MANY INSTRUMENTS AND JUDGES

Michal Kolesár[*]

April 3, 2024

In this lecture, we'll consider some issues that arise in the linear instrumental variables (IV) model when the number of instruments K is large. We may find ourselves in a setting with many instruments for two main reasons. First, when we take an original, low-dimensional instrument, and interact it with controls. A classic example is Angrist and Krueger (1991). More generally, we could take an underlying low-dimensional set of instruments and covariates, and take a series expansion of it in an effort to find a transformation that leads to a strong first stage—see Belloni et al. (2012).

The second setting in which this problem arises is when we use fixed effects (group indicators) as instruments: if there are many effects, we have many instruments. This setting has been quite common in recent empirical work, and is sometimes referred to as "judges" or "examiners" designs. The basic idea is as follows. Suppose we're interested in the effect of incarceration on some economic outcome, as in Kling (2006). We can't just regress the outcome of interest on the length of sentence, since it's unlikely, even if we control for observables, that the length of sentence is as good as randomly assigned. However, we can exploit the fact that (perhaps conditional on time and location dummies) the judge assigned to a defendant is as good as random. If some judges tend to hand out longer sentences than others, this will generate variation in the sentencing length that's as good as random, and we can use we can use judge indicators as instruments. If there are many judges, then we have many instruments.

Apart from studying the effects of incarceration on economic outcomes (e.g. Kling 2006; Aizer and Doyle 2015), this design has been used in a number of other settings: we may exploit random assignment of judges to bankruptcy cases, criminal cases, or patent cases, doctors to shifts, or social workers to cases—see the introduction to Frandsen, Lefgren, and Leslie (2023) for a partial list of applications.

We'll consider the effects of many instruments on:

1. the bias of the two-stage least squares (TSLS) estimator, and consider alternatives to TSLS
2. standard errors
3. the interpretation of the first-stage F statistic

[*] Email: mkolesar@princeton.edu.

We'll see that the conclusion in Bound, Jaeger, and Baker (1995) critique of Angrist and Krueger (1991) that "the natural experiment afforded by the interaction between compulsory school attendance laws and quarter of birth does not give much usable information regarding the causal effect of education on earnings" is not quite correct: with the right estimator and right standard errors, we can still extract useful information from the data and report a reliable assessment of its accuracy.

1. SETUP AND ESTIMATION

We use the same notation as in the previous lecture, with the reduced form and the first stage given by

$$Y_i = Z_i'\delta + W_i'\psi_Y + u_{Yi}, \tag{1}$$

$$D_i = Z_i'\pi + W_i'\psi_D + u_{Di}. \tag{2}$$

The parameter of interest is given by

$$\beta := \frac{\delta' Q\pi}{\pi' Q\pi},$$

where $Q = E[\tilde{Z}_i\tilde{Z}_i']$, with $\tilde{Z}_i = Z_i - E[Z_i W_i']E[W_i W_i']^{-1}W_i$. To measure the overall strength of the instruments, define

$$r_n = n\pi' Q\pi = nE[(\tilde{Z}_i'\pi)^2].$$

You can think of r_n as the effective sample size, in the sense that it scales the sample size n by instrument strength, the denominator in the definition of β. As discussed in the previous lecture, under homogeneous treatment effects, β corresponds to the causal effect of D_i on Y_i, and the reduced-form coefficients are proportional to each other, $\delta = \pi\beta$. Under heterogeneous treatment effects, β has the interpretation as a weighted average of local average treatment effects (LATEs) if we make a monotonicity assumption.

Remark 1 (Notation). For any matrix A, let $H_A = A(A'A)^{-1}A'$ denote the hat matrix (also called a projection matrix), and let $\ddot{A} = A - H_W A$ denote the residuals after projecting A onto the covariates. Thus, for instance, $\ddot{Z}_i$ is the sample analog of $\tilde{Z}_i$. Let $X_i = (Z_i', W_i')'$ collect the right-hand side (RHS) (or "exogenous") variables.

