Improved Jive Estimators for Overidentified Linear Models with and without Heteroskedasticity

Daniel A. Ackerberg
UCLA

Paul J. Devereux
University College Dublin, CEPR and IZA

October 8, 2008

Abstract

We introduce two simple new variants of the Jackknife Instrumental Variables (JIVE) estimator for overidentified linear models and show that they are superior to the existing JIVE estimator, significantly improving on its small sample bias properties. We also compare our new estimators to existing Nagar (1959) type estimators. We show that, in models with heteroskedasticity, our estimators have superior properties to both the Nagar estimator and the related B2SLS estimator suggested in Donald and Newey (2001). These theoretical results are verified in a set of Monte-Carlo experiments and then applied to estimating the returns to schooling using actual data.

Econlit Subject Descriptors: C310, J240

1 Introduction[1]

It is well known (see, e.g. Staiger and Stock (1997)) that in overidentified models first stage overfitting can generate small-sample bias in the Two Stage Least Squares (2SLS) estimator. We analyze the Jackknife Instrumental Variables (JIVE) estimator that has been proposed to address this overfitting problem by Phillips and Hale (1977), Angrist, Imbens and Krueger (1995, 1999), and Blomquist and Dahlberg (1999). While the small sample bias of JIVE does not depend on the degree of overidentification, it does increase in the number of *included exogenous variables* in the second stage equation. This number can be quite large in empirical analysis, e.g., in Angrist and Krueger's (1991) study of the returns to education, there are up to 60 such variables. We suggest two very simple but significant improvements to the JIVE estimator that eliminate this bias term, reducing both its small sample bias and variability. We call these new estimators the Improved JIVE (IJIVE) estimator and the Unbiased IJIVE (UIJIVE) estimator.

We then compare our IJIVE and UIJIVE estimators to Nagar's (1959) bias corrected estimator for overidentified models. Nagar's estimator has recently been investigated by Hahn and Hausman (2002) and Donald and Newey (2001). Interestingly, we show the IJIVE and Nagar estimators are similar in spirit and have very similar properties under homoskedasticity. However, we show that the IJIVE estimator has superior properties to Nagar's estimator when the residuals are heteroskedastic. In particular, the IJIVE and UIJIVE estimators (as well as the original JIVE estimator) are consistent under many instrument asymptotics, while the Nagar estimator (as well as Donald and Newey's related Bias-Adjusted 2SLS (B2SLS) estimator) is not. As Chao and Swanson (2004) have recently shown that the LIML estimator is not consistent under many instrument asymptotics with heteroskedasticity, our estimators also have better properties than LIML with heteroskedasticity.[2]

We report two sets of Monte-Carlo experiments that verify our theoretical results. In the first set, we show that our IJIVE and UIJIVE estimators clearly dominate the JIVE estimator, particularly when

there are many covariates in the system. In the second set of experiments, we compare our IJIVE and UIJIVE estimators to, among others, the Nagar, B2SLS and LIML estimators. As expected, we find that the IJIVE and UIJIVE estimators are superior in the presence of heteroskedasticity. Finally, we apply our new estimators to the Angrist and Krueger (1991) returns to schooling specification and find reasonable differences between our estimators and the standard JIVE estimator.

2 The JIVE Estimator

Consider the following simultaneous equations model:

$$\begin{aligned} Y &= X^*\beta^* + W\gamma^* + \epsilon_i \\ X^* &= Z^*\pi^* + W\delta^* + \eta_i. \end{aligned} \tag{1}$$

The endogenous variable Y is an N by 1 vector, X^* is an N by L_1 matrix of endogenous explanatory variables, W is an N by L_2 matrix of exogenous variables, and Z^* is an N by K_1 matrix of exogenous instruments that are excluded from the main equation. β^* is L_1 by 1, γ^* is L_2 by 1, π^* is K_1 by L_1, and δ^* is L_2 by L_1. We assume $K_1 \geq L_1$. Let $L = L_1 + L_2$ and $K = K_1 + L_2$. The number of overidentifying restrictions in this model is $K_1 - L_1 = K - L$.

Define the N by L matrix $X = \left[X^* \; W\right]$, the N by K matrix $Z = \left[Z^* \; W\right]$, $\beta = \begin{bmatrix} \beta^* \\ \gamma^* \end{bmatrix}$, and

$\pi = \begin{bmatrix} \pi^* & 0_{K_1 x L_2} \\ \delta^* & I_{L_2} \end{bmatrix}$. We can now write our model as

$$\begin{aligned} Y &= X\beta + \epsilon \\ X &= Z\pi + \eta, \end{aligned}$$

where β is an L vector, π is a K by L matrix, ϵ is an N vector, and η is an N by L matrix. We assume ϵ and η are independent across i and mean independent of W and Z. We also assume initially that ϵ and η are homoskedastic with $L+1$ by $L+1$ variance matrix $\Sigma_{\epsilon\eta}$. This homoskedasticity assumption is relaxed in section 3. We denote the probability limits of $Z'Z/N$ and $X'X/N$ as Σ_z and Σ_x respectively.

The 2SLS estimator is $(X'P_zX)^{-1}(X'P_zY)$ where $P_z = Z(Z'Z)^{-1}Z'$. While β_{2SLS} is consistent as N goes to infinity, it is now well known (see Nagar (1959), Phillips and Hale (1977), Staiger and Stock (1997), and others) that it has poor finite sample properties when there are many instruments Z^* relative to the dimension of X^*.[3] This bias is caused by overfitting in the first stage - with a large number of instruments, $X'P_z$ approaches X' and the 2SLS estimator approaches the biased OLS estimator. A first order approximation to this bias (to order $1/N$) given in Angrist, Imbens, and Krueger (1995) is $(K-L-1)(\pi'\Sigma_z\pi)^{-1}\sigma_{\epsilon\eta}/N$, where $\sigma_{\epsilon\eta}$ is an L vector of the covariances between ϵ and each of the L elements of η.[4]

The JIVE estimator of Phillips and Hale (1977, henceforth PH), Angrist, Imbens, and Krueger (1995, 1999, henceforth AIK), and Blomquist and Dahlberg (1999, henceforth BD) works as follows: Let $Z(i)$ and $X(i)$ denote matrices equal to Z and X with the ith row removed. Define $\widehat{X}_{JIVE}$ to be the N x L dimensional matrix with ith row equal to $Z_i\pi(i)$, where $\pi(i) = (Z(i)'Z(i))^{-1}(Z(i)'X(i))$. The JIVE

estimator is[5]

$$\beta_{JIVE} = (\widehat{X}'_{JIVE}X)^{-1}(\widehat{X}'_{JIVE}Y). \tag{2}$$

Note the intuition behind the JIVE estimator: In forming the "predicted value" of X for observation i, one uses a π coefficient estimated on all observations other than i. This eliminates the overfitting problems in the first stage. The JIVE estimator can be written in very compact form that doesn't require iterating over observations to compute. Following PH and defining $D_{P_z} = diag(P_z)$ and $C_{JIVE} = (I - D_{P_z})^{-1}(P_z - D_{P_z})$, we can write the JIVE estimator as[6]

$$\beta_{JIVE} = (X^{'}C'_{JIVE}X)^{-1}(X^{'}C'_{JIVE}Y). \tag{3}$$

2.1 The Improved JIVE (IJIVE) Estimator

PH and AIK use Edgeworth expansions show that the small sample bias (to order $1/N$) of JIVE is approximately equal to

$$(-L_1 - L_2 - 1)(\pi'\Sigma_z\pi)^{-1}\sigma_{\epsilon\eta}/N \tag{4}$$

and is generally less than that of 2SLS. As L_1 is the number of included *endogenous* variables in the second stage equation and L_2 is the number of included *exogenous* variables (including the constant term) in the second stage equation, in most applications the L_2 term will be the primary source of small sample bias in the JIVE estimator.

Somewhat surprisingly, it turns out that one can eliminate this L_2 bias term by simply partialing out W (including the constant term) from Y, X, and Z *before* implementing the JIVE estimator.[7] We denote the partialled out JIVE estimator as the IJIVE (Improved JIVE) estimator. More precisely, with $\widetilde{Y}$, $\widetilde{X}$,

and $\widetilde{Z}$ now representing variables partialled out with respect to W (e.g. $\widetilde{Z} = M_w Z$), we have

$$\beta^*_{IJIVE} = (\widetilde{X}' C'_{IJIVE} \widetilde{X})^{-1} (\widetilde{X}' C'_{IJIVE} \widetilde{Y}), \tag{5}$$

where

$$C_{IJIVE} = (I - D_{\widetilde{P}_Z})^{-1} (\widetilde{P_Z} - D_{\widetilde{P}_Z}) \tag{6}$$

and $\widetilde{P_Z} = \widetilde{Z}(\widetilde{Z}'\widetilde{Z})^{-1}\widetilde{Z}'$. In Appendix 1, we show using an Edgeworth expansion that β^*_{IJIVE} has a small sample bias of $(-L_1 - 1)(\pi' \Sigma_{\widetilde{Z}} \pi)^{-1} \sigma_{\epsilon\eta}/N$, i.e. our IJIVE estimator eliminates the L_2 bias term.[8]

This bias term proportional to $(-L_1 - 1)$ can compare very favorably to the bias of the original JIVE estimator proportional to $(-L_1 - L_2 - 1) = (L - 1)$. In the median application, $L_1 = 1$, whereas L is at least 2 (constant term plus endogenous variable) and often much larger. For example, in Angrist and Krueger's (1991) study, $L_1 = 1$ and $L = 61$ in specifications with year and state controls. While small sample variability is hard to explicitly calculate, our monte carlo results suggest that β^*_{IJIVE} has significantly lower variance than β^*_{JIVE}, making it even more attractive.

