the finite sample behavior of the estimators proposed in this paper, focussing on robust measures of bias. In particular, we report quantiles of the Monte Carlo sampling distribution along with the median absolute error. Mean squared error is not likely to be as useful a standard for comparison because neither LIML or the JIVE estimators have moments.

We also report coverage rates for 95 percent confidence intervals computed using the usual asymptotic approximation to the distribution of OLS, 2SLS, and LIML (i.e., the estimate plus or minus 1.96 times the asymptotic standard error.) For the JIVE estimators, we report confidence intervals based on asymptotic standard errors for a just–identified IV estimator using $\hat{\mathbf{X}}_{jive1}$ and $\hat{\mathbf{X}}_{jive2}$ as instruments. The justification for this is pragmatic: if the usual approximation works in the sense of providing accurate coverage for the approximately unbiased LIML and JIVE estimators, there would seem to be little reason to report more sophisticated approximations such as those developed by Stock and Staiger (1994).

In fact, Bekker (1994) finds that some theoretically more accurate approximations to the limiting distribution of LIML based on group-asymptotics provide little or no improvement over the usual asymptotic approximation in most cases. This is not true for 2SLS, however. The results of our simulations confirm and extend this: asymptotic confidence intervals for the approximately unbiased LIML and JIVE estimators turn out to be remarkably accurate while conventional asymptotic confidence intervals for 2SLS are quite poor.

6.2 Models and Results

We begin with a model where there is a single overidentifying restriction. The second model is similar, with the modification that there are a large number of instruments relative to the number of regressors. In both of these first two models, the errors are homoscedastic and the first stage regression is linear, so that LIML is the maximum likelihood estimator. In the third model, the first stage is nonlinear and heteroscedastic. Here LIML is less likely to have good small sample properties since it is no longer the maximum likelihood estimator. In both of the latter two models, 2SLS should be badly biased because of the large number of overidentifying restrictions. The last model sets the true reduced form coefficients to zero for all instruments in an attempt to ascertain how misleading the estimators might be in this non–identified case.

The models and results are as follows:

Model 1

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \epsilon_i$$

$$X_{i1} = \pi_0 + \sum_{j=1}^{2} \pi_j \cdot Z_{ij} + \eta_i$$

with $\beta_1 = 1$, $\beta_0 = 0$, $\pi_0 = 0$, $\pi_1 = 0.3$, and $\pi_2 = 0$. Here, K=3 and L=2, and

$$\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & 0.20 \\ 0.20 & 0.25 \end{pmatrix} \right).$$

All Z_{ij} are independent, normally distributed random variables with mean zero and unit variance.

Table 1 presents quantiles of the sampling distributions of the estimators, as well as the median absolute error and coverage rates. In this set of simulations, LIML, JIVE1 and

Table 1: MODEL 1: N=100, L=2, K=3; 5,000 REPLICATIONS

	Quantiles Around β_1					median	coverage rate
estimator	0.10	0.25	0.50	0.75	0.90	absolute error	95% conf. interval
OLS	0.50	0.55	0.59	0.64	0.67	0.59	0.00
2SLS	-0.19	-0.06	0.04	0.14	0.22	0.11	0.91
LIML	-0.26	-0.13	0.00	0.11	0.19	0.12	0.96
JIVE1	-0.40	-0.20	-0.05	0.07	0.17	0.13	0.96
JIVE2	-0.40	-0.20	-0.05	0.07	0.17	0.13	0.96

JIVE2 all have median absolute error close to that of 2SLS, which is the estimator with the minimum median absolute error. But confidence interval coverage is actually more accurate for JIVE and LIML than for 2SLS. One reason LIML is better is that it has a more symmetric distribution. It is not surprising that LIML does very well, however, since it comes from the normal likelihood function and in this example the disturbances are in fact normal. Note that confidence interval coverage for JIVE is as good as that for LIML, in spite of some asymmetry in the Monte Carlo sampling distribution of JIVE.

Figure 1 presents the distribution functions of the sampling distributions for the five estimators. The approximate median unbiasedness of 2SLS, LIML, JIVE1 and JIVE2 for this example shows up in the proximity of the intersection of the distribution function and the vertical line at zero to the intersection of the distribution function and the horizontal line at 0.5.

Model 2

Model 2 adds 18 worthless instruments to the design in Model 1. This is a situation where we expect the performance of 2SLS to deteriorate.