Example 1 (Judges design). To make the above setup concrete, let us consider a setting in which individual i is assigned judge Q_i, which is random conditional on i's geographic location G_i. Then the covariates and instruments are both indicators: $W_{i\ell} = \mathbb{1}\{G_i = \ell\}$, and $Z_{ik} = \mathbb{1}\{Q_i = k\}$. Let us assume that each judge only works in one location, and that the treatment of interest D_i is an indicator for incarceration. We're interested in the effect of incarceration on some economic outcome. There are L locations and $K + L$

judges; in each location we drop one judge to avoid collinearity—call this judge the reference judge. Then $\psi_{D,\ell}$ measures the average sentencing rate of the reference judge in location ℓ, and π_k measures the sentencing rate of judge k relative to the reference judge in the same location. ⊠

Remark 2. The monotonicity assumption in Example 1 requires that judges agree on the ranking of the defendants, they just disagree on the cutoff at which they start sentencing them. Such an assumption can be problematic. For example, Chan, Gentzkow, and Yu (2019) point out that decisions of physicians differ due to both skill and preferences. They look at radiologists who diagnose whether a patient has pneumonia, and define a type II error as a patient who was not diagnosed, but who have a subsequent pneumonia diagnosis in the next 10 days. They show that radiologists who diagnose at higher rates actually have higher rather than lower type II error rates. Similarly, Kleinberg et al. (2018) find that the increase in crime associated with judges who are more likely to release defendants on bail is about the same as if these more lenient judges randomly picked the extra defendants to release on bail. See also Frandsen, Lefgren, and Leslie (2023). However, since failure of monotonicity only affects the interpretation of β, to keep the statistical issues separate from identification issues, we put these issues aside here.

To capture the effect of the potentially large number of instruments, $K = \dim(Z_i)$, or covariates $L = \dim(W_i)$ on the finite-sample behavior of estimators in the asymptotic approximation, we consider asymptotics in which both K and L are allowed to grow with sample size. Under such asymptotics, we obtain the following result:

Lemma 3. The TSLS estimator suffers from own observation bias towards ordinary least squares (OLS). In particular, suppose $E[u_i \mid X_i] = 0$, with $\Omega(X_i) = E[u_i u_i' \mid X_i]$ bounded, $E[\tilde{Z}_i \mid W_i] = 0$, and $L/n \to 0$.

Then consistency of TSLS is in general requires $K/r_n \to 0$. Under homoskedastic errors,

$$\hat{\beta}_{TSLS} = \beta + \frac{(\Omega_{YD} - \Omega_{DD}\beta)K}{r_n + \Omega_{DD}K} + o_p(1) = (1-w)\beta + w\beta_{OLS} + o_p(1), \quad w = \frac{K}{r_n/\Omega_{DD} + K},$$

where $\beta_{OLS} = (\Omega_{YD} - \Omega_{DD}\beta)/\Omega_{DD} + \beta$ is the probability limit of OLS.

Proof. We sketch the argument, see Evdokimov and Kolesár (2018) for a precise proof. Recall from extremum estimation theory that typically, under regularity conditions, an estimator $\hat{\beta}$ that minimizes a sample objective function $\hat{Q}_n(\beta)$ converges in probability to the minimizer of $\beta_n = \operatorname{argmin}_b Q_n(b) = E[\hat{Q}_n(b)]$ in the sense that $\hat{\beta}_n - \beta_n \xrightarrow{p} 0$, provided that (uniformly) $\hat{Q}_n(b) - Q_n(b) \xrightarrow{p} 0$. In our case, the TSLS objective function is

$$\begin{aligned}\hat{Q}_n(b) &= (Y - Db)'H_{\ddot{Z}}(Y - Db) \\ &= (\delta - \pi b)'\ddot{Z}'\ddot{Z}(\delta - \pi b) + 2(\delta - \pi b)H_{\ddot{Z}}(u_Y - u_D b) + (u_Y - u_D b)H_{\ddot{Z}}(u_Y - u_D b),\end{aligned}$$

where the second line follows from $H_{\ddot{Z}}(Y - Db) = \ddot{Z}(\delta - \pi b) + H_{\ddot{Z}}(u_Y - u_D b)$. Using the assumption $E[u_i \mid Z_i, W_i] = 0$, we have[1]

$$Q_n(b) = (\delta - \pi b)' E[\ddot{Z}'\ddot{Z}](\delta - \pi b) + \sum_i E[(u_Y - u_D b)^2 H_{\ddot{Z},ii}].$$