Initial partialling out of W removes the L_2 term from the bias. We now address the $-L_1 - 1$ term. Define

$$C_{UIJIVE} = (I - D_{\widetilde{P}_Z} + \omega I)^{-1} (\widetilde{P_Z} - D_{\widetilde{P}_Z} + \omega I),$$

where $\omega = (L_1 + 1)/N$. In the appendix, we show that

$$\beta^*_{UIJIVE} = (\widetilde{X}' C'_{UIJIVE} \widetilde{X})^{-1} (\widetilde{X}' C'_{UIJIVE} \widetilde{Y})$$

is approximately unbiased (to order $1/N$). Intuitively, the additional ωI terms change the trace of C, subtracting out the bias in the IJIVE estimator[9]. In our Monte Carlo results, we also note that β^*_{UIJIVE}

tends to have lower dispersion than both β^*_{JIVE} and β^*_{IJIVE}.

2.2 The IJIVE Estimator and Nagar's Estimator

The Nagar (1959) estimator can be written as

$$\beta^*_{Nagar} = (\widetilde{X}' C'_{Nagar} \widetilde{X})^{-1} (\widetilde{X}' C'_{Nagar} \widetilde{Y}), \tag{7}$$

where $C_{Nagar} = \widetilde{P_Z} - \lambda \widetilde{M_Z}$, $\lambda = (K_1/(N-K_1))$, and $\widetilde{M_Z} = I - \widetilde{P_Z}$. Defining $\widehat{\lambda} = K_1/N$ gives

$$C_{Nagar} = (1-\widehat{\lambda})^{-1} (\widetilde{P_Z} - \widehat{\lambda} I). \tag{8}$$

This $\widehat{\lambda}$ formulation of the Nagar estimator is convenient in that it allows easy comparison to the IJIVE estimator. Note the similarities between (6) and (8): Since the trace of $\widetilde{P_Z}$ is equal to K_1, the average value of the diagonal elements of $\widetilde{P_Z}$ is K_1/N. As such, the Nagar and IJIVE estimators differ only in that in IJIVE the actual diagonal elements of $\widetilde{P_Z}$ are subtracted from $\widetilde{P_Z}$ while, in Nagar, the average value of the diagonal elements of $\widetilde{P_Z}$ is subtracted from $\widetilde{P_Z}$. Likewise, in the denominator, the IJIVE estimator subtracts the actual value of the diagonal of $\widetilde{P_Z}$ while Nagar subtracts the average value of the diagonal elements of $\widetilde{P_Z}$. Under homoskedasticity, the Nagar estimator has the same approximate small sample bias as IJIVE, i.e. $(-L_1 - 1)(\pi' \Sigma_{\widetilde{Z}} \pi)^{-1} \sigma_{\epsilon\eta}/N$.

Donald and Newey (2001) also suggest a variant of the Nagar estimator, which they term the Bias-Adjusted 2SLS (B2SLS) estimator. This estimator is identical to the Nagar estimator except that $\widehat{\lambda} = (K_1 - (L_1 + 1))/N$. This adjustment is analagous to the adjustment of IJIVE to UIJIVE, and it reduces the approximate small sample bias of the Nagar estimator to zero.

3 Many Instruments Asymptotics under Heteroskedasticity

While the IJIVE (UIJIVE) and Nagar (B2SLS) estimators have very similar properties under homoskedasticity, they diverge under heteroskedasticity. The heteroskedasticity we consider is in the instruments Z, i.e. we allow heteroskedasticity of ϵ and η in Z. Changing notation slightly, we now assume that the exogenous variables (W) have already been partialled out of the model so we have

$$Y = X\beta + \epsilon \tag{9}$$

$$X = Z\pi + \eta. \tag{10}$$

In this formulation, Y is an N vector, X is an $N \times L_1$ matrix, Z is an $N \times K_1$ matrix, ϵ is an N vector, and η is an $N \times L_1$ matrix. To keep notation simple, in this section we assume X is one dimensional, i.e. $L_1 = 1$. The results can easily be generalized.

Small sample bias calculations like those above are difficult under heteroskedasticity. As a result, we turn to many-instrument asymptotics (also called group asymptotics) previously used by Bekker (1994), Angrist and Krueger (1995), AIK, Chao and Swanson (2005), and Newey (2004)[10]. The basic idea of many-instrument asymptotics is to allow the number of instruments K_1 to go to infinity at the same rate as the number of observations. More specifically, we assume (as Newey (2004)) that $K_1/N \to \alpha$ ($0 < \alpha < 1$) as $N \to \infty$. This is intended to approximate a situation where the number of instruments is relatively large and overfitting might be problematic.

The estimators under consideration can all be written as

$$\begin{aligned}\widehat{\beta} &= (X'C'X)^{-1}(X'C'Y) \\ &= (\frac{1}{N}X'C'X)^{-1}(\frac{1}{N}X'C'X)\beta + (\frac{1}{N}X'C'X)^{-1}\frac{1}{N}X'C'\epsilon\end{aligned}$$

with the appropriate C matrix. We simply assume that under the asymptotic sequence studied, the quantities $(1/N)X'C'X$ and $(1/N)X'C'\epsilon$ converge in probability to the limit of their (assumed finite) expectations. Clearly, these very high-level assumptions put restrictions on what is happening to both the Z matrix and the heteroskedasticity as $K_1 \to \infty$. It would be preferable to develop more primitive, lower level, assumptions on the processes that generate these higher level assumptions. Chao and Swanson (2004) do exactly this, thus providing a deeper, more complete, proof of the consistency of JIVE under many-instrument asymptotics with heteroskedasticity (and guaranteeing that at least some such sequences do satisfy our high-level assumptions). However, for our purposes, starting with these high level assumptions is sufficient to demonstrate the fundamental and important differences between IJIVE (UIJIVE) and Nagar (B2SLS) under heteroskedasticity.[11]

Under these assumptions, it follows that

$$p\lim(\widehat{\beta}) - \beta = \frac{p\lim\left[X'C'\epsilon/N\right]}{p\lim\left[X'C'X/N\right]} = \frac{\lim_{N,K_1\to\infty} E\left[X'C'\epsilon/N\right]}{\lim_{N,K_1\to\infty} E\left[X'C'X/N\right]} \tag{11}$$

so consistency of $\widehat{\beta}$ depends on the behavior of $E\left[X'C'\epsilon/N\right]$ as N and K_1 increase. To analyze this behavior, first note that

$$X'C'_{Nagar} = X'(P_z - \lambda M_z) = (Z\pi + \eta)'(P_z - \lambda M_z) = \pi' Z' + \eta' C'_{Nagar} \tag{12}$$

and

$$\begin{aligned} X'C'_{IJIVE} &= X'(P_z - D_{P_z})(I - D_{P_z})^{-1} \\ &= (Z\pi + \eta)'(P_z - D_{P_z})(I - D_{P_z})^{-1} \\ &= \pi' Z'(I - D_{P_z})(I - D_{P_z})^{-1} + \eta' C'_{IJIVE} \\ &= \pi' Z' + \eta' C'_{IJIVE}. \end{aligned} \tag{13}$$

Thus, with C representing either C_{Nagar} or C_{IJIVE}, the numerator of (11) is

$$E\left[X'C'\epsilon/N\right] = E\left[\pi' Z'\epsilon/N\right] + E\left[\eta' C'\epsilon/N\right] \tag{14}$$

$$= E\left[\eta' C'\epsilon/N\right] \tag{15}$$

since Z and ϵ are uncorrelated. Therefore, consistency under many instrument asymptotics relies on the term

$$E\left[\eta' C'\epsilon/N\right] = \frac{1}{N}E(trace(\eta' C'\epsilon)) = \frac{1}{N}Etrace(C'\epsilon\eta') = \frac{1}{N}trace(C'E(\epsilon\eta')) \tag{16}$$

either equaling zero or vanishing asymptotically.

First note that with homoskedasticity, (16) is zero for both C_{IJIVE} and C_{NAGAR}. For C_{IJIVE}, this is because 1) C_{IJIVE} has a zero diagonal and 2) $E(\epsilon\eta')$ is a diagonal matrix due to independence across observations. Thus, the diagonal of $C'E(\epsilon\eta')$ is identically zero. For C_{NAGAR}, since the diagonal

elements of $E(\epsilon\eta')$ all equal $\sigma_{\epsilon\eta}$ we have

$$\begin{aligned}\frac{1}{N}trace(C'_{NAGAR}E(\epsilon\eta')) &= \frac{1}{N}trace((P_z - \lambda M_z)E(\epsilon\eta')) \\ &= \frac{1}{N}\sigma_{\epsilon\eta}\left[trace(P_z) - \frac{K_1}{N-K_1}trace(M_z)\right] \\ &= \frac{1}{N}\sigma_{\epsilon\eta}\left[K_1 - \frac{K_1}{N-K_1}(N-K_1)\right] \\ &= 0.\end{aligned} \tag{17}$$

Thus, under homoskedasticity, both IJIVE and Nagar are consistent under many instruments asymptotics.