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \epsilon_i$$

$$X_{i1} = \pi_0 + \sum_{j=1}^{20} \pi_j \cdot Z_{ij} + \eta_i$$

with $\beta_1 = 1$, $\beta_0 = 0$, $\pi_0 = 0$, $\pi_1 = 0.3$, and $\pi_j = 0$ for $j = 2, 3, \ldots, 20$. Here, K=21 and L=2, and

$$\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & 0.20 \\ 0.20 & 0.25 \end{pmatrix}\right).$$

All Z_{ij} are independent, normally distributed random variables with mean zero and unit variance.

Table 2 presents Monte Carlo statistics for this model. In this set of simulations, LIML, JIVE1 and JIVE2 are all superior to 2SLS and OLS in terms of median absolute error. Unlike 2SLS and OLS, the three other estimators are essentially median unbiased and the asymptotic confidence intervals have very good coverage. LIML is less dispersed than both JIVE1 and JIVE2 with the latter having thick tails. The asymptotic coverage for 2SLS is also poor. Again, it is not surprising that LIML does very well here since in this example the disturbances are normally distributed.

Figure 2 presents the distribution functions of the sampling distributions for the five estimators.

Table 2: MODEL 2: N=100, L=2, K=21; 5,000 REPLICATIONS

	Quantiles Around β_1					median	coverage rate
estimator	0.10	0.25	0.50	0.75	0.90	absolute error	95% conf. interval
OLS	0.51	0.55	0.59	0.63	0.67	0.59	0.00
2SLS	0.14	0.21	0.28	0.35	0.41	0.28	0.31
LIML	-0.31	-0.14	0.00	0.11	0.20	0.13	0.94
JIVE1	-0.61	-0.28	-0.04	0.12	0.23	0.17	0.94
JIVE2	-0.63	-0.29	-0.04	0.11	0.23	0.17	0.94

Model 3

The third model has the same basic structure as before, except that the relationship between X_i and Z_i is nonlinear and heteroscedastic. But as in model 2, there are 20 linear instruments so that nonlinearities in the first stage are ignored in the estimation.

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \epsilon_i$$

$$X_{i1} = \pi_0 + \sum_{j=1}^{20} \pi_j \cdot Z_{ij} + 0.3 \cdot \sum_{j=2}^{20} Z_{ij}^2 + \eta_{i0} \cdot \sum_{j=2}^{20} Z_{ij}^2/19$$

with $\beta_1 = 1$, $\beta_0 = 0$, $\pi_0 = 0$, $\pi_1 = 0.3$, and $\pi_j = 0$ for $j = 2, 3, \ldots, 20$. In the estimation, K=21 and L=2, and

$$\begin{pmatrix} \epsilon_i \\ \eta_{i0} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{pmatrix} \right).$$

Table 3 presents Monte Carlo statistics for this model. As expected, OLS and 2SLS are still biased, as evidence by the fact that almost all probability is concentrated on one side of the true value of β_1 for these estimators. Moreover, in spite of the low median absolute error of 2SLS in this case, the asymptotic coverage of 2SLS is very poor.

JIVE1 and LIML do not do as well in Model 3 as in Models 1 and 2. But JIVE2 is the best estimator in terms of median-bias and median absolute error. It is clearly superior to LIML and even to JIVE1 in this model, both in terms of bias and (slightly) in terms of asymptotic coverage. The medians of JIVE1, LIML, and 2SLS are all similar. The big

difference in spread between JIVE1 and JIVE2 is surprising and only in this sort of nonlinear example have we seen such a difference. It is important to note, however, that in contrast with 2SLS, even the highly dispersed JIVE1 generates an asymptotic confidence interval with reasonably accurate coverage. The lack of dispersion in 2SLS, reflected correctly in the 2SLS asymptotic standard errors, actually leads to highly misleading inferences.

Table 3: MODEL 3: N=100, L=2, K=21; 5,000 REPLICATIONS

	Quantiles Around β_1					median	coverage rate
estimator	0.10	0.25	0.50	0.75	0.90	absolute error	95% conf. interval
OLS	0.12	0.14	0.17	0.20	0.23	0.17	0.03
2SLS	0.04	0.10	0.16	0.22	0.27	0.16	0.57
LIML	-0.59	-0.15	0.10	0.32	0.80	0.25	0.97
JIVE1	-0.69	-0.13	0.16	0.43	0.95	0.32	0.97
JIVE2	-0.41	-0.13	0.04	0.16	0.33	0.15	0.95

Figure 3 presents the distribution functions of the sampling distributions for the five estimators in this model.