This is minimized at

$$\beta_n = \frac{\pi' E[\ddot{Z}'\ddot{Z}]\delta + \sum_i E[u_{Yi} u_{Di} H_{\ddot{Z},ii}]}{\pi' E[\ddot{Z}'\ddot{Z}]\pi + \sum_i E[u_{Di}^2 H_{\ddot{Z},ii}]} = \beta + \frac{\pi' E[\ddot{Z}'\ddot{Z}](\delta - \pi\beta) + \sum_i E[(u_{Yi} - u_{Di}\beta) u_{Di} H_{\ddot{Z},ii}]}{\pi' E[\ddot{Z}'\ddot{Z}]\pi + \sum_i E[u_{Di}^2 H_{\ddot{Z},ii}]}$$

Under regularity conditions, replacing $E[\ddot{Z}'\ddot{Z}]$ in this expression by nQ will have a negligible effect, so that we may write

$$\beta_n = \beta + \frac{\sum_i E[(u_{Yi} - u_{Di}\beta) u_{Di} H_{\ddot{Z},ii}]}{r_n + \sum_i E[u_{Di}^2 H_{\ddot{Z},ii}]} + o_p(1),$$

By boundedness of $\Omega(X_i)$, $\sum_i E[(u_{Yi} - u_{Di}\beta) u_{Di} H_{\ddot{Z},ii}]$ is bounded by a constant times $\sum_i H_{\ddot{Z},ii} = K$, so the bias is of the order $K/(r_n + K)$, which gives the rate for the consistency result. □

The key thing here is that the TSLS bias scales with K/r_n, rather than K/n: thus, the TSLS bias can be substantial even if K/n is very small if the instruments are not very strong.

The bias comes from the fact that the single constructed instrument $\hat{Z}_{\text{TSLS},i} = Z_i'\hat{\pi} + W_i'\hat{\psi}_D$ used by TSLS puts positive weight on own treatment status D_i, which means that the constructed instrument is slightly endogenous. The total net weight, holding overall instrument strength constant, scales linearly with K, so that the TSLS bias scales with the number of instruments.

Remark 4. Recall from the previous lecture note that the oracle who knows the first stage would use the single instrument $\hat{Z}_i^* = Z_i'\pi + W_i'\psi_D$, estimating β as

$$\hat{\beta}^* = \frac{\hat{Z}^{*\prime}\ddot{Y}}{\hat{Z}^{*\prime}\ddot{D}} = \frac{\hat{D}^{*\prime}Y}{\hat{D}^{*\prime}D}, \qquad \hat{D}^* = \ddot{Z}\pi.$$

Here $\hat{D}^*$ is the single instrument $\hat{Z}^*$ with the covariates partialled out. We also discussed that $\hat{Z}^*$ (or equivalently $\hat{D}^*$) has the interpretation as the optimal instrument under some conditions. The TSLS estimator replaces the unknown $\hat{D}^*$ by the estimate $H_{\ddot{Z}}D = \ddot{Z}\hat{\pi}$, $\hat{\pi} = (\ddot{Z}'\ddot{Z})^{-1}\ddot{Z}'D$. Under standard asymptotics that hold π and K fixed, so that $r_n \asymp n$, this has no effect on consistency of the estimator. However, when K is large, the noise in $\hat{\pi}$ may become non-negligible relative to the signal r_n. Lemma 3 makes it precise when this happens.

Remark 5 (Partial solutions). The first solution TSLS inconsistency, dating back to Bekker (1994), the first paper to analyze the many instrument problem, is to use limited information maximum likelihood (LIML), or variants thereof. Indeed, one can show that under homogeneous treatment effect and homoskedastic errors, we only need $\sqrt{K}/r_n \to 0$ for consistency, a substantially weaker requirement. To ensure consistency under heteroskedastic errors, further adjustments are needed. However, as discussed in the pre-

[1] The expectation of a quadratic form $Y'HY$ is $E[Y'HY] = E[\text{tr}(HYY')] = \text{tr}(HE[YY'])$, since trace is a linear operator.

vious lecture, an issue with LIML-like estimators is that they are not robust to heterogeneous treatment effects.

The second solution is to directly correct for the TSLS bias by subtracting an estimate of its bias. This is easy to do under homoskedastic errors, and leads to variants of the bias-corrected TSLS estimator that dates back to Nagar (1959). However, its consistency does depend on homoskedasticity. We can think of the jackknife estimators, discussed next, as heteroskedasticity-robust analogs.

