Introducing Two-Stage Least Squares

In the previous chapter we defined the IV regression estimand as

$$\underline{\boldsymbol{\beta}} = E[\underline{\boldsymbol{\Pi}}\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']^{-1}E[\underline{\boldsymbol{\Pi}}\underline{\boldsymbol{Z}}_i Y_i]. \tag{1}$$

Here Y_i is a (scalar) outcome; $\underline{\boldsymbol{X}}_i = [\boldsymbol{X}_i', \boldsymbol{W}_i']'$ is a $(J+K)\times 1$ vector stacking the $J\times 1$ vector of endogenous variables $\boldsymbol{X}_i$ and the $K\times 1$ vector of controls $\boldsymbol{W}_i$; $\underline{\boldsymbol{Z}}_i = [\boldsymbol{Z}_i', \boldsymbol{W}_i']'$ is an $(L+K)\times 1$ vector stacking the $L\times 1$ vector of exogenous instruments and the controls, and $\underline{\boldsymbol{\Pi}} = E[\underline{\boldsymbol{X}}_i\underline{\boldsymbol{Z}}_i'][\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i']^{-1}$ is an $(J+K)\times(L+K)$ matrix collecting the first-stage coefficients from regressing each element of $\underline{\boldsymbol{X}}_i$ on $\underline{\boldsymbol{Z}}_i$. We previously discussed how such estimands may be used to overcome different endogeneity challenges and identify parameters of interest under two key assumptions: instrument *validity* and *relevance*.

As with linear (non-instrumented) regression, the IV regression estimand is a relatively simple function of simple moments: in particular, $E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i']$, $E[\underline{\boldsymbol{X}}_i\underline{\boldsymbol{Z}}_i']$, and $E[\underline{\boldsymbol{Z}}_i Y_i]$. Just as our ordinary least squares estimator "plugged in" sample analogs of population moments into the regression formula, so too can we estimate IV regressions as

$$\widehat{\underline{\boldsymbol{\beta}}} = \left(\frac{1}{N}\sum_i \widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i'\right)^{-1}\frac{1}{N}\sum_i \widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}_i Y_i \tag{2}$$

$$\text{where } \widehat{\underline{\boldsymbol{\Pi}}} = \frac{1}{N}\sum_i \underline{\boldsymbol{X}}_i\underline{\boldsymbol{Z}}_i'\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i'\right)^{-1}. \tag{3}$$

We call $\widehat{\underline{\boldsymbol{\beta}}}$ the *two-stage least squares* (2SLS) estimator of $\underline{\boldsymbol{\beta}}$.

As with OLS, we can use matrix notation to more compactly write $\widehat{\underline{\boldsymbol{\beta}}}$. Let $\underline{\boldsymbol{Z}} = [\underline{\boldsymbol{Z}}_1, \ldots, \underline{\boldsymbol{Z}}_N]'$ be the $N\times(L+K)$ matrix of instrument (and control) observations, let $\underline{\boldsymbol{X}} = [\underline{\boldsymbol{X}}_1, \ldots, \underline{\boldsymbol{X}}_N]'$ be the $N\times(J+K)$ matrix of endogenous variable (and control) observations, and as before let $\boldsymbol{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N]'$ be the $N\times 1$ vector of outcome observations. Then, similar to before, we can write

$$\widehat{\underline{\boldsymbol{\Pi}}} = \underline{\boldsymbol{X}}'\underline{\boldsymbol{Z}}\left(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\right)^{-1}. \tag{4}$$

$$\text{and } \widehat{\underline{\boldsymbol{\beta}}} = \left(\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\right)^{-1}\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\boldsymbol{Y}. \tag{5}$$