Model 4

The fourth model has the same basic structure as model 2 but all coefficients in the reduced form are set to zero.

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \varepsilon_i$$

$$X_{i1} = \pi_0 + \sum_{j=1}^{20} \pi_j \cdot Z_{ij} + \eta_i$$

with $\beta_1 = 1$, $\beta_0 = 0$, and $\pi_j = 0$ for all j. Again, L=2 and K=21, and

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & 0.20 \\ 0.20 & 0.25 \end{pmatrix} \right).$$

Table 4 presents Monte Carlo statistics for this model. The two JIVE estimators and LIML are much more dispersed than either OLS or 2SLS in this case, suggesting that a

Table 4: MODEL 4: N=100, L=2, K=21; 5,000 REPLICATIONS

	Quantiles Around β_1					median	coverage rate
estimator	0.10	0.25	0.50	0.75	0.90	absolute error	95% conf. interval
OLS	0.72	0.76	0.80	0.84	0.87	0.80	0.00
2SLS	0.62	0.71	0.80	0.89	0.97	0.80	0.00
LIML	-1.14	0.18	0.81	1.42	2.69	1.01	0.71
JIVE1	-0.40	0.41	0.80	1.21	2.07	0.88	0.71
JIVE2	-0.35	0.41	0.80	1.20	2.05	0.88	0.71

researcher would not be misled by JIVE or LIML estimates into thinking that the instruments generate reliable inferences regarding the coefficient of interest. It is also interesting to note that the correlation between JIVE and LIML in this model is very low, unlike in models where the instruments are valid. This suggests that a comparison of JIVE and LIML could provide a useful check on the validity of inferences in applications with weak instruments.

Figure 4 presents the distribution functions of the sampling distributions of the five estimators for this model.

7. RETURNS TO EDUCATION USING QUARTER OF BIRTH AS INSTRUMENT

In this section, we return to the Angrist and Krueger (1991) application that has motivated some of the recent literature on instrumental variables estimates with many weak instruments. Angrist and Krueger (1991) estimated schooling coefficients using quarter of birth as an instrument in a sample of 329,500 men born 1930-39 from the 1980 census. The dependent variable is the log weekly wage. In one version of this model, there are 30 instruments created by interacting quarter and year of birth. In a second version there are 180 instruments constructed by adding interactions of 50 state and quarter of birth dummies to the 30 original instruments. The appendix to Angrist and Krueger (1991) provides a detailed description of the data.

Table 5 reports schooling coefficients generated by different estimators applied to the

Angrist and Krueger data. Exogenous covariates are listed in the table (these are either state effects or state and year effects.) Table 5 shows that all IV estimators give very similar

Table 5: ANGRIST–KRUEGER DATA

no of instr.	state effects	year effect	ols	2sls	liml	jive1	jive2
30	no	yes	0.071	0.084	0.093	0.096	0.096
			(0.0003)	(0.016)	(0.018)	(0.022)	(0.022)
180	yes	yes	0.067	0.093	0.106	0.119	0.119
			(0.0003)	(0.009)	(0.011)	(0.064)	(0.064)

results. This is important because Bound, Jaeger and Baker (1995) and Angrist and Krueger (1995) note that if the instruments were in fact uncorrelated with schooling, 2SLS could still give results very close to OLS. In contrast, the two JIVE estimators and LIML would not be expected to give similar or statistically significant estimates in such circumstances.

Another notable finding in Table 5 is that the asymptotic standard errors of the JIVE estimators are quite large for the 180-instrument specification. The fact that the coverage provided by asymptotic confidence intervals appears to be pretty good in the Monte Carlo study suggests that the reported standard errors are an accurate reflection of the large sampling variance of JIVE in this case. The JIVE standard errors are actually larger than those reported for a similar specification using an estimator that estimates the first and second stage parameters in separate half samples (USSIV, Angrist and Krueger 1995). The USSIV estimator works as follows: suppose the data are split into half samples, with data matrices $(\mathbf{Y}_1, \mathbf{X}_1, \mathbf{Z}_1)$ and $(\mathbf{Y}_2, \mathbf{X}_2, \mathbf{Z}_2)$ Then USSIV is

$$\hat{\beta}_u = \left(\hat{\mathbf{X}}_{21}'\mathbf{X}_1\right)^{-1}\left(\hat{\mathbf{X}}_{21}'\mathbf{Y}_1\right),$$

where $\hat{\mathbf{X}}_{21} = \mathbf{Z}_1(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{Y}_2$. The USSIV estimator has bias properties similar to those of JIVE.