If the bias is caused by using D_i in constructing the single instrument $\hat{Z}_{\text{TSLS},i}$, then an obvious solution is to not use it. There are multiple ways of implementing this idea. The first option is to construct predictors $\hat{\pi}_{-i}$ and $\hat{\psi}_{D,-i}$ of π and ψ_D based on a regression of D onto (Z, W), but with the ith observation removed, to construct a single instrument

$$\hat{Z}_{\text{JIVE1},i} = Z_i' \hat{\pi}_{-i} + W_i' \hat{\psi}_{D,-i}.$$

In the judges design, $\hat{Z}_{\text{JIVE1},i}$ is just the average sentencing rate of the judge assigned to me, using all observations except myself,

$$\hat{Z}_{\text{JIVE1},i} = \frac{\sum_{j \neq i} D_i \mathbb{1}\{Q_j = Q_i\}}{\sum_{j \neq i} \mathbb{1}\{Q_j = Q_i\}}.$$

Then run an IV regression of Y_i onto D_i and W_i, using $\hat{Z}_{\text{JIVE1},i}$ as an instrument for D_i. The resulting estimator is known as jackknife instrumental variables estimator (JIVE1), and can be written as

$$\hat{\beta}_{\text{JIVE1}} = \frac{\hat{Z}_{\text{JIVE1}}' \ddot{Y}}{\hat{Z}_{\text{JIVE1}}' \ddot{D}} = \frac{\hat{D}_{\text{JIVE1}}' Y}{\hat{D}_{\text{JIVE1}}' D}, \quad \hat{Z}_{\text{JIVE1}} = \ddot{H} D, \quad \hat{D}_{\text{JIVE1}} = (I - H_W)\hat{Z},$$
$$\ddot{H} = (I - \operatorname{diag}(H_X))^{-1}(H_X - \operatorname{diag}(H_X)),$$

In particular, we don't need to run n first-stage regressions to construct $\hat{Z}_{\text{JIVE1},i}$ for each i; we can do everything in one step using matrix algebra.[2] This is the version of jackknife discussed in Angrist, Imbens, and Krueger (1999), and Blomquist and Dahlberg (1999), and goes back to Phillips and Hale (1977). However, a problem with this implementation is that when we project out the covariates from $\hat{Z}_i$, we re-introduce the own observation bias: the instrument $\hat{D}_i$ with the covariates projected out does depend on D_i. In the judges design, we adjust $\hat{Z}_{\text{JIVE1},i}$ by the average sentencing rate in the location of i,

$$\hat{D}_{\text{JIVE1},i} = \hat{Z}_{\text{JIVE1},i} - \frac{\sum_j \hat{Z}_j \mathbb{1}\{G_j = G_i\}}{\sum_j \mathbb{1}\{G_j = G_i\}} = \hat{Z}_{\text{JIVE1},i} - \frac{\sum_j D_j \mathbb{1}\{G_j = G_i\}}{\sum_j \mathbb{1}\{G_j = G_i\}},$$

since $\sum_{j \colon Q_j = Q_i} \hat{Z}_j = \sum_{j \colon Q_j = Q_i} D_j$. Because the average sentencing rate in location i depends on D_i, this re-introduces the own observation bias. Indeed, we generally need

[2] Though implementing this estimator (and the estimators below) in practice does require computing the diagonal of the projection matrix H_X, which may challenging when the dimensionality of the data is very large.

$L/r_n \to 0$ for consistency (see Evdokimov and Kolesár (2018) for a formal result), which can be a stringent requirement if there are many covariates. Furthermore, under homoskedastic errors, the bias goes in the opposite direction than the TSLS bias:

$$\hat{\beta}_{\text{JIVE1}} = \beta - \frac{(\Omega_{YD} - \Omega_{DD}\beta)L}{r_n - \Omega_{DD}L} + o_p(1) = (1+\lambda)\beta - \lambda\beta_{OLS} + o_p(1),\ \lambda = \frac{L}{r_n/\Omega_{DD} - L}.$$

In many applications the JIVE1 bias can be *bigger* in magnitude than that of TSLS—see Section 4.