Things are different under heteroskedasticity. While $\Omega_{\epsilon\eta} = E(\epsilon\eta')$ is still diagonal, the elements on the diagonal are now generally unequal and functions of Z_i. However, since C_{IJIVE} has a zero diagonal, we still have

$$\frac{1}{N}trace(C'_{IJIVE}E(\epsilon\eta')) = 0. \tag{18}$$

In contrast

$$\begin{aligned}\frac{1}{N}trace(C'_{NAGAR}E(\epsilon\eta')) &= \frac{1}{N}trace((P_z - \lambda M_z)\Omega_{\epsilon\eta}) \\ &= \frac{1}{N}\sum_i (P_{z,ii} - \frac{K_1}{N-K_1}M_{z,ii})\sigma_{\epsilon_i\eta_i} \\ &= \frac{1}{N}\sum_i \left[(\frac{N}{N-K_1})P_{z,ii} - \frac{K_1}{N-K_1}\right]\sigma_{\epsilon_i\eta_i} \\ &= \frac{1}{N}\sum_i (\frac{N}{N-K_1})\widetilde{P}_{z,ii}\sigma_{\epsilon_i\eta_i},\end{aligned} \tag{19}$$

where $\sigma_{\epsilon_i\eta_i}$ is the ith diagonal element of $\Omega_{\epsilon\eta}$, i.e. the covariance between ϵ_i and η_i. $P_{z,ii}$ is the ith diagonal element of P_z, $\widetilde{P}_{z,ii}$ is $P_{z,ii} - (1/N)\sum_i P_{z,ii}$, and the last line relies on the fact that $\frac{1}{N}\sum_i P_{z,ii} = Trace(P_z)/N = K_1/N$. This term is *not* generally zero and does not disappear as N

and K_1 increase. The size of this term depends on how the covariance term is "correlated" with deviations from average $P_{z,ii}$. Since both $\sigma_{\epsilon_i \eta_i}$ and $\widetilde{P}_{z,ii}$ are functions of Z_i, we expect such correlation generally. Thus, unlike the IJIVE estimator, the Nagar estimator is generally not consistent under many instruments asymptotics with heteroskedasticity. The intuition behind this result follows directly from the discussion in the prior section. Unlike IJIVE, which subtracts off the exact diagonal elements of P_z, the Nagar estimator subtracts off the expectation of this diagonal and the difference between the expectation and realized value will generally be correlated with $\sigma_{\epsilon_i \eta_i}$.[12]

One should consult the working paper version for details, but it is fairly easy to show that we get the same theoretical results with the UIJIVE and B2SLS estimators - UIJIVE is consistent under many-instrument asymptotics, while B2SLS is not. The intuition behind this extension is that UIJIVE and B2SLS only differ from their respective IJIVE and Nagar counterparts by terms that are of order $(L_1 + 1)/N$, which disappear under many-instrument asymptotics. In summary, we have shown that JIVE type estimators have superior properties to Nagar-type estimators under heteroskedasticity. These superior properties will be very evident in our Monte-Carlo experiments.

4 Monte Carlo Analysis

We perform two sets of Monte Carlo analyses. The first set compares our IJIVE and UIJIVE estimators to the JIVE estimator. We focus in particular on what happens as the number of exogenous explanatory variables W increases, as this influences the small sample improvements of our estimators. As expected, the IJIVE and UIJIVE estimators perform considerably better than JIVE.

The second set of results compares the IJIVE and UIJIVE estimators to other existing small sample instrumental variables estimators under heteroskedasticity. Again, as expected, we find that IJIVE and

UIJIVE perform very well in comparison to the other estimators (in particular with respect to Nagar, B2SLS, and LIML), and argue that in many situations, these may be the estimators of choice.

4.1 Base Model

Our base model assumes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 W_i + \epsilon_i \\ X_i &= \pi_0 + \pi_1 Z_i + \pi_2 W_i + \eta_i, \end{aligned}$$

where X_i is a scalar endogenous variable and W_i is a set of included exogenous variables that are distributed i.i.d. $N(0, 1)$. We vary the dimension of W_i in our experiments. In all cases, we set $\beta_0 = \pi_0 = 0$, $\beta_1 = 1$, and all the elements of β_2 and π_2 equal to 1.

For simplicity, our instruments Z_i are a set of mutually exclusive dummy variables.[13] One can think of these dummy variables as representing groups, similar to, e.g., Angrist and Krueger's (1991) returns to education specifications where the instruments are groups defined by quarter of birth interacted with state of birth. With our initial sample size of 100, we assume there are 20 such groups - five observations in each group. Hence, Z_i is 19 dimensional (one of the groups is captured by the constant term). The parameters on these instruments π_1 are distributed i.i.d. $N(0, 0.1)$ and are redrawn across experimental draws to integrate over a range of potential first stage models. Note that the variance of this π_1 vector determines the strength of the instruments. At our setting of 0.1, the instruments are relatively weak, but not weak enough to preclude meaningful inference.[14] We assume initially that the errors are

homoskedastic and have the distribution

$$\begin{pmatrix} \epsilon_i \\ \\ \eta_i \end{pmatrix} \mid Z_i \sim N\left(\begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25 & 0.2 \\ \\ 0.2 & 0.25 \end{pmatrix}\right). \tag{20}$$

We perform 10000 Monte-Carlo replications. In all tables we report quantiles (10%, 25%, 50%, 75%, 90%) of the distribution of the estimator of β_1 around the true β_1. The 50% quantile is thus the median bias of the estimator. We also report the median absolute error of the estimator. Mean biases and mean squared errors of our estimators are problematic because JIVE and Nagar type estimators are known not to have second moments. This makes their means extremely sensitive to outliers and makes mean squared errors meaningless. To address this issue we trimmed the distributions of all the estimators (at the 5th and 95th percentiles) and report mean bias and mean absolute error for these trimmed distributions. For potential 90% confidence intervals, we report both infeasible "true" coverage rates[15] and feasible coverage rates using standard asymptotic approximations. For the homoskedastic case, we simply follow AIK and use the following asymptotic variance for the JIVE estimators[16]:

$$\widehat{\sigma}^2 \left[X'CX \left(X'C'CX \right)^{-1} X'C'X \right]^{-1},$$

where C is either C_{JIVE}, C_{IJIVE}, or C_{UIJIVE}.

4.2 Experiment 1

In our first set of experiments, we examine the performance of the various JIVE estimators as the dimension of the exogenous variables increases. The first panel of Table 1 contains results when dim(W)=0, i.e. when the only non-excluded exogenous variable is the constant term. The large median and mean

biases of the OLS and 2SLS estimates suggest that there is both a significant endogeneity problem and a significant overfitting problem. Confidence interval coverage is very poor for these estimators. With only one non-excluded exogenous variable, we would expect the JIVE estimator to do quite well at reducing median and trimmed mean bias, and it does. However, even with only the constant term as an included exogenous variable, the IJIVE and UIJIVE estimators appear to perform better. Both the means and the medians of IJIVE are about 0.03 closer to the truth than JIVE and the UIJIVE estimator is approximately mean unbiased (although in terms of median bias, it is similar to JIVE). The distributions of the IJIVE and UIJIVE estimators are also slightly tighter than the JIVE estimator. Also note that feasible confidence interval coverage for the IJIVE and UIJIVE estimators (as well as the JIVE estimator) are reasonable.

Moving through the panels of Table 1 corresponds to adding more exogenous variables to the system. The basic trend is that the JIVE's performance quickly deteriorates while the IJIVE and UIJIVE estimators continue to perform well. By the last panel, where dim(W) is equal to 10, the JIVE estimator has considerable bias - median bias is -0.31 and the mean bias is -0.46. This bias is even larger than that of 2SLS. In addition, the variance of the JIVE estimator increases tremendously. In contrast, the IJIVE and UIJIVE estimators continue to perform well, both in terms of bias and variance. It is interesting to compare the IJIVE and UIJIVE estimators. While IJIVE does slightly better at median bias, UIJIVE does slightly better in terms of trimmed mean bias. UIJIVE also tends to perform a bit better in the variance measures (Median and Mean Absolute Error). On the other hand, UIJIVE's confidence interval coverage is a bit worse.[17] However, the differences between IJIVE and UIJIVE are small compared to the differences between JIVE and IJIVE/UIJIVE.

For more Monte Carlo results in the homoskedastic case, one can also consult our comment (Ackerberg and Devereux (2006)) on Davidson and MacKinnon (2006 (DM)). DM show reasonably large bias

and variance advantages of the LIML estimator (see below) over the original JIVE estimator. In our comment, we show that across a wide range of specifications, a large portion of these bias and variance advantages disappear when using the IJIVE and UIJIVE estimators introduced in this paper. This is particularly interesting because in the DM experiment the only included exogenous variable is a constant term. The difference between JIVE and IJIVE/UIJIVE would be even greater with a higher dimensional W.