The apparently unfavorable comparison with USSIV is puzzling because JIVE is asymptotically equivalent to the efficient 2SLS estimator while USSIV is not. But the USSIV standard errors reported by Angrist and Krueger (1995) turn out to be incorrect because they fail to take account of the random split into half samples. A random split clearly generates additional sampling variance even in a single data set. This mistake highlights another advantage of JIVE: there is no need to take account of a random sample split when calculating sampling variance.

Finally, we note that in a comment on an earlier version of this paper, Blomquist and Dahlberg (1994) present an extensive Monte Carlo comparison of JIVE and USSIV, along with another split-sample estimator discussed by Angrist and Krueger (1995) called SSIV. They find that JIVE is typically the minimum mean squared error estimator in the group of approximately unbiased estimators that they consider.

8. Conclusion

In this paper we present two alternatives to 2SLS, LIML and other k–class estimators for models with endogenous regressors. These estimators perform much better than 2SLS in models with many weak instruments, and have finite sample properties similar to those of LIML. Moreover, simulations and a theoretical argument based on group-asymptotics suggest that JIVE estimators combine the attractive bias properties of LIML with the robustness of 2SLS. The JIVE estimators therefore seem to provide useful alternatives in applications where there is concern about the number of instruments.

Instrumental variables is one special case in a larger class of generalized method of moments models where a weight matrix is estimated in an initial stage and a weighted set of restrictions is imposed in a second stage. In some cases, using the same data set to estimate the weight matrix and to impose the moment restrictions leads to poor small sample properties. In this context, Altonji and Segal (1994) discuss a sample splitting approach similar to that used by Angrist and Krueger (1995). The jackknife idea developed here for

instrumental variables extends to moment estimators such as those considered by Altonji and Segal. Developing this extension appears to be a natural avenue for future research.

BIBLIOGRAPHY

ALTONJI, J., AND L. SEGAL, (1994), "Small-Sample Bias in GMM Estimation of Covariance Structures," NBER technical working paper, June.

ANDERSON, T.W., N. KUNITOMO, AND T. SAWA, (1982), "Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator," *Econometrica*, 50, 1009-1027.

ANGRIST, J., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

ANGRIST, J. AND A. KRUEGER, (1991), "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics*, 106, 979-1014.

ANGRIST, J. AND A. KRUEGER, (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association* 87, June.

ANGRIST, J. D., AND A. KRUEGER, (1995), "Split Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, April.

BEKKER, P.,, (1994), "Alternative Approximations to the Distributions of Instrumental Variables Estimators", *Econometrica*, 62, 657–682.

BLOMQUIST, S., AND M. DAHLBERG, (1994), "Small Sample Properties of Jackknife Instrumental Variables Estimators: Experiments with Weak Instruments", mimeo, Uppsala University, August.

BOUND, J., D. JAEGER, AND R. BAKER, (1995), "Problems with Instrumental Variables Estimation when the Correlation between Instruments and the Endogenous Explanatory Variable is Weak", forthcoming, em Journal of the American Statistical Association.

BUSE, A. (1992), "The Bias of Instrumental Variables Estimators", *Econometrica*, 60, 173–180.

COOK, R. D., (1979), "Influential Observations in Linear Regressions", *Journal of the American Statistical Association*, 74, 169–174.

FISHER, F., (1966), "The Relative Sensitivity to Specification Error of Different k-class Estimators," *Journal of the American Statistical Association*, Vol 61, 345–356.

FISHER, F., (1967), "Approximate Specification and the Choice of a k–class Estimator," *Journal of the American Statistical Association*, Vol 62, 1265–1276.

MADDALA, G. S., AND J. JEONG, (1992), "On the Exact Small Sample Distribution of the Instrumental Variables Estimator", *Econometrica*, 60, 181–183.

NAGAR, A. L., (1959), "The Bias and Moment Matrix of the General k–Class Estimators of the Parameters in Simultaneous Equations", *Econometrica*, 27, 575–595.

NELSON, C., AND R. STARZ, (1990), "The Distribution of the Instrumental Variables Estimator and its t–ratio when the Instrument is a Poor One", *Journal of Business*.