There are two solutions to this problem. The first is to *first* partial out the covariates, and then do the leave-one-out prediction, leading to the improved improved jackknife instrumental variables estimator (IJIVE1) estimator proposed in Ackerberg and Devereux (2009). That is: (i) compute the residuals $\ddot{Z}, \ddot{D}, \ddot{Y}$ from regressing Z, D, Y onto W, and then (ii) compute $\hat{D}_{\text{IJIVE1},i} = \ddot{Z}_i\hat{\pi}_{-i}$, where $\hat{\pi}_{D,-i}$ is the estimate from the regression of $\ddot{D}$ onto $\ddot{Z}$, with observation i excluded. In one step, this can be written as:

$$\hat{\beta}_{\text{IJIVE1}} = \frac{\ddot{D}'\ddot{H}'\ddot{Y}}{\ddot{D}'\ddot{H}'\ddot{D}}, \qquad \ddot{H} = (I - \operatorname{diag}(H_{\ddot{Z}}))^{-1}(H_{\ddot{Z}} - \operatorname{diag}(H_{\ddot{Z}})).$$

The second solution is to exclude own observation both when calculating the severity of the judge assigned to i, and also when calculating the average sentencing rate in the location of i,

$$\hat{D}_{\text{UJIVE},i} = \hat{Z}_{\text{JIVE1},i} - \frac{\sum_{j\neq i} D_j \mathbb{1}\{G_j = G_i\}}{\sum_{j\neq i}\mathbb{1}\{G_j = G_i\}}$$

Kolesár (2013) calls this estimator unbiased jackknife instrumental variables estimator (UJIVE). With general instruments and covariates, the estimator takes the form

$$\hat{\beta}_{\text{UJIVE}} = \frac{D'\ddot{H}'Y}{D'\ddot{H}'D},$$
$$\ddot{H} = (I - \operatorname{diag}(H_X))^{-1}(H_X - \operatorname{diag}(H_X)) - (I - \operatorname{diag}(H_W))^{-1}(H_W - \operatorname{diag}(H_W)).$$

This implements the following procedure in one step: (i) run the regression of D onto (W, Z) with ith observation removed to compute $\hat{\pi}_{-i}$ and $\hat{\psi}_{D,-i}$. Compute $\hat{Z}_{\text{UJIVE},i} = Z_i'\hat{\pi}_{-i} + W_i'\hat{\psi}_{D,-i}$. (ii) adjust $\hat{Z}_{\text{UJIVE},i}$ for covariates by regressing D onto W, with ith observation removed to compute $\hat{\tau}_{-i}$, constructing $\hat{D}_{\text{UJIVE},i} = \hat{Z}_{\text{UJIVE},i} - W_i\hat{\tau}_{-i}$. Then use this as a single instrument, $\hat{\beta}_{\text{UJIVE}} = \hat{D}'_{\text{UJIVE}}Y/\hat{D}'_{\text{UJIVE}}D$.

Both jackknife estimators, UJIVE and IJIVE1 are consistent so long as:

1. The instruments are not too weak, in the sense that $\sqrt{K}/r_n \to 0$ (without this condition, no estimator can be consistent).

2. There aren't too many covariates: IJIVE1 requires $LK/r_n n \to 0$, while consistency of UJIVE only requires $\sqrt{L}/r_n \to 0$.

2. INFERENCE

Let us define $\pi_\Delta = \delta - \pi\beta$, $u_{\Delta,i} = u_{Yi} - u_{Di}\beta$, $\gamma = E[W_i W_i']^{-1}E[W_i(Y_i - D_i\beta)] = E[W_i W_i']^{-1}E[W_i Z_i]\pi_\Delta + \psi_Y - \psi_D\beta$, and $\epsilon_i = \tilde{Z}_i'\pi_\Delta + u_{\Delta,i}$. Then we can write the "structural" equation as

$$Y_i = D_i\beta + W_i'\gamma + \epsilon_i.$$

We saw in the previous lecture that the oracle estimator satisfied

$$\mathcal{V}_{1,n}^{-1/2}(\hat{\beta}^* - \beta) \Rightarrow \mathcal{N}(0,1), \qquad \mathcal{V}_{1,n} = \frac{E[\epsilon_i^2(\tilde{Z}_i'\pi)^2]}{r_n E[(\tilde{Z}_i'\pi)^2]}.$$