4.3 Experiment 2

In our second set of experiments, we fix $\dim(W) = 0$ and add heteroskedasticity to the model to compare the performance of IJIVE and UIJIVE to the Nagar and B2SLS estimators. For comparision purposes, we also consider three other estimators, the LIML estimator, the Jackknife 2SLS estimator of Hahn, Hausman, and Kursteiner (2002) (HHK), and the Pseudo Empirical Likelihood (PEL) estimator of Guggenberger and Hahn (2005) (GH).

The LIML estimator can be conveniently written as

$$\beta_{LIML} = (X'C'_{LIML}X)^{-1}(X'C'_{LIML}Y), \tag{21}$$

where

$$C_{LIML} = (I - \lambda M_z), \tag{22}$$

and λ is the smallest characteristic root of $\left([Y\ X]'\,P_z\,[Y\ X]\right)\left([Y\ X]'\,M_z\,[Y\ X]\right)^{-1}$. The LIML estimator has known optimality properties under correct specification, but this requires i.i.d. normality and linearity. Chao and Swanson (2004) show that LIML is not consistent under many instruments asymptotics with heteroskedasticity.

The Jackknife 2SLS estimator suggested by HHK performs a jackknife bias correction on the 2SLS estimator. This involves estimating the 2SLS model $N+1$ times – once on the full sample, and once on each subsample of $N-1$. Assuming that the bias is linear in $1/N$, a linear combination of the full sample estimator and the average of the $N-1$ sample estimators produces an unbiased estimate. This is

$$\beta_{J2SLS} = N\beta_{2SLS} - (N-1)\frac{1}{N}\sum_{n}\beta_{2SLS-n},$$

where β_{2SLS-n} is the 2SLS estimate on the dataset without observation n. HHK show that relative to Nagar (and implicitly JIVE) type estimators, β_{J2SLS} has considerably less variance, more bias, and lower mean squared error.

The PEL estimator introduced by GH is an analytic version of the Generalized Empirical Likelihood (GEL) class of estimators. A number of papers have demonstrated attractive theoretical properties of this class of estimators, even under heteroskedasticity, e.g. Qin and Lawless (1994), Kitamura and Sturtzer (1997), Imbens, Spady, and Johnson (1998), and Newey and Smith (2004). The primary problem with GEL estimators is that they do not have a closed form solution, requiring iterative techniques to minimize an objective function, making them more time consuming and complicated to program and run, as well as making them potentially sensitive to optimization issues such as local extrema. They are particularly hard to study in Monte-Carlo experiments because of this. GH introduce an estimator that has the same third order bias properties as GEL estimators, but that has a closed form solution (see GH for this closed form). This is the estimator that we consider in our experiments, with the caveat that it is conceivable that the non-analytic GEL estimators might perform better in small samples.

To construct feasible confidence intervals with heteroskedasticity, we use a White-adjusted asymp-

totic variance formula, i.e.

$$Var(\widehat{\beta}) = (X'C'X)^{-1}\left(\sum_{i=1}^{N}\widehat{e}_i^2\,[CX]_i\,[CX]_i^{'}\right)(X'C'X)^{-1'},$$

where $\widehat{e}_i$ are the estimated residuals and $[CX]_i$ is the ith row of CX. This formula is used for 2SLS, Nagar, B2SLS, IJIVE, UIJIVE, and LIML with the appropriate C matrix.[18] There are no available asymptotic approximations for J2SLS and PEL under heteroskedasticity, so we simply report the infeasible CIs.

We introduce heteroskedasticity by allowing the variance matrix in (20) to differ across groups. We also change the group sizes - specifically, we assume that there are 2 groups of 23 observations and 18 groups of 3 observations (in total there are still 20 groups and 100 observations). This is important because in the special case where group sizes are identical, $diag(P_z)$ is constant and the Nagar and B2SLS *are* consistent under many-instrument asymptotics even with heteroskedasticity. For the heteroskedasticity, we allow the variance matrix (20) to differ across the two types of groups (large (23) and small (3)).

Table 2 presents results from these heteroskedastic models. In the first panel, results for the model without heteroskedasticity are presented. Of interest here is how Nagar, B2SLS, J2SLS, and LIML perform compared to IJIVE and UIJIVE. As expected, without heteroskedasticity the Nagar and B2SLS estimators look almost identical to their IJIVE and UIJIVE counterparts. J2SLS, as expected, has considerably lower spread and lower mean and median absolute error, but worse bias and confidence interval coverage. LIML performs quite well, with almost no bias and the lowest mean and median absolute errors. Surprisingly, PEL performs relatively poorly with high mean and median bias. On the other hand, this is consistent with some of the monte-carlo results in GH. Again, it is possible that the non-analytic

GEL estimators might do better than the PEL in small samples.

The next four panels of the table show the results with heteroskedastic errors. In all cases, we hold the variances in (20) constant at 0.25, only changing the covariances.[19] In panel B, we set the covariance to 0 for the 2 large groups, keeping it at 0.2 for the 18 small groups. One can see the heteroskedasticity dramatically impact the performance of Nagar, B2SLS, and LIML. All perform very poorly with median and trimmed mean biases in excess of 0.2, almost as high as 2SLS. They also have poor confidence interval coverage, lower than 70% for 90% confidence intervals. In contrast, the IJIVE and UIJIVE estimators continue to have very small biases, reasonably low values of mean and median absolute error, and good confidence interval coverage. The J2SLS estimator also seems to be unaffected by the heteroskedasticity, continuing to have higher mean and median biases than IJIVE and UIJIVE (and worse confidence interval converage), but lower dispersion. PEL also does not appear to be particularly affected by the heteroskedasticity, although its performance was not good to start with.

In panel C we reverse the heteroskedasticity, making the covariance 0 for the 18 small groups and 0.2 for the 2 large groups. In this example, there is less of an overfitting problem, evidenced by the fact that standard 2SLS starts to perform very well. This probably occurs because for the small groups (where the overfitting problem is most severe) there is no endogeneity problem because of the zero covariance. Even though 2SLS now performs quite well, Nagar, B2SLS, and LIML are still seriously biased, almost as bad as OLS. Again, both of our JIVE estimators perform quite well in comparison.

Panels D and E perform two perturbations of these experiments starting from the setup in panel B. In panel D, we simply weaken the degree of heteroskedasticity, setting the covariance to 0.1 for the two large groups. We still see quite sizeable biases in the Nagar, B2SLS, and LIML estimators. In panel E we try to approximate our many-instruments asymptotic arguments by simultaneously increasing both the number of observations (to 500) and the number of instruments (to 100). As suggested by the many

instruments asymptotics, the small biases of IJIVE and UIJIVE quickly disappear while their mean and median absolute errors become very small. In contrast, the biases in Nagar, B2SLS, LIML remain large and their coverage rates become extremely poor. In sum, our Monte-Carlo results confirm our theoretical preditions that the IJIVE and UIJIVE estimators are considerably more robust to heteroskedasticity than are Nagar, B2SLS, and LIML.

5 Application to Return to Education

In a paper that motivated much of the recent literature on overidentified models, Angrist and Krueger (1991) estimate the return to schooling using quarter of birth as an instrument in a sample of 329,500 men born between 1930-39 (from the 1980 Census). We estimate two of their specifications. In the first specification, there are 30 instruments created by interacting quarter and year of birth and the control variables are a set of year indicators (so $K = 30$ and $L = 11$). The second specification contains 180 instruments constructed by adding interactions of 50 state and quarter of birth dummies to the original 30 instruments. In this second specification, both state and year fixed effects are included as controls (so $K = 180$ and $L = 61$).

In Table 3, we report estimates for the two specifications along with asymptotic standard errors. The OLS, 2SLS, LIML, and JIVE coefficients and standard errors are exactly the same as those reported in AIK (1999). Comparing the JIVE estimates to IJIVE and UIJIVE we see that, as expected, the IJIVE and UIJVE estimates are smaller and have lower standard errors. This is particularly the case in the second specification, where L is quite high. These results are again consistent with 1) our theory and Monte Carlo evidence showing that JIVE is biased away from OLS and 2) our Monte Carlo evidence suggesting that JIVE has higher variance than IJIVE and UIJIVE. LIML seems to perform adequately

here, and Nagar and B2SLS generate point estimates and standard errors almost identical to IJIVE and UIJIVE. This suggests that in these data there is not a significant heteroskedasticity problem affecting the performance of Nagar, B2SLS, and LIML. While heteroskedasticity does not seem to be an issue in this particular application, our theoretical and Monte Carlo results definitely suggest that for robustness, one should prefer the UIJIVE and JIVE estimators.[20]

6 Conclusions

In this paper, we have suggested two simple but significant improvements to the JIVE estimator that reduce both its small sample bias and variability. These estimators are similar to Nagar's estimator when errors are homoskedastic but have superior theoretical properties to Nagar's estimator with heteroskedastic errors. In particular, we show that the IJIVE estimator (as well as the UIJIVE estimator) is consistent under many instruments asymptotics, while the Nagar estimator (as well as Donald and Newey's related B2SLS estimator) is not. We verify these theoretical results with two sets of Monte Carlo experiments. The first shows that our IJIVE and UIJIVE estimators clearly dominate the JIVE estimator, particularly when there are many exogenous variables in the system. The second shows that the IJIVE and UIJIVE estimators are superior to Nagar and B2SLS in the presence of heteroskedasticity. We also compare our estimators to three other estimators that have been advocated in the literature; the Limited Information Maximum Likelihood (LIML) estimator, the Jackknife Two Stage Least Squares (J2SLS) estimator, and the PEL estimator. We find that LIML is also sensitive to heteroskedasticity. While the J2SLS estimator typically has considerably lower variance and lower mean squared error than our estimators, it also typically has more bias. This combination of lower variance and higher bias can generate poor confidence interval coverage. Our work suggests that the estimators of choice when one is worried about robustness

to heteroskedasticity in a situation with overidentification depends on one's goals. If one is interested in minimizing mean squared error, J2SLS may be appropriate. If one is concerned with limiting bias, hypothesis testing, and confidence interval coverage, IJIVE/UIJIVE may be the estimators of choice.