To see where the 2SLS estimator gets its name, let's multiply and divide inside the parentheses of $\widehat{\underline{\boldsymbol{\beta}}}$ by $\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}$:

$$\begin{aligned}\widehat{\underline{\boldsymbol{\beta}}} &= \left(\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\left(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\right)^{-1}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\right)^{-1}\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\boldsymbol{Y} \\ &= \left(\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\widehat{\underline{\boldsymbol{\Pi}}}'\right)^{-1}\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\boldsymbol{Y} \\ &= \left(\widehat{\underline{\boldsymbol{X}}}'\widehat{\underline{\boldsymbol{X}}}\right)^{-1}\widehat{\underline{\boldsymbol{X}}}'\boldsymbol{Y}\end{aligned} \tag{6}$$

where we define $\widehat{\underline{\boldsymbol{X}}} = \underline{\boldsymbol{Z}}\widehat{\underline{\boldsymbol{\Pi}}}'$ as the first stage fitted values. This shows that the 2SLS estimator can be written in terms of two OLS estimators. The first-stage OLS regresses $\underline{\boldsymbol{X}}_i$ on $\underline{\boldsymbol{Z}}_i$ in the sample and takes its fitted values. The second-stage OLS regresses $\boldsymbol{Y}_i$ on these fitted values in the sample. This parallels how we first encountered the population IV regression, as a series of two population regressions. Intuitively, 2SLS isolates exogenous variation in $\underline{\boldsymbol{X}}$ by projecting it on the exogenous $\underline{\boldsymbol{Z}}$ and using only this component of variation.

Another way to look at 2SLS is via the estimated reduced-form and first-stage regressions. Note that by multiplying both pieces of $\widehat{\underline{\boldsymbol{\beta}}}$ by $\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}})^{-1} = \boldsymbol{I}$ we can also write

$$\begin{aligned}\widehat{\underline{\boldsymbol{\beta}}} &= \left(\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}})^{-1}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\right)^{-1}\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}(\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}})^{-1}\underline{\boldsymbol{Z}}'\boldsymbol{Y} \\ &= \left(\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\widehat{\underline{\boldsymbol{\Pi}}}'\right)^{-1}\widehat{\underline{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\hat{\underline{\boldsymbol{\rho}}},\end{aligned} \tag{7}$$

where $\underline{\hat{\boldsymbol{\rho}}} = (\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}})^{-1}\underline{\boldsymbol{Z}}'\boldsymbol{Y}$ is an OLS estimate of the reduced form regression, of Y_i on $\underline{\boldsymbol{Z}}$. This way of writing $\underline{\widehat{\boldsymbol{\beta}}}$ shows that 2SLS can also be seen as a $\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}$-weighted OLS regression, of $\underline{\hat{\boldsymbol{\rho}}}$ on $\underline{\widehat{\boldsymbol{\Pi}}}'$. Make sure you understand what I mean by that: note $\underline{\hat{\boldsymbol{\rho}}}$ and $\underline{\widehat{\boldsymbol{\Pi}}}'$ are not matrices of N observations running along the rows, but instead they have L rows (one for each instrument). What 2SLS is doing is projecting the reduced-form relationship between Y_i and $\underline{\boldsymbol{Z}}'$, across different instruments, on the first-stage relationship between $\underline{\boldsymbol{X}}$ and $\underline{\boldsymbol{Z}}'$. The simplest version of this is when $\underline{\widehat{\boldsymbol{\Pi}}}$ is square (and invertible), as in the just-identified case. Then equation (8) reduces to

$$\begin{aligned}\underline{\widehat{\boldsymbol{\beta}}} &= \left(\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\underline{\widehat{\boldsymbol{\Pi}}}'\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\underline{\hat{\boldsymbol{\rho}}} \\ &= \underline{\widehat{\boldsymbol{\Pi}}}'^{-1}\left(\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{Z}}\underline{\hat{\boldsymbol{\rho}}} \\ &= \underline{\widehat{\boldsymbol{\Pi}}}'^{-1}\underline{\hat{\boldsymbol{\rho}}}. \end{aligned} \tag{8}$$

Loosly, this is "reduced form over first stage" — it is exactly this in the just-identified case with only one instrument/endogenous variable. If, for example, we regress a Y_i on a scalar X_i and instrument with a binary Z_i, we obtain the so-called Wald IV estimator

$$\hat{\beta} = \frac{\widehat{Cov}(Z_i, Y_i)/\widehat{Var}(Z_i)}{\widehat{Cov}(Z_i, X_i)/\widehat{Var}(Z_i)} = \frac{\widehat{E}[Y_i \mid Z_i = 1] - \widehat{E}[Y_i \mid Z_i = 0]}{\widehat{E}[X_i \mid Z_i = 1] - \widehat{E}[X_i \mid Z_i = 0]}, \tag{9}$$

where we use the fact that Z_i is binary to arrive at the second equality.[1]

