ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Selection On Unobservables Designs:

Part 6, Weak Instruments

Weak instruments — that is to say, instruments that are only weakly correlated with the treatment of interest — pose a special set of problems. First, and most importantly, a weak first stage implies that any bias in the reduced form will be amplified in the IV estimate. This is true regardless of the number of instruments one uses. When using many weak instruments, however, a finite sample issue arises and 2SLS becomes biased towards the OLS estimate (conventional standard errors are also inaccurate). Though these issues have been known to some degree since the development of IV, they were brought to the attention of applied researchers by Bound, Jaeger, and Baker (1995) (henceforth BJB 1995).

1 Omitted Variables Bias

Consider a case with a single endogenous variable, d_i, one or more instruments, z_i, and no covariates.[1] We are interested in the causal relationship between d_i and y_i, summarized as

$$y_i = \alpha + \beta d_i + \varepsilon_i$$

We have an instrument z_i that we use to predict d_i

$$d_i = z_i\gamma + u_i$$

Consider the consistency of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$. For OLS,

$$\text{plim } \hat{\beta}_{OLS} = \frac{\text{Cov}(d_i, y_i)}{\text{Var}(d_i)} = \frac{\text{Cov}(d_i, \beta d_i + \varepsilon_i)}{\text{Var}(d_i)} = \beta + \frac{\sigma_{d\varepsilon}}{\sigma_{dd}}$$

[1] As per BJB 1995, the core results remained unchanged by the addition of covariates.

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{RF}}{\hat{\beta}_{FS}} = \frac{Cov(y,z)/Var(z)}{Cov(D,z)/Var(z)} = \frac{Cov(y,z)}{Cov(D,z)} = \frac{Cov(D\beta+\varepsilon,z)}{Cov(D,z)}$$

RF: $y = \beta_{RF} z + \varepsilon$ or $\beta_{RF}\hat{x} + \varepsilon$

FS: $x = \beta_{FS} z + \varepsilon$

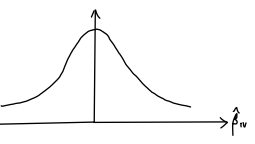
BJB Quarter of birth: Q1 vs Q2-Q4
0 " 1

- $\hat{\beta}_{FS} = 0.1$ Statistically strong (large t stat)
  but small economic relevance (Δ in z (quarter) doesn't induce much Δ in x)

- log(family income) 2.4% ↑ for families w/ kids Q2-Q4 vs Q1
  ↳ selection bias concern.

- Assume intergenerational income correlation = 0.4
  ⇒ Children born in Q2-Q4, we expect wages to be 0.4·0.024 = 0.01 (1%)
  ⇒ bias of 1% in RF
  ⇒ $\hat{\beta}_{IV}$ "bias" = $\frac{0.01}{0.1} = 0.1$ (inflated by weak first stage)
  ↳ bias may be large enough to explain nearly all of $\hat{\beta}_{IV}$ magnitude
  (make $\hat{\beta}_{IV}$ insignificant)

⟹ Want to show balance tables
- normally would show diff in means of covariates between z=1 and z=0
- run reg. of each x on z and report coef
- for z = {0,1}, " diff in means
- for continuous z, shows how strongly z affects x ... why do we want to see this?



- if true FS coef are 0 ⟹ ... portions of D are basically random and don't

	OLS	random instruments generated
coef		0.06
		0.014
		t>4

① Correct for bias in coef

② Correct for SEs

① 2SLS → assume normality in errors → MLE (LIML) will be different than OLS when we have many weak instruments
↳ approx. unbiased at median

The plim for 2SLS relies on the fact that $\hat{d}_i$ plims to $z_i\gamma$,

$$\text{plim } \hat{\beta}_{2SLS} = \frac{\text{Cov}(\hat{d}_i, y_i)}{\text{Var}(\hat{d}_i)} = \frac{\text{Cov}(\hat{d}_i, \beta d_i + \varepsilon_i)}{\text{Var}(\hat{d}_i)} = \frac{\text{Cov}(z_i\gamma, \beta(z_i\gamma + u_i) + \varepsilon_i)}{\text{Var}(\hat{d}_i)} = \beta + \frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{\hat{d}\hat{d}}}$$

If there is zero covariance between d and ε then OLS will consistently estimate β. If there is zero covariance between z and ε then 2SLS will consistently estimate β (note that 2SLS is never unbiased because it is a ratio of two random variables). What happens when these covariances are nonzero, however? Under what conditions will one estimator be more or less inconsistent than the other?