The variance $\mathcal{V}_{1,n}$ is the variance estimated by the conventional robust standard errors, such as those in Stata. These are also the correct standard errors for TSLS, UJIVE, or IJIVE1 if (i) there is no treatment effect heterogeneity, and (ii) $K/r_n \to 0$ for UJIVE, or IJIVE1, and $K^2/r_n \to 0$ for TSLS. If (i) fails, then, as discussed last time, we don't achieve the oracle variance, but instead the correct asymptotic variance is given by

$$\mathcal{V}_{2,n} = \frac{E[((\tilde{Z}_i'\pi_\Delta)u_{D,i} + \epsilon_i(\tilde{Z}_i'\pi))^2]}{r_n E[(\tilde{Z}_i'\pi)^2]}.$$

What about (ii)? If $K/r_n \to 0$, but $K^2/r_n \not\to 0$, then TSLS will be consistent, but asymptotically biased, and inference based on TSLS will be difficult. However, inference based on UJIVE or IJIVE1 is simple in this case: use an estimate of $\mathcal{V}_{2,n}$, or, if we want to impose homogeneity of treatment effects, $\mathcal{V}_{1,n}$. If K/r_n doesn't converge to zero, then we need to account for the presence of many instruments in the asymptotic variance formula (Evdokimov and Kolesár 2018, Theorem 5.4)

$$(\mathcal{V}_{2,n} + \mathcal{V}_{MI,n})^{-1/2}(\hat{\beta}_{\text{UJIVE}} - \beta) \Rightarrow \mathcal{N}(0,1),$$
$$\mathcal{V}_{MI,n} = \frac{1}{r_n^2}\sum_{i\neq j}(H^2_{\ddot{Z},ij}u^2_{\Delta,i}u^2_{D,j} + H^2_{\ddot{Z},ij}u_{\Delta,i}u_{D,i}\cdot u_{\Delta,j}u_{D,j}), \tag{3}$$

The same result holds with $\hat{\beta}_{\text{UJIVE}}$ replaced with $\hat{\beta}_{\text{IJIVE1}}$. This result generalizes that in Chao et al. (2012), who give a formula for JIVE1, assuming constant treatment effects and assuming that L is fixed with sample size.

Under homoskedastic errors, $\Omega(X_i) = \Omega = E[u_i u_i']$, this additional many instrument term in the variance formula simplifies to[3]

$$\mathcal{V}_{MI,n} = \frac{K}{r_n^2}(E[(u_{Yi} - u_{Di}\beta)^2]\cdot E[u_{Di}^2] + E[(u_{Yi} - u_{Di}\beta)u_{Di}]^2)(1 + o_p(1)).$$

[3] This follows since $\sum_{i,j} H^2_{\ddot{Z},ij} = \operatorname{tr}(H_{\ddot{Z}}) = K$, and $\sum_i H^2_{\ddot{Z},ii} \leq \max_i H_{\ddot{Z},ii}\sum_i H_{\ddot{Z},ii} = o(1)K$, provided that $\max_i H_{\ddot{Z},ii} \to 0$ (this is a familiar leverage condition).

For correct inference, in addition for accounting for the additional $\mathcal{V}_{MI,n}$ term in the asymptotic variance, we need to ensure that our estimate of $\mathcal{V}_{1,n} + \mathcal{V}_{MI,n}$, or, preferably, $\mathcal{V}_{2,n} + \mathcal{V}_{MI,n}$ is consistent even under the many instrument asymptotics. The issue is that under standard asymptotics, the usual estimates of π, β, and γ are all consistent, so we can just use plug-in estimators. However, simple plugin estimators are not consistent once we let K and L increase with the sample size.

To see the issue, suppose, for simplicity, that the reduced-form errors are homoskedastic, and that there is no treatment effect heterogeneity. Then

$$\mathcal{V}_{1,n} = \mathcal{V}_{2,n} = \frac{E[\epsilon_i^2]}{r_n} \approx \frac{E[\epsilon_i^2]}{\pi'\ddot{Z}'\ddot{Z}\pi}.$$

The natural estimator, used by default in Stata for inference based on TSLS is given by

$$\frac{\frac{1}{n}\sum_i \hat{\epsilon}_{i,\text{TSLS}}^2}{D'H_{\ddot{Z}}D}.$$

Using footnote 1, it follows that $E[D'H_{\ddot{Z}}D] = E[\pi\ddot{Z}'\ddot{Z}\pi] + KE[u_{Di}^2] \approx r_n + K\Omega_{DD}$. This leads to a downward bias in the standard errors. See Section 4 for an illustration.