References

[1] Ackerberg Daniel A. and Paul J. Devereux , "Improved JIVE estimators for Overidentified Linear Models with and without Heteroskedasticity," mimeo (2003).

[2] Ackerberg Daniel A. and Paul J. Devereux , "Comment on 'The Case Against JIVE," *Journal of Applied Econometrics* 21:6, September/October (2006), 835-838.

[3] Angrist Joshua D. and Alan B. Krueger, "Split sample instrumental variables estimates of the return to schooling," *Journal of Business and Economic Statistics* 13:2, April (1995), 225-35.

[4] Angrist, Joshua D. and Alan B. Krueger, "Does compulsory school attendance affect schooling and earnings?," *Quarterly Journal of Economics* 106:4, November (1991), 979-1014.

[5] Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger, "Jackknife Instrumental Variables Estimation," NBER technical working paper 172 (1995).

[6] Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger, "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics* 14:1, January/February (1999), 57-67.

[7] Bekker, Paul A., "Alternative approximations to the distributions of instrumental variables estimators," *Econometrica*, 62:3, May (1994), 657-682.

[8] Blomquist Soren and Matz Dahlberg, “Small sample properties of LIML and jackknife IV estimators: experiments with weak instruments,” *Journal of Applied Econometrics* 14:1, January/February (1999), 69-88.

[9] Chao, John C. and Norman R. Swanson, “Consistent estimation with a large number of weak instruments,” *Econometrica* 73:5, September (2005), 1673-1692.

[10] Chao, John C. and Norman R. Swanson, “Estimation and Testing using Jackknife IV in Heteroskedastic Regressions with Many Weak Instruments,” Rutgers University Working Paper, (2004).

[11] Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen, “Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments,” Working Paper (2007).

[12] Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen, “Instrumental Variables Estimation with Heteroskedasticity and Many Instruments,” Working Paper (2008).

[13] Davidson, Russell and James G. MacKinnon, “The Case Against JIVE,” *Journal of Applied Econometrics* 21:6, September/October (2006), 827-833.

[14] Devereux, Paul J., “Improved Errors-in-Variables Estimators for Grouped Data,” *Journal of Business and Economic Statistics* 25:3, July (2007a), 278-287.

[15] Devereux, Paul J., “Small Sample Bias in Synthetic Cohort Models of Labor Supply,” *Journal of Applied Econometrics* 22:4, (2007b), 839-848.

[16] Donald Stephen G. and Whitney K. Newey, “Choosing the number of instruments,” *Econometrica* 69:5, September (2001), 1161-1191.

[17] Greene, William H., *Econometric Analysis*, 5th Edition, Prentice Hall, NJ (2003).

[18] Guggenberger Patrik and Jinyong Hahn, “Finite Sample Properties of the 2-step Empirical Likelihood Estimator,” *Econometric Reviews* 24:3, (2005), 247-263.

[19] Hahn, Jinyong and Jerry Hausman, “Notes on Bias in Estimators for Simultaneous Equation Models,” *Economics Letters* 75:2, April (2002), 237-241.

[20] Hahn, Jinyong, Hausman, Jerry, and Kuersteiner, Guido, “Estimation with Weak Instruments: Accuracy of Higher Order Bias and MSE Approximations,” *Econometrics Journal* 7:1, (2004), 272-306.

[21] Hansen, Christian, Jerry Hausman, and Whitney K. Newey, “Estimation with Many Instrumental Variables,” CeMMAP Working Paper 19/06 (2006).

[22] Imbens, Guido W., Richard H. Spady, and Phillip Johnson, “Information Theoretic Approaches to Inference in Moment Condition Models,” *Econometrica* 66:2, March (1998), 333-357.

[23] Kitamura, Yuichi and Michael Stutzer, “An Information-Theoretic Alternative to Generalized Method of Moments Estimation,” *Econometrica* 65:4, July (1997), 861-874.

[24] Nagar, A.L., “The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations,” *Econometrica* 27:4, October (1959), 575-595.

[25] Newey, Whitney K., “Many Instrument Asymptotics,” mimeo, MIT (2004).

[26] Newey, Whitney K. and Richard J. Smith, “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica* 72:1, January (2004), 219-255.

[27] Phillips, Gary D. and C. Hale, “The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems” *International Economic Review* 18:1, February (1977), 219-228.

[28] Qin, Jing and Jerry Lawless, “Empirical Likelihood and General Estimating Equations,” *The Annals of Statistics* 22:1, (1994), 300-325.

[29] Staiger Douglas and James H. Stock, “Instrumental variables regression with weak instruments,” *Econometrica* 65:3, May (1997), 557-586.

7 Appendix 1: Higher Order Asymptotic Proofs

This appendix examines the higher order asymptotic properties of the IJIVE and UIJIVE estimators introduced in the paper. In all these proofs, we utilize the following result from AIK (1995):

Lemma 1 *:*

Assume the following model

$$Y = X\beta + \epsilon \tag{23}$$

$$X = Z\pi + \eta \tag{24}$$

where the error terms ϵ and η are i.i.d.. Assume that we can write an estimator $\widehat{\beta}$ in the form

$$\widehat{\beta} = (X'C'X)^{-1}(X'C'Y) \tag{25}$$

where C is an NxN matrix such that the elements of C are of stochastic order $Op(1/\sqrt{N})$ and

$$CX = Z\pi + C\eta$$

Then, with $P_{z\pi} = Z\pi(\pi'Z'Z\pi)^{-1}\pi'Z'$ and $\Sigma_z = p\lim(Z'Z/N)$, the approximate bias of $\widehat{\beta}$ to order $1/N$ equals

$$(\pi'\Sigma_z\pi)^{-1}E\frac{1}{N}\left[\eta'(C' - C'P_{z\pi} - 1)\epsilon\right].$$

■

With this lemma in hand, consider a partialled out version of (1), i.e.

$$\widetilde{Y} = \widetilde{X}\beta^* + \widetilde{\epsilon} \tag{26}$$

$$\widetilde{X} = \widetilde{Z}\pi^* + \widetilde{\eta} \tag{27}$$

where $\widetilde{Y} = M_w Y$, $\widetilde{X} = M_w X^*$, $\widetilde{Z} = M_w Z^*$, $\widetilde{\epsilon} = M_w \epsilon$, $\widetilde{\eta} = M_w \eta$ and $M_w = I - W(W'W)^{-1}W'$.

Consider a C matrix equal to either C_{IJIVE} or C_{UIJIVE} where $C_{IJIVE} = (I - D_{\widetilde{P}_z})^{-1}(\widetilde{P}_z - D_{\widetilde{P}_z})$ and $C_{UIJIVE} = (I - D_{\widetilde{P}_z} + \Psi I)^{-1}(\widetilde{P}_z - D_{\widetilde{P}_z} + \Psi I)$. Here $\widetilde{P}_z = \widetilde{Z}(\widetilde{Z}'\widetilde{Z})^{-1}\widetilde{Z}'$, $D_{\widetilde{P}_z}$ is the diagonal matrix with the diagonal equal to the diagonal of $\widetilde{P}_z$, and $\Psi = \frac{L_1+1}{N}$. Note that for both IJIVE and UIJIVE, $C\widetilde{X} = \widetilde{Z}\pi + C\widetilde{\eta}$ (see derivation (13) in text). Thus, we can write either estimator as

$$\widehat{\beta}^* = (\widetilde{X}'C'\widetilde{X})^{-1}(\widetilde{X}'C'\widetilde{Y}) \tag{28}$$

and the lemma implies that the approximate bias of $\widehat{\beta}^*$ to order $1/N$ equals

$$(\pi'\widetilde{\Sigma_z}\pi)^{-1}E\frac{1}{N}\left[\widetilde{\eta}'(C' - C'\widetilde{P_{z\pi}} - 1)\widetilde{\epsilon}\right], \tag{29}$$

where $\widetilde{P_{z\pi}} = \widetilde{Z}\pi(\pi'\widetilde{Z}'\widetilde{Z}\pi)^{-1}\pi'\widetilde{Z}'$ and $\widetilde{\Sigma_z} = p\lim(\widetilde{Z}'\widetilde{Z}/N)$. Note that this approximate bias is L_1 dimensional – corresponding to the number of columns of $\widetilde{X}$ and the dimension of $\widetilde{\eta}$. The lth element of this approximate bias is

$$(\pi'\widetilde{\Sigma_z}\pi)^{-1}E\frac{1}{N}\left[\widetilde{\eta}_l'(C' - C'\widetilde{P_{z\pi}} - 1)\widetilde{\epsilon}\right] \tag{30}$$

where $\widetilde{\eta}_l$ is the lth column of $\widetilde{\eta}$.