The ratio of the inconsistency in the IV estimator to the inconsistency in the OLS estimator is:

$$\frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}} \cdot \frac{\sigma_{dd}}{\sigma_{\hat{d}\hat{d}}} = \frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}} \cdot \frac{1}{R^2_{FS}}$$

R^2_{FS} is the R^2 of the first stage; the equality holds because $R^2_{FS} = SSR/SST = \sigma_{\hat{d}\hat{d}}/\sigma_{dd}$. If we had covariates in the model, the R^2_{FS} term would be the partial R^2 from the first stage, i.e., the R^2 from running d_i on z_i after the covariates have been partialled out from both.[2]

From the result above, we see that the relative inconsistency of IV vis a vis OLS depends on two quantities. First, it depends on the covariance of $\hat{d}_i$ and ε_i relative to the covariance of d_i and ε_i. If the covariance of the error term and $\hat{d}_i$ increases (relative to the covariance of the error term and d_i), then the inconsistency of IV increases — this is quite intuitive. More interestingly, the relative inconsistency of IV also depends on the inverse of the R^2 (or partial R^2, if you have covariates) of the first stage. Thus, if the first stage is weak (i.e., low R^2), any violation of the exclusion restriction will be amplified, and IV can become very inconsistent. A first stage (partial) R^2 of 0.1, for example, will inflate the ratio $\frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}}$ by a factor of 10. Except that things aren't quite that simple.

[2]With covariates in the model, the $\sigma_{\hat{d}\varepsilon}$ and $\sigma_{d\varepsilon}$ terms are also calculated after the covariates have been partialled out from d_i and z_i.

The complication is that $\sigma_{\hat{d}\varepsilon}$ is itself affected by the strength of the first stage. If the first stage is weak, then by definition the variance of $\hat{d}$ will be relatively low, and so the covariance $\sigma_{\hat{d}\varepsilon}$ will tend to be low as well. For tractability, and because it covers the preponderance of meaningful cases, suppose that z_i contains only one instrument. In that case:

$$\frac{\sigma_{\hat{d}\varepsilon}}{\sigma_{d\varepsilon}} \cdot \frac{1}{R^2_{FS}} = \frac{\text{Cov}(\gamma z, \varepsilon)}{\text{Cov}(d, \varepsilon)} \cdot \frac{\text{Var}(d)}{\text{Var}(\gamma z)} = \frac{\sigma_{z\varepsilon}/\sigma_z\sigma_\varepsilon}{\sigma_{d\varepsilon}/\sigma_d\sigma_\varepsilon} \cdot \frac{\sigma_d}{\sigma_{\gamma z}} = \frac{\rho_{z\varepsilon}}{\rho_{d\varepsilon}} \cdot \frac{1}{R_{FS}}$$

The last expression is more useful in the sense that it is expressed in terms that do not depend on the units of measurement for any of the variables in question. The second term, $\frac{1}{R_{FS}}$, confirms that a weak first stage does exacerbate the relative inconsistency of IV vis a vis OLS, but the degree of bias is not as strong as originally implied. With a first stage (partial) R^2 of 0.1, for example, IV will be less inconsistent than OLS as long as the correlation between the instrument, z_i, and the error term, ε_i, is approximately three times less than the correlation between d_i and ε_i. With a first stage (partial) R^2 of 0.01, however, the correlation between z_i and ε_i needs to be ten times less than the correlation between d_i and ε_i in order for IV to be preferable to OLS.