The remedy to the bias in the denominator is simple: use the denominator of IJIVE1 or UJIVE to estimate r_n. Getting the numerator right is trickier—see Evdokimov and Kolesár (2018).

[[TODO: describe estimator of the many instrument term]].

3. FIRST STAGE F

In some papers, due to the dimensionality of the problem, researchers calculate the instrument $\hat{Z}_i$ manually, often as a leave-one-out prediction, effectively computing JIVE1 by hand. Often, they then forget that $\hat{Z}_i$ is a constructed instrument, and compute the first stage F statistic, as well as all other results, as if $K = 1$. This will of course overstate the actual instrument strength.

When using the (correctly computed) first-stage F statistic for diagnostics, remember that the $F > 10$ rule of thumb is a test of the hypothesis that the TSLS bias, relative to the bias of OLS exceeds 0.1. So if F is small, this is an indication that TSLS is biased. Since we have argued that it is good practice to use jackknife estimators precisely to eliminate the TSLS bias, a small F statistic is not necessarily a concern when UJIVE or IJIVE1 are used.

In particular, with homoskedastic errors,

$$E[F] = \frac{E[\hat{\pi}\ddot{Z}'\ddot{Z}\hat{\pi}]}{KE[u_{Di}^2]} = \frac{\pi E[\ddot{Z}'\ddot{Z}]\pi}{KE[u_{Di}^2]} + 1 \approx \frac{r_n}{KE[u_{Di}^2]} + 1.$$

Indeed, we have seen that if $r_n/K \to 0$, TSLS will be inconsistent; but jackknife estimators will remain consistent so long as $r_n/\sqrt{K} \to \infty$, and we can still do inference using jackknife estimators, provided we account for the additional many instrument term in the asymptotic variance.

4. SUMMARY AND ILLUSTRATION

If there are many instruments, in the sense that the number of instruments K is non-negligible relative to the sample size scaled by instrument strength r_n, the TSLS estimator will be biased. Instead, use IJIVE1 or UJIVE. To ensure reliable inference, the standard errors need to account for an additional many instrument term in the asymptotic variance, and avoid the downward bias that's present in the default standard errors estimator based on TSLS.

To illustrate these takeaways, we consider the dataset from Angrist and Krueger (1991), who use quarter of birth as in instrument in a regression of log earnings on education. Let us focus on the subsample of men born in 1930–39 from 1980 census. We consider four specifications. The first three correspond to a version of Table III, and to Tables V and VII in Angrist and Krueger (1991). In the first specification, we just use quarter of birth as an instrument. In the second, we interact it with 10 year of birth indicators, for a total of 30 instruments. In the third, we also interact it with state of birth indicators, for a total of $30 + 50 \times 3 = 180$ instruments. Finally, we consider a specification in which we use a triple interaction of year with state and with quarter of birth. We drop Alaska and Hawaii, which have some empty cells, or cells with only one observation (for which the jackknifing would be impossible). This reduces the number of observations by 324, and we're left with 329,185 observations. This leaves us with a total of $49 \times 10 \times 3 = 1470$ instruments.

The ratio K/n is very small in all of these specifications. However, the ratio of r_n to K is fairly small, as is the first-stage F, indicating that there will likely be a problem with TSLS. Indeed, we see that

1. The estimate gets close to OLS in Panel 4, in line with the theory above
2. The estimate of $\mathcal{V}_{1,n}$ and $\mathcal{V}_{2,n}$ appears to be biased downward in multiple panels, in line with the theory above.

Question 1. Do these two observations explain the findings of Bound, Jaeger, and Baker (1995)?