First examine the first term of this approximate bias. We have[21]

$$
\begin{aligned}
E\left(\frac{\widetilde{\eta}_l' C' \widetilde{\epsilon}}{N}\right) &= E\left(\frac{\eta_l' M_w C' M_w \epsilon}{N}\right) && (31)\\
&= trace\left(\frac{M_w C' M_w E(\epsilon \eta_l')}{N}\right)\\
&= \sigma_{\epsilon\eta_l} trace\left(\frac{M_w C' M_w}{N}\right)\\
&= \sigma_{\epsilon\eta_l} trace\left(\frac{(I-P_w) C' (I-P_w)}{N}\right)\\
&= \sigma_{\epsilon\eta_l} trace\left(\frac{(I-P_w) C'}{N}\right) && (32)\\
&= -\sigma_{\epsilon\eta_l}\left[trace\left(\frac{C'}{N}\right) - trace\left(\frac{P_w C'}{N}\right)\right] && (33)
\end{aligned}
$$

For IJIVE, the first of these two terms ($trace\left(C'_{IJIVE}\right)/N$) is zero. For UIJIVE,

$$
\begin{aligned}
trace\left(\frac{C'_{UIJIVE}}{N}\right) &= trace\left(\frac{(\widetilde{P}_z - D_{\widetilde{P}_z} + \Psi I)(I - D_{\widetilde{P}_z} + \Psi I)^{-1}}{N}\right) && (34)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\Psi}{1-\widetilde{P_{z,ii}} + \Psi}\right)
\end{aligned}
$$

where $\widetilde{P_{z,ii}}$ is the ith diagonal element of $\widetilde{P}_z$. The first two terms of an expansion of $\frac{1}{N}\sum_{i=1}^{N}\left(\Psi/(1-\widetilde{P_{z,ii}}+\Psi)\right)$ around the mean of $\widetilde{P_{z,ii}}$ (K_1/N) are

$$
\frac{1}{N}\sum_{i=1}^{N}\left(\frac{L_1+1}{N-K_1+L_1+1}\right) + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\Psi}{(1-\frac{K_1}{N}+\Psi)^2}\right)\left(\widetilde{P_{z,ii}} - \frac{K_1}{N}\right) \qquad (35)
$$

and the omitted terms in the expansion are all of higher order than $1/N$. The first term in (35) equals $(L_1+1)/N$ to order $1/N$ because $(L_1+1)/(N-K_1+L_1+1) - (L_1+1)/N$ is of order $1/N^2$, and

the second term is zero since the trace of $\widetilde{P}_z$ is identically K_1.[22] Hence, to order $1/N$ we have

$$trace\left(\frac{C'_{UIJIVE}}{N}\right) = \frac{L_1+1}{N}$$

For IJIVE, the second term of equation (33) is

$$\begin{aligned} -\sigma_{\epsilon\eta_l} trace\left(\frac{P_w C'_{IJIVE}}{N}\right) &= -\sigma_{\epsilon\eta_l} trace\left(\frac{P_w(\widetilde{P}_z - D_{\widetilde{P}_z})(I - D_{\widetilde{P}_z})^{-1}}{N}\right) \\ &= \sigma_{\epsilon\eta_l} trace\left(\frac{P_w D_{\widetilde{P}_z}(I - D_{\widetilde{P}_z})^{-1}}{N}\right) \\ &= \sigma_{\epsilon\eta_l}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widetilde{P_{z,ii}}P_{w,ii}}{1-\widetilde{P_{z,ii}}}\right) \end{aligned}$$

Note we are using the facts that $trace(C'_{IJIVE}) = 0$ and $P_w\widetilde{P}_z = 0$. In the last line, $P_{w,ii}$ is the ith diagonal element of P_w, and $\sigma_{\epsilon\eta_l}$ is the covariance between ϵ and the lth element of η.

To determine the order of this term, we expand the summands around the mean of $\widetilde{P_{z,ii}}$, K_1/N, and the mean of $P_{w,ii}$, L_2/N. The first three terms of this expansion of $(1/N)\sum_{i=1}^{N}(\widetilde{P_{z,ii}}P_{w,ii})/(1-\widetilde{P_{z,ii}})$ equal

$$\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{\frac{K_1}{N}\frac{L_2}{N}}{1-\frac{K_1}{N}}\right)+\left(\frac{\frac{K_1}{N}}{1-\frac{K_1}{N}}\right)\left(P_{w,ii}-\frac{L_2}{N}\right)+\left(-\frac{\frac{K_1}{N}\frac{L_2}{N}}{(1-\frac{K_1}{N})^2}+\frac{\frac{L_2}{N}}{1-\frac{K_1}{N}}\right)\left(\widetilde{P_{z,ii}}-\frac{K_1}{N}\right)\right] \quad (36)$$

where the omitted terms of the expansion are all of order higher than $1/N$. Since the traces of P_w and $\widetilde{P}_z$ are identically L_2 and K_1 respectively, the second and third terms of (36) sum to zero, and the first term

is of order $1/N^2$. Therefore, to order $1/N$,

$$E\left(\frac{\widetilde{\eta}_l' C'_{IJIVE}\widetilde{\epsilon}}{N}\right) = 0$$

Thus, the first term of the approximate bias for IJIVE is zero. The exact same approach can be used to show that $-\sigma_{\epsilon\eta_l} trace\left((P_w C'_{UIJIVE})/N\right)$ is 0 to order $1/N$. Therefore, to order $1/N$,

$$E\left(\frac{\widetilde{\eta}_l' C'_{UIJIVE}\widetilde{\epsilon}}{N}\right) = \frac{L_1 + 1}{N} \tag{37}$$

Now move to the second term of the approximate bias (29). Noting that $\widetilde{P_{z\pi}} M_w = \widetilde{P_{z\pi}}$ and $\widetilde{P_{z\pi}} P_w = 0$, we have

$$\begin{aligned}
E\left(\frac{\widetilde{\eta}_l' C'\widetilde{P_{z\pi}}\widetilde{\epsilon}}{N}\right) &= E\left(\frac{\eta_l' M_w C'\widetilde{P_{z\pi}} M_w \epsilon}{N}\right) \\
&= E\left(\frac{\eta_l' M_w C'\widetilde{P_{z\pi}}\epsilon}{N}\right) \\
&= E\left(\frac{\eta_l' C'\widetilde{P_{z\pi}}\epsilon}{N}\right) - E\left(\frac{\eta_l' P_w C'\widetilde{P_{z\pi}}\epsilon}{N}\right) \\
&= \sigma_{\epsilon\eta_l}\left[trace\left(\frac{C'\widetilde{P_{z\pi}}}{N}\right) - trace\left(\frac{C'\widetilde{P_{z\pi}} P_w}{N}\right)\right] \\
&= \sigma_{\epsilon\eta_l}\left[\frac{L_1}{N}\right]
\end{aligned}$$

For IJIVE, the last equality results uses the fact that

$$\begin{aligned} trace\left(C'_{IJIVE}\widetilde{P_{z\pi}}\right) &= trace\left(\widetilde{P_{z\pi}}C'_{IJIVE}\right) \\ &= trace\left(\widetilde{P_{z\pi}}(\widetilde{P}_z - D_{\widetilde{P}_z})(I - D_{\widetilde{P}_z})^{-1}\right) \\ &= trace\left(\widetilde{P_{z\pi}}(I - D_{\widetilde{P}_z})(I - D_{\widetilde{P}_z})^{-1}\right) \\ &= trace(\widetilde{P_{z\pi}}) = L_1 \end{aligned} \tag{38}$$

For UIJIVE, the equality results uses the fact that $\widetilde{P_{z\pi}}P_w = 0$ and $trace(C'_{UIJIVE}\widetilde{P_{z\pi}}) = L_1$ (proof is similar to the above).

With these results in hand, return to (29). Stacking the bias terms for the L_1 elements of β^*, we obtain an approximate bias of IJIVE to order $1/N$ of

$$(\pi'\widetilde{\Sigma_z}\pi)^{-1}\sigma_{\epsilon\eta}\left[\frac{-L_1 - 1}{N}\right]$$

where $\sigma_{\epsilon\eta}$ is the L vector of covariances between ϵ and the first L_1 elements of η.

Again, stacking the bias terms for the L_1 elements of β, the approximate bias of β^*_{UIJIVE} to order $1/N$ is

$$(\pi'\widetilde{\Sigma_z}\pi)^{-1}\sigma_{\epsilon\eta}\left[\frac{L_1 + 1 - L_1 - 1}{N}\right] = 0 \tag{39}$$

Notes

[1]Corresponding author: Daniel A. Ackerberg, Dept. of Economics, UCLA, Los Angeles, CA 90095. Thanks to Jin Hahn for very helpful discussions. All errors are our own. The first author acknowledges generous support under NSF grant #0339850.