PHILLIPS, P.C.B., (1983), "Exact Small Sample Properties in the Simultaneous Equations Model," Chapter 8 in Z. Griliches and M. Intriligator, eds., *The Handbook of Econometrics*, Amsterdam: North Holland.

SAWA, T., (1973), "An Almost Unbiased Estimator in Simultaneous Equations Systems," *International Economic Review*, 14.

STAIGER, D., AND J. STOCK, (1994), "Instrumental Variables Regression with Weak Instruments", NBER Technical Working Paper, No. 151, January.

STOKER, T., (1995), "Smoothing Bias in the Measurement of Marginal Effects", forthcoming, *Journal of Econometrics*.

Figure 1: Distribution Functions for OLS, 2SLS, LIML, JIVE1, JIVE2 (Model 1)

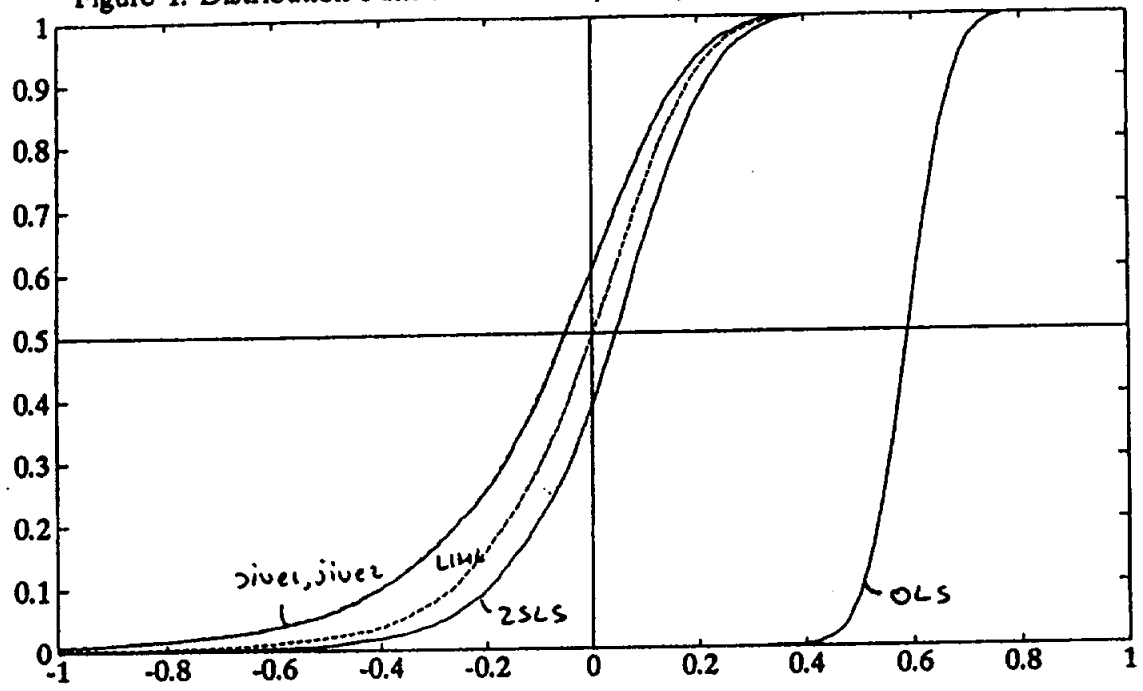

Figure 2: Distribution Functions for OLS, 2SLS, LIML, JIVE1, JIVE2 (Model 2)