So, if the first stage is relatively weak, then you should think carefully about whether your exclusion restriction ($\text{Cov}(z, \varepsilon) = 0$) holds. Even a modest correlation between the instrument and the structural error term can make IV highly inconsistent if the first stage (partial) R^2 is low. This is true regardless of whether you have one instrument or many instruments.

BJB 1995 analyze the potential for omitted variables bias in Angrist and Krueger (1991) using the just-identified case. Quarter of birth is parameterized as a single indicator variable that equals zero if an individual is born in the first quarter and unity if an individual is born in the second through fourth quarters. With this parameterization, Angrist and Krueger report a first stage coefficient of 0.1 — people born in the first quarter have 0.1 years less education than those born in the second through fourth quarters. This is a fairly small effect, but the coefficient is highly significant since the sample numbers in the hundreds of thousands.

BJB note that the difference in mean log per capita family income for young children born in the second through fourth quarters versus those born in the first quarter is 0.024 — families of children born in the first quarter have per capita income that is about 2.4% lower than families of children born in the second through fourth quarters. Using an intergenerational correlation coefficient of 0.4 (the standard in the literature at that time — now it is estimated to be even higher), BJB infer that omitted factors might lead to a difference in mean log income of 0.01 between individuals born in the second through fourth quarters and individuals born in the first quarter. Though this differential is quite small, it is important to remember that the first stage is also very small, with a coefficient of 0.1. Thus the bias in the IV estimate will be 10 times the bias in the reduced form estimate — a reduced form bias of 0.01 translates to an IV bias of 0.10. Interestingly, this is very close to the return to education that Angrist and Krueger estimate using the quarter of birth instrument. I am not claiming that their estimate is necessarily wrong, but the relatively weak first stage does mean that the quarter of birth design is not quite as clean as it first appears.

1.1 Testing for Covariate Balance

In many IV applications, it is informative to test for balance of covariates across the instrument. This is similar to testing for covariate balance when stratifying on the propensity score — the idea is that if observable factors determining Y_i are balanced across the instrument, then unobservable factors are also likely to be balanced. In terms of testing statistical significance, it is fine to simply regress each covariate on the instrument and examine the significance of the coefficient on Z_i — we are just estimating the reduced form relationship between Z_i and each covariate here. If anything, this will yield a conservative test (in that it assumes the first stage is relatively strong). However, when interpreting the magnitude of this reduced form coefficient, it is important to keep in mind the weak instrument results above. Even a small bias in the reduced form can translate into a large bias in the IV.

2 Finite Sample Bias

The first issue — that a weak first stage can amplify any correlation between the instrument and the structural equation error term — is important. Nevertheless, for reasons that are unclear, much of the focus in the last decade regarding weak instruments has pertained to the second issue that BJB 1995 raise — finite sample bias.[3] This issue is, in my opinion, somewhat overblown, but it is important in a subset of cases, so you should be aware of it.

Recall that I said that IV/2SLS is consistent but not unbiased. This occurs because the IV estimator is the ratio of two random variables (the reduced form and the first stage), so we cannot compute the expectation (in fact, there is no guarantee that it even exists!). For an arbitrarily large sample, the bias of IV disappears, but of course no real sample is arbitrarily large. The problem is that the first stage is estimated with error — if we knew the true value of the first stage coefficient(s), we could plug these values in and IV/2SLS would be unbiased.

To fix ideas, consider a case in which there is no population first stage — the instruments have zero effect on d. We know that IV partitions the variation in d into two components: the variation in d induced by z and the complement of that variation. In this case, however, because z has no effect on d, the distinction between the two components is entirely arbitrary. In the population, the first stage will be zero, and the component of d that is correlated with z will contain nothing. In any finite sample, however, the first stage will not be zero.

If there is only one (weak) instrument, the IV estimate will be highly unstable. The IV coefficient is the ratio of the reduced form coefficient over the first stage coefficient. Since the first stage coefficient is centered at zero, the IV coefficient can easily realize very large positive or negative values — its distribution may be approximated by a Cauchy distribution. However, unless d is endogenous to a degree that is uncommon in empirical research, there is little chance that finite sample bias will be an issue with just one (or a small number of)

[3] I suspect the relative focus on the second issue rather than the first is due in part to the fact that the first issue is fairly straightforward — there is not much else to be said.

instrument(s). This is because the IV standard errors will be very large, and the researcher will correctly conclude that it is not possible to conduct precise statistical inference.