[2]Chao and Swanson (2004) also independently prove that the original JIVE estimator is consistent under Bekker-type asymptotics with heteroskedasticity and Chao et al. (2007) provide further results. Chao et al. (2008) propose and study new, related, estimators. These are the only studies we know of, other than ours, that examine the theoretical properties of these types of estimators under heteroskedasticity.

[3]A conceptually distinct issue is the case where instruments are weak in terms of having very low correlation with the endogenous variable. This causes biases in 2SLS even in just-identified models. The estimators we propose are not designed to deal with this "weak instruments" problem.

[4]Note that since W is in both Z and X, the last L_2 of the η's are identically zero, and thus the last L_2 of the L covariances in $\sigma_{\epsilon\eta}$ are identically zero.

[5]AIK also suggest a alternative version of the JIVE estimator, with

$$\widetilde{\pi}(i) = (Z'Z)^{-1}(Z(i)'X(i)).(N/(N-1))$$

As all evidence suggests that these two JIVE estimators have similar properties, we focus on the version in equation (2).

[6]While this formulation is quite compact, it does require the manipulation of an NxN matrix, which can be problematic in large datasets. Davidson and MacKinnon (2006) provide an alternative formulation of the JIVE estimator that avoids this (this can also be applied to our IJIVE and UIJIVE estimators).

[7]For intuition on why partialling out W beforehand eliminates the L_2 bias term, see a prior version of

this paper, Ackerberg and Devereux (2003) available at http://www.econ.ucla.edu/ackerber/newjive14all.pdf.

[8]Note that because of the partialling out of W, this bias vector is L_1dimensional – $\sigma_{\epsilon\eta}$ is now the vector of correlations between ϵ and only the first L_1 elements of η. Note also that the first L_1 by L_1 block of $(\pi'\Sigma_Z\pi)^{-1}$ is identical to $(\pi'\Sigma_{\widetilde{Z}}\pi)^{-1}$. This, combined with the fact that the last L_2 elements of $\sigma_{\epsilon\eta}$ in the JIVE bias formula are zero, implies that the first L_1 elements of the $(\pi'\Sigma_Z\pi)^{-1}\sigma_{\epsilon\eta}$ component in the JIVE bias are equal to the $(\pi'\Sigma_{\widetilde{Z}}\pi)^{-1}\sigma_{\epsilon\eta}$ component in the IJIVE bias.

[9]As we note below, this correction is similar to the correction used by Donald and Newey (2001) on the Nagar estimator to get an approximately unbiased estimator.

[10]AIK show that the original JIVE estimator is consistent under these asymptotics under the assumption of homoskedasticity.

[11]Because of their superior properties, we focus on the consistency of the IJIVE and UIJIVE estimators with heteroskedasticity under many-instrument asymptotics. Note that this proof of consistency trivially applies to the original JIVE estimator as well.

[12]Note that in one special case, this bias term disappears – when the diagonal of P_z is constant (i.e. $\widetilde{P}_{z,ii} = 0\ \forall i$). This special case occurs, for example, when instruments are group dummies and there are equal numbers of observations in each group (in this case, $P_{z,ii} = K_1/N\ \forall i$). See Devereux (2007a) and Devereux (2007b) for analysis of grouping models.

[13]In the working paper version (Ackerberg and Devereux (2003)), we examine both discrete and continuous instruments and find similar results.

[14]The average first stage F-statistic in our base model is around 3. While this indicates that the instruments are weak, they are significant at conventional levels. The reason we chose this level for our Monte-Carlos is that it makes the differences between the various estimators more obvious in the experiments - these differences get smaller as the the F-statistic increases (see Ackerberg and Devereux (2006)

for Monte-Carlos at higher F-statistic levels).

[15]This true coverages rates simply use the empirical distribution of the estimates across replications to form confidence intervals, similar to what would be done with bootstrapped confidence intervals.

[16]To derive this formula, note that each of the JIVE estimators can be rewritten as a standard, just-identified, IV estimator (i.e. $\widehat{\beta} = \left(\widehat{X}'X\right)^{-1}\widehat{X}'Y$) where the instrument $\widehat{X} = CX$ is the "first stage predicted value" using the appropriate C (C_{JIVE}, C_{IJIVE}, or C_{UIJIVE}). Under conventional asymptotics (and regularity conditions), error in the estimation of the instrument $\widehat{X}$ disappears (see Greene (2003), pp 76-78 and Theorem 5.3). Thus, the estimator is asymptotically normal with the standard IV variance $Var(\widehat{\beta}) = \widehat{\sigma}^2\left(X'P_{\widehat{X}}X\right)^{-1}$, which expands to the formula in the main text. A developing literature (e.g. Hansen, Hausman and Newey (2005), Chao and Swanson (2004)) develop more sophisticated alternative asymptotic approximations that are likely to perform better than those we use. However, for our estimators of choice, these standard asymptotic approximations seem to do reasonably well, at least in our particular Monte-Carlo experiments.

[17]Note that these specifications only have one endogenous explanatory variable. Given our theoretical results, one might expect the relative performance of the UIJIVE estimator (vs. IJIVE) to increase as the number of endogenous variables increases. Also note that the small median biases apparent in IJIVE and UIJIVE when dim(W)=10 are coming from high order terms - when one increases N to, e.g., 200 the biases quickly disappear (this is not the case with the JIVE biases).

[18]This is again derived under standard asymptotics, i.e. using White-adjusted standard errors for a just-identified IV estimator where the instrument is $\widehat{X} = CX$. Again, recent work (Chao and Swanson (2004)) develops alternative approximations that could perform better. In any case, if coverage rates are poor for our *infeasible* confidence intervals, it is likely that even the best feasible asymptotic approximation will also be poor.

[19]Allowing heteroskedasticity in the covariances (rather than the variances) seems to have more of a detrimental effect on the performance of the Nagar, B2SLS, and LIML estimators.

[20]We have also tried implementing the J2SLS estimator. J2SLS is very time-consuming in this application because of the large sample size and the large number of explanatory variables and instruments (unlike JIVE type estimators, for which there is a shortcut, the J2SLS estimator actually requires running $N+1$ separate regressions). For the first specification (which took 4 days to run), we obtained a point estimate of 0.092. The second specification would take considerably longer to run.

[21]Note that for all these proofs, the bias terms are of dimension L_1, the number of endogenous variables X^*. Note that η is $L_1 x N$ dimensional. The proofs should be interpreted as element by element computations of small sample bias, e.g. $\sigma_{\epsilon\eta}$ is the covariance of ϵ with one of the L_1 elements of η.

[22]Note that we could have used an alternative Ψ,

$$\Psi = \left(\frac{L_1+1}{N}\right)\left(\frac{N-K_1}{N-L_1-1}\right)$$

to get the first term in the expansion exactly equal to $(L_1+1)/N$. This differs from the first term using the current Ψ only in a term of order $1/N^2$. While the alternative Ψ would make the proof a bit cleaner, we chose the current Ψ to make our UIJIVE adjustment analagous to Donald and Newey's B2SLS adjustment to Nagar.

Table 1 - Number of Exogenous Included Variables

	10%	25%	Median	75%	90%	Median Absolute Error	Trimmed Mean Bias	Trimmed Mean Abs. Error	90% C.I. Coverage (Infeasible)	90% C.I. Coverage (Feasible)
Panel A: Dim(W) = 0										
OLS	0.4809	0.5290	0.5817	0.6369	0.6836	0.5818	0.5825	0.5825	0.0000	0.0000
2SLS	0.1136	0.1866	0.2694	0.3503	0.4265	0.2695	0.2689	0.2689	0.2913	0.2615
JIVE	-0.5813	-0.2467	-0.0314	0.1209	0.2316	0.1684	-0.0828	0.1965	0.9129	0.9064
IJIVE	-0.4868	-0.2007	-0.0039	0.1383	0.2450	0.1622	-0.0464	0.1777	0.9017	0.8901
UIJIVE	-0.3696	-0.1377	0.0358	0.1649	0.2680	0.1552	0.0040	0.1583	0.8774	0.8582
Panel B: Dim(W) = 1										
OLS	0.4807	0.5286	0.5818	0.6358	0.6822	0.5818	0.5820	0.5820	0.0000	0.0000
2SLS	0.1120	0.1900	0.2712	0.3513	0.4238	0.2713	0.2701	0.2701	0.2811	0.2571
JIVE	-0.6511	-0.2855	-0.0537	0.1036	0.2237	0.1735	-0.1124	0.2133	0.9147	0.9175
IJIVE	-0.4690	-0.1892	-0.0015	0.1376	0.2491	0.1574	-0.0408	0.1730	0.9011	0.8859
UIJIVE	-0.3503	-0.1270	0.0384	0.1640	0.2698	0.1516	0.0094	0.1546	0.8765	0.8537
Panel C: Dim(W) = 5										
OLS	0.4807	0.5283	0.5807	0.6350	0.6833	0.5807	0.5815	0.5815	0.0000	0.0000
2SLS	0.1166	0.1939	0.2754	0.3584	0.4379	0.2754	0.2761	0.2761	0.2887	0.2530
JIVE	-1.1700	-0.5118	-0.1606	0.0487	0.1963	0.2400	-0.2793	0.3487	0.9166	0.9513
IJIVE	-0.4439	-0.1662	0.0180	0.1581	0.2691	0.1612	-0.0206	0.1717	0.8924	0.8706
UIJIVE	-0.3344	-0.1052	0.0548	0.1836	0.2885	0.1585	0.0266	0.1572	0.8689	0.8348
Panel D: Dim(W) = 10										
OLS	0.4772	0.5261	0.5818	0.6365	0.6837	0.5818	0.5812	0.5812	0.0000	0.0000
2SLS	0.1251	0.2023	0.2839	0.3651	0.4474	0.2839	0.2843	0.2843	0.2743	0.2444
JIVE	-2.1177	-0.8620	-0.3059	0.0010	0.3047	0.4383	-0.4607	0.6746	0.8985	0.9602
IJIVE	-0.4018	-0.1460	0.0386	0.1733	0.2887	0.1634	0.0036	0.1690	0.8770	0.8542
UIJIVE	-0.2963	-0.0888	0.0725	0.1978	0.3074	0.1608	0.0482	0.1582	0.8521	0.8199