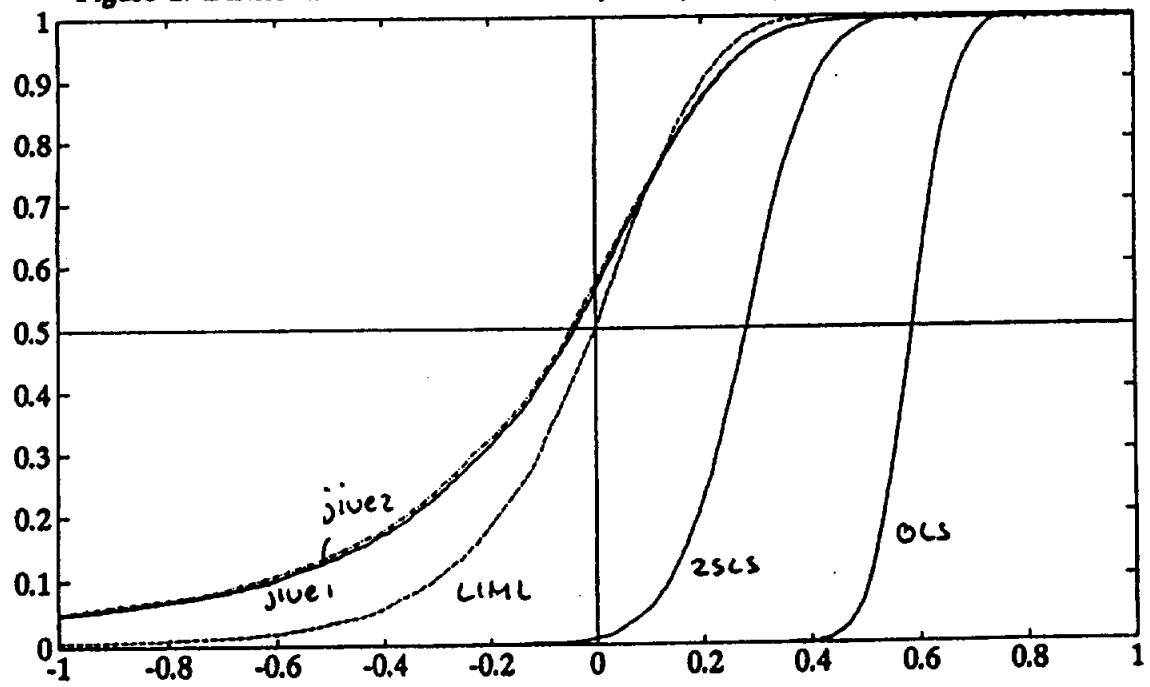

Figure 3: Distribution Functions for OLS, 2SLS, LIML, JIVE1, JIVE2 (Model 3)

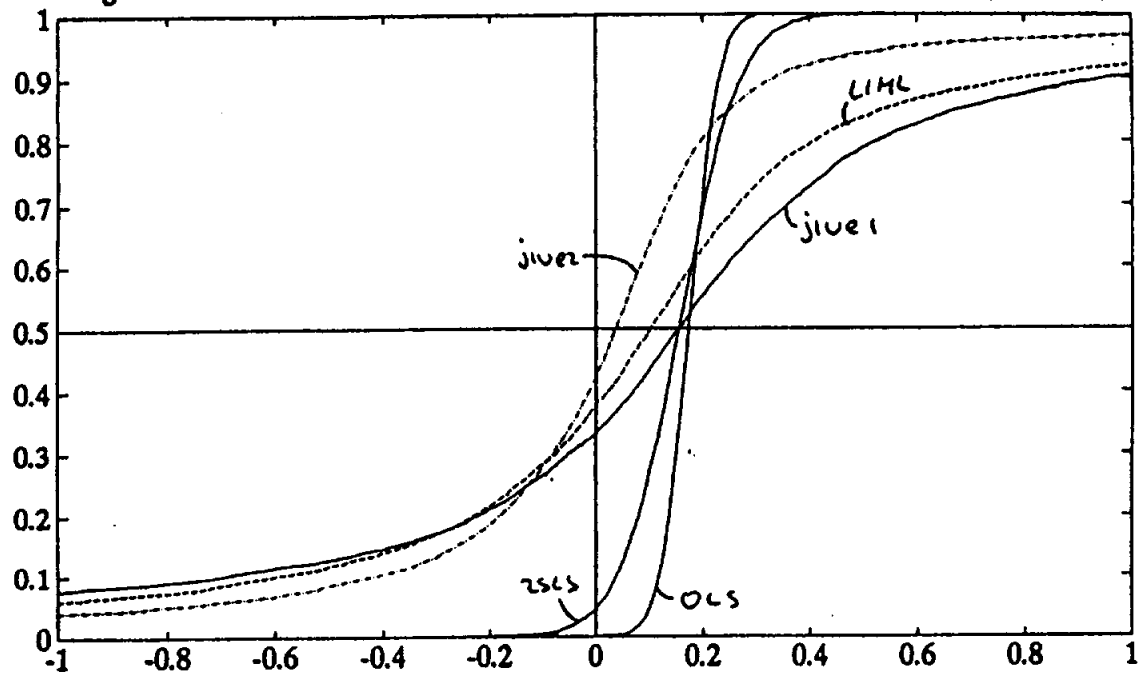

Figure 4: Distribution Functions for OLS, 2SLS, LIML, JIVE1, JIVE2 (Model 4)