We also see that JIVE1 is rather erratic. This stems from the issue that the dimension of controls L is high relative to the strength of identification. In Panel 4, the JIVE1-based estimate of r_n is in fact negative—while, as discussed in Section 2.1, TSLS overestimates r_n, and hence underestimates the standard error, JIVE1 underestimates it, and hence

Table 1: Application to Angrist and Krueger (1991)

Estimator	Estimate	$\hat{\mathcal{V}}_1^{1/2}$	$\hat{\mathcal{V}}_2^{1/2}$	$\sqrt{\hat{\mathcal{V}}_2 + \hat{\mathcal{V}}_{MI}}$	$\hat{r}_n / K$
Panel A: OLS					
OLS	0.0670	0.0004			
Panel B: Instrument is QOB. $F = 34.0$					
TSLS	0.1026	0.0195	0.0198		366.0
JIVE1	0.1039	0.0203	0.0206	0.0209	351.6
UJIVE	0.1036	0.0201	0.0204	0.0207	355.2
Panel C: Instrument is $QOB \times YOB$. $F = 4.9$					
TSLS	0.0891	0.0162	0.0176		52.6
JIVE1	0.0959	0.0224	0.0244	0.0273	38.3
UJIVE	0.0938	0.0204	0.0222	0.0211	41.9
Panel D: Instrument is $QOB \times YOB + QOB \times SOB$, $F = 2.6$					
TSLS	0.0928	0.0097	0.0112		26.2
JIVE1	0.1211	0.0205	0.0243	0.0273	12.7
UJIVE	0.1096	0.0160	0.0187	0.0211	16.1
Panel E: Instrument is $QOB \times YOB \times SOB$, $F = 1.1$					
TSLS	0.0721	0.0049	0.0067		11.6
JIVE1	0.0320	0.0307	0.0425	0.0515	-1.9
UJIVE	0.1110	0.0397	0.0548	0.0663	1.4

tends to overestimate the standard errors. This explains why the JIVE1–based standard errors tend to be the largest.

In contrast, UJIVE is quite stable. The many-instrument term in the standard error formula doesn't matter in the first two specification, but increases the standard errors by 12% and 21%, respectively, in the last two specifications. Allowing for heterogeneity in the treatment effects has a similar effect.

REFERENCES

Ackerberg, Daniel A., and Paul J. Devereux. 2009. "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity." *Review of Economics and Statistics* 91, no. 2 (May): 351–362. https://doi.org/10.1162/rest.91.2.351.

Aizer, Anna, and Joseph J. Doyle Jr. 2015. "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges." *The Quarterly Journal of Economics* 130, no. 2 (May): 759–803. https://doi.org/10.1093/qje/qjv003.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics* 14 (1): 57–67. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1<57::AID-JAE501>3.0.CO;2-G.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (November): 979–1014. https://doi.org/10.2307/2937954.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62, no. 3 (May): 657–681. https://doi.org/10.2307/2951662.

Belloni, Alexandre, Daniel L. Chen, Victor Chernozhukov, and Christian B. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica* 80, no. 6 (November): 2369–2429. https://doi.org/10.3982/ECTA9626.

Blomquist, Soren, and Matz Dahlberg. 1999. "Small Sample Properties of LIML and Jackknife IV Estimators: Experiments with Weak Instruments." *Journal of Applied Econometrics* 14, no. 1 (January): 69–88. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1<69::AID-JAE521>3.0.CO;2-7.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90, no. 430 (June): 443–450. https://doi.org/10.1080/01621459.1995.10476536.

Chan, David, Matthew Gentzkow, and Chuan Yu. 2019. *Selection with Variation in Diagnostic Skill: Evidence from Radiologists.* Working Paper 26467. Cambridge, MA: National Bureau of Economic Research, November. https://doi.org/10.3386/w26467.

Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen. 2012. "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments." *Econometric Theory* 12, no. 1 (February): 42–86. https://doi.org/10.1017/S0266466611000120.

Evdokimov, Kirill, and Michal Kolesár. 2018. "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects." January. https://www.princeton.edu/~mkolesar/papers/het_iv.pdf.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review* 113, no. 1 (January): 253–277. https://doi.org/10.1257/aer.20201860.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133, no. 1 (February): 237–293. https://doi.org/10.1093/qje/qjx032.

Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96, no. 3 (May): 863–876. https://doi.org/10.1257/aer.96.3.863.

Kolesár, Michal. 2013. "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity." Working paper, Princeton University, November. https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf.

Nagar, Anirudh Lal. 1959. "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations." *Econometrica* 27, no. 4 (October): 575–595. https://doi.org/10.2307/1909352.

Phillips, G. D. A., and C. Hale. 1977. "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems." *International Economic Review* 18, no. 1 (February): 219–228. https://doi.org/10.2307/2525779.