With many weak instruments, finite sample bias becomes problematic. With a large number of instruments, the amount of variation in d that $\hat{d}$ captures becomes nontrivial — we "overfit" the first stage, so to speak. If the instruments have no effect on d, however, the partitioning of d into the component determined by z and its complement is meaningless — the variation in d that we think is caused by z is no different than remaining variation in d that we throw away. It is thus unsurprising that with a large number of weak instruments, $\hat{\beta}_{2SLS}$ becomes biased towards $\hat{\beta}_{OLS}$.[4] To complicate matters, the 2SLS standard errors become biased downwards as well — $\hat{\beta}_{2SLS}$ is not as precisely estimated as it appears to be.

Angrist and Krueger (1995) demonstrate the many weak instruments problem by generating 180 instruments ((3 quarter of birth dummies $\times$ 10 year of birth dummies = 30 dummies) plus (3 quarter of birth dummies $\times$ 50 state of birth dummies = 150 dummies)) and then replacing the actual quarter of birth with random draws from a discrete uniform distribution with four points of support. Recall that the IV estimate for Angrist and Krueger (1991) is approximately 0.10 (this is using a small number of instruments) while the OLS estimate is approximately 0.07. The IV estimate produced by the large number of randomly generated instruments is 0.06 with a standard error of 0.014 — very close to the OLS estimate and statistically significant. An inattentive researcher might mistakenly believe that this estimate is informative.

What can be done to address the finite sample bias issue when working with many weak instruments?[5] The simplest solution is to simply reduce the number of instruments being used. Since it is difficult to find one good instrument, let alone many good instruments, the "many weak instruments" issue often occurs when the researcher interacts the primary instrument with a number of other covariates. In these cases it is straightforward to eliminate

[4] To take an extreme case, if you had as many instruments as observations, you could fit d perfectly, and $\hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$. Things become problematic long before that happens, however.

[5] It is not exactly clear what constitutes a weak first stage. Staiger and Stock (1997) recommend caution when dealing with first stage F-statistics of less than 10.

the interaction terms — Angrist and Krueger (1991), for example, parameterize the QOB instrument as a single variable (first quarter versus second through fourth quarters) and as 180 different variables.

If it is not possible to reduce the number of instruments, there are two issues to be addressed. First, the bias towards the OLS estimate must be corrected. Second, the standard errors must be corrected. An easy way to address the first issue is to use the Limited Information Maximum Likelihood Estimator (LIML). LIML is derived by assuming that the residuals in the structural and first stage equations are normally distributed and then estimating β and γ via maximum likelihood methods.[6] Angrist and Krueger (2001) note that LIML is approximately unbiased in that the median of its sampling distribution is often close to the parameter being estimated.[7] Although this does not completely eliminate finite sample bias, LIML generally performs better than 2SLS in cases with many instruments. Conventional LIML standard errors can still be too small, however. Imbens suggests implementing a correction derived in Bekker (1994). Multiply the conventional LIML standard errors by:

$$1 + \frac{K/N}{1 - K/N} \cdot (\sum_{i=1}^{N} (\tilde{z}_i \gamma)^2 / N)^{-1} \cdot (\begin{bmatrix} 1 & \beta_1 \end{bmatrix} \Omega^{-1} \begin{bmatrix} 1 \\ \beta_1 \end{bmatrix})^{-1}$$

where K is the number of instruments, N is the sample size, $\tilde{z}_i$ is a $1 \times K$ row vector of demeaned instruments, γ is a $K \times 1$ column vector of first stage coefficients, β_1 is the coefficient on d in the structural equation, and Ω is the variance-covariance matrix of the reduced form residual and the first stage residual (the variance of the reduced form residual, v_i, is the upper left element, and the variance of the first stage residual, u_i, is the bottom right element). In practice, we replace these population coefficients and moments with their

[6] In just-identified cases, LIML is numerically identical to 2SLS/IV.