Notes: First five columns show quantiles of the distribution of the estimator over the 10000 replications. Trimming for the "Trimmed Mean Bias" and "Trimmed Mean Abs. Error" columns is done at the 5th and 95th quantiles. In the last two columns, C.I. refers to confidence interval. Infeasible C.I.'s use the empirical distribution of the estimates to form the confidence intervals. Feasible C.I.'s use the asymptotic approximation formula on page 9 (page 11 for Table 2).

Table 2 - Heteroskedasticity

	10%	25%	Median	75%	90%	Median Absolute Error	Trimmed Mean Bias	Trimmed Mean Abs. Error	90% C.I, Coverage (Infeasible)	90% C.I, Coverage (Feasible)
Panel A: Baseline - $\sigma(\text{large}) = 0.2, \sigma(\text{small}) = 0.2$										
OLS	0.4787	0.5394	0.5988	0.6582	0.7076	0.5988	0.5976	0.5976	0.0000	0.0000
2SLS	0.1166	0.1973	0.2865	0.3795	0.4652	0.2866	0.2887	0.2887	0.3404	0.2388
J2SLS	-0.1498	-0.0134	0.1159	0.2375	0.3439	0.1587	0.1096	0.1548	0.8282	-
IJIVE	-0.5920	-0.2213	-0.0019	0.1534	0.2816	0.1775	-0.0564	0.2045	0.9009	0.8780
UIJIVE	-0.4023	-0.1331	0.0487	0.1885	0.3082	0.1690	0.0138	0.1743	0.8802	0.8351
NAGAR	-0.5631	-0.2214	-0.0078	0.1453	0.2667	0.1720	-0.0596	0.1984	0.9007	0.8921
B2SLS	-0.3850	-0.1312	0.0428	0.1809	0.2934	0.1634	0.0110	0.1683	0.8807	0.8483
LIML	-0.3380	-0.1488	-0.0015	0.1196	0.2203	0.1313	-0.0220	0.1370	0.9006	0.9166
PEL	-0.0534	0.0663	0.1800	0.2901	0.3902	0.1914	0.1757	0.1884	0.7110	-
Panel B: $\sigma(\text{large}) = 0.0, \sigma(\text{small}) = 0.2$										
OLS	0.2082	0.2642	0.3248	0.3849	0.4396	0.3248	0.3244	0.3244	0.0267	0.0267
2SLS	0.0915	0.1767	0.2731	0.3671	0.4593	0.2733	0.2732	0.2732	0.3994	0.3066
J2SLS	-0.1525	-0.0208	0.1100	0.2395	0.3527	0.1558	0.1064	0.1556	0.8406	-
IJIVE	-0.5626	-0.2045	0.0011	0.1603	0.3020	0.1783	-0.0444	0.2013	0.8997	0.8745
UIJIVE	-0.4230	-0.1468	0.0318	0.1761	0.3060	0.1654	0.0012	0.1741	0.8870	0.8527
NAGAR	-0.1089	0.0577	0.2157	0.3821	0.5634	0.2464	0.2198	0.2487	0.8271	0.6800
B2SLS	-0.0644	0.0813	0.2260	0.3781	0.5347	0.2445	0.2300	0.2463	0.7869	0.6345
LIML	-0.0572	0.0797	0.2251	0.3742	0.5312	0.2430	0.2289	0.2436	0.7848	0.6228
PEL	-0.0325	0.0942	0.2287	0.3667	0.4994	0.2361	0.2306	0.2393	0.7261	-
Panel C: $\sigma(\text{large}) = 0.2, \sigma(\text{small}) = 0.0$										
OLS	0.1515	0.2099	0.2722	0.3358	0.3944	0.2722	0.2729	0.2729	0.1062	0.1026
2SLS	-0.1683	-0.0779	0.0176	0.1120	0.1942	0.0954	0.0165	0.0932	0.8969	0.8816
J2SLS	-0.2599	-0.1292	0.0069	0.1369	0.2588	0.1329	0.0030	0.1317	0.8991	-
IJIVE	-0.3904	-0.1843	-0.0074	0.1568	0.3334	0.1674	-0.0160	0.1765	0.9008	0.9199
UIJIVE	-0.3138	-0.1431	0.0128	0.1635	0.3203	0.1545	0.0088	0.1569	0.8998	0.9118
NAGAR	-0.8318	-0.4574	-0.2242	-0.0527	0.0783	0.2455	-0.2780	0.2975	0.9143	0.8652
B2SLS	-0.6658	-0.3808	-0.1833	-0.0271	0.0918	0.2049	-0.2193	0.2434	0.9080	0.8642
LIML	-0.5715	-0.3655	-0.1914	-0.0513	0.0684	0.2065	-0.2143	0.2312	0.8804	0.8734
PEL	-0.3416	-0.1889	-0.0442	0.0898	0.2102	0.1404	-0.0525	0.1437	0.8968	-
Panel D: $\sigma(\text{large}) = 0.1, \sigma(\text{small}) = 0.2$										
OLS	0.3460	0.4014	0.4604	0.5189	0.5731	0.4604	0.4599	0.4599	0.0002	0.0001
2SLS	0.1033	0.1864	0.2790	0.3739	0.4648	0.2791	0.2808	0.2808	0.3824	0.2747
J2SLS	-0.1556	-0.0210	0.1138	0.2362	0.3510	0.1577	0.1073	0.1563	0.8392	-
IJIVE	-0.5925	-0.2177	-0.0007	0.1573	0.2998	0.1794	-0.0527	0.2070	0.9001	0.8764
UIJIVE	-0.4208	-0.1465	0.0391	0.1814	0.3088	0.1688	0.0040	0.1777	0.8841	0.8458
NAGAR	-0.3251	-0.0766	0.1020	0.2512	0.3982	0.1914	0.0778	0.1923	0.8634	0.8015
B2SLS	-0.2151	-0.0231	0.1325	0.2716	0.4060	0.1897	0.1186	0.1861	0.8337	0.7508
LIML	-0.2107	-0.0316	0.1142	0.2506	0.3856	0.1742	0.1045	0.1731	0.8534	0.7830
PEL	-0.0465	0.0730	0.2005	0.3247	0.4498	0.2124	0.2002	0.2115	0.7476	-
Panel E: $\sigma(\text{large}) = 0.0, \sigma(\text{small}) = 0.2, \text{N} = 500$										
OLS	0.2563	0.2811	0.3068	0.3374	0.3630	0.3069	0.3087	0.3087	0.0000	0.0000
2SLS	0.1577	0.2027	0.2433	0.2836	0.3220	0.2435	0.2426	0.2426	0.0170	0.0130
J2SLS	-0.0239	0.0246	0.0803	0.1344	0.1853	0.0853	0.0800	0.0861	0.7410	-
IJIVE	-0.1644	-0.0802	-0.0028	0.0608	0.1182	0.0703	-0.0115	0.0711	0.9060	0.8770
UIJIVE	-0.1537	-0.0720	0.0031	0.0655	0.1228	0.0684	-0.0045	0.0695	0.8970	0.8680
NAGAR	0.0578	0.1197	0.1820	0.2452	0.3013	0.1822	0.1815	0.1815	0.3850	0.3490
B2SLS	0.0618	0.1227	0.1835	0.2473	0.3023	0.1838	0.1839	0.1839	0.3660	0.3260
LIML	0.0725	0.1281	0.1899	0.2508	0.3062	0.1902	0.1893	0.1893	0.3080	0.2850
PEL	0.0966	0.1441	0.2076	0.2652	0.3233	0.2076	0.2068	0.2068	0.2460	-

Notes: See Table 1

Table 3 - Angrist - Kreuger Data

	OLS	2SLS	LIML	JIVE	IJIVE	UIJIVE	Nagar	B2SLS
Panel A: Year Effects, K=30, L=11								
Estimate	0.071	0.089	0.093	0.096	0.094	0.093	0.094	0.093
S.E.	0.0003	0.016	0.018	0.022	0.019	0.019	0.019	0.019
Panel B: Year and State Effects, K=180, L=61								
Estimate	0.067	0.093	0.106	0.121	0.110	0.109	0.109	0.109
S.D.	0.0003	0.009	0.012	0.020	0.012	0.012	0.012	0.012