[7] An alternative is to use a split-sample IV estimator that estimates the first stage on a different sample than the second stage. In practice, you would estimate the vector of first stage coefficients, $\hat{\gamma}$, using the first sample. Then you would apply this coefficient vector to the instruments in the second sample to construct the fitted value of the treatment variable. This solves the first stage "over-fitting" problem by essentially forcing the first stage to do an out-of-sample forecast. In practice, however, it's often easier to estimate LIML than it is to estimate some of the SSIV estimators, and there doesn't appear to be a strong consensus among econometricians in favor of the latter over the former.

sample counterparts. $\hat{\beta}_1$ comes from the LIML estimate of β_1, and $\hat{\gamma}$, $\hat{v}_i$, and $\hat{u}_i$ come from OLS estimates of the first stage and reduced form equations.

Intuitively, the standard error adjustment increases in K because additional instruments make it more likely that we will "overfit" the first stage. It decreases in the second term, which is the inverse of the regression sum of squares for the first stage, because a larger regression sum of squares implies a stronger set of instruments. Finally, it increases in (the magnitude of) Ω because Ω determines the degree of endogeneity (recall that the first stage residual is the potentially endogenous part of d, while the reduced form residual is the unexplained part of y, so a high covariance between the two components implies a high degree of endogeneity).

3 Summary

In conclusion, "weak" instruments pose two separate problems.

1. The first stage may be weak in an economic sense — that is, the instrument may explain only a tiny fraction of the variation in the treatment. We typically diagnose this by examining the first-stage coefficients and/or the partial R^2 of the first-stage regression. What constitutes an "economically weak" first stage varies by context. If the first stage is economically weak, then even small violations of the exclusion restriction can cause large biases, regardless of the number of instruments.

2. The first stage may be weak in a statistical sense. This only matters when you have multiple instruments (especially "many" instruments), and we typically diagnose it by examining the first-stage F-statistic on the instruments (not on the instruments plus covariates). A common rule of thumb is that the F-stat should be greater than 10. In practice this means that if your first-stage F-stat on the instruments is less than 10 you're in big trouble, if it's between 10 and 20 you're in the zone of concern, and if it's over 20 you're probably okay. However, if your data feature heteroskedasticity

or clustering (as is often the case), the vanilla F-stat is not appropriate. Instead you should use, for example, an "effective" F-stat, proposed by Olea and Pflueger (2013). $F_{eff} = \frac{1}{N}\frac{\mathbf{D'ZZ'D}}{tr(\hat{\mathbf{V}})}$, where $\hat{\mathbf{V}}$ is the robust or clustered first-stage covariance matrix, and variables are residualized with respect to covariates if necessary.

4 Additional References

Angrist, J. and A. Krueger. "Split-Sample Instrumental Variables Estimates of the Return to Schooling." *Journal of Business and Economic Statistics*, 1995, 33, 225-235.

Bekker, P. "Alternative Approximations to the Distribution of Instrumental Variables Estimators." *Econometrica*, 1994, 62, 657-681.

Olea, J. and C. Pflueger. "A Robust Test for Weak Instruments." *Journal of Business and Economic Statistics*, 2013, 31, 358-369.

Staiger, D. and J. Stock. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 1997, 65, 557-586.

Want to document the other pathways that the instrument might be related to the outcome, and then test those pathways and show that they aren't significant relationships.

Multiple treatments:
Anderson & Piscke in mostly harmless show modification of first stage to address multiple treatments

Setup: D_1, D_2, Z_1, Z_2

Anderson & Piske: $\hat{D}_2 = \hat{\gamma}_{20} + \hat{\gamma}_{21} Z_1 + \hat{\gamma}_{22} Z_2$

- Residualize D_1 wrt $\hat{D}_2$, then est FS with $\tilde{D}_1$ on Z_1, Z_2

→ deals with multicollinearity