As with OLS, the Frisch-Waugh-Lovell theorem can help us make sense of "big" 2SLS regressions, with many controls. Typically the most useful way to apply FWL is via the reduced-form and first-stage regressions. If we estimate

$$Y_i = \boldsymbol{Z}_i'\widehat{\boldsymbol{\rho}} + \boldsymbol{W}_i'\widehat{\boldsymbol{\kappa}} + \hat{\nu}_i \tag{10}$$

$$\boldsymbol{X}_i = \widehat{\boldsymbol{\Pi}}\boldsymbol{Z}_i + \widehat{\boldsymbol{\mu}}\boldsymbol{W}_i + \hat{\boldsymbol{\eta}}_i \tag{11}$$

then $\underline{\hat{\boldsymbol{\rho}}} = [\hat{\boldsymbol{\rho}}', \hat{\boldsymbol{\kappa}}']'$ and

$$\underline{\widehat{\boldsymbol{\Pi}}} = \begin{bmatrix} \widehat{\boldsymbol{\Pi}} & \hat{\boldsymbol{\mu}} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}, \tag{12}$$

using the definition of $\underline{\widehat{\boldsymbol{\Pi}}}$ and the fact that the first-stage OLS regressions of $\boldsymbol{W}_i$ on $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$ will put a zero coefficient on the former and the identity matrix on the latter. When we are just-identified, such that $\underline{\widehat{\boldsymbol{\beta}}} = \underline{\widehat{\boldsymbol{\Pi}}}^{-1}\underline{\hat{\boldsymbol{\rho}}}$, it follows from the partitioned inverse formula that $\widehat{\boldsymbol{\beta}}$ (the 2SLS coefficient on $\boldsymbol{X}_i$ alone) equals $\widehat{\Pi}^{-1}\widehat{\boldsymbol{\rho}}$; thus, we can study $\widehat{\boldsymbol{\beta}}$ just from the reduced-form OLS coefficient on $\boldsymbol{Z}_i$ while controlling for $\boldsymbol{W}_i$ ($\widehat{\boldsymbol{\rho}}$) and the first-stage OLS coefficient on $\boldsymbol{Z}_i$ while controlling for $\boldsymbol{W}_i$ ($\widehat{\Pi}$). It follows by the Frisch-Waugh-Lovell theorem that we can study the partialled-out reduced form and first stage to learn about the IV estimate; you'll see an example of this on the last question of PS #4.

2SLS Asymptotics

As with OLS, it is straightforward to show that 2SLS is $\sqrt{N}$-consistent under fairly weak conditions. By substituting in the population regression $Y_i = \underline{\boldsymbol{X}}_i'\underline{\boldsymbol{\beta}} + U_i$ we have, in matrix form,

$$\begin{aligned}\underline{\widehat{\boldsymbol{\beta}}} &= \left(\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\boldsymbol{Y} \\ &= \left(\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\left(\underline{\boldsymbol{X}}\underline{\boldsymbol{\beta}} + \boldsymbol{U}\right) \\ &= \underline{\boldsymbol{\beta}} + \left(\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\underline{\boldsymbol{X}}\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\underline{\boldsymbol{Z}}'\boldsymbol{U}. \end{aligned} \tag{13}$$

[1] The Wald IV estimator is due to Abraham Wald, who also gave us the Wald test studied in Chapter 6. It was originally developed (in 1940) to address measurement error concerns, but has since taken on a broader role in modern IV methods.

It thus remains for us to characterize the asymptotic distribution of

$$\begin{aligned}\sqrt{N}\left(\underline{\widehat{\boldsymbol{\beta}}}-\underline{\boldsymbol{\beta}}\right) &= \sqrt{N}\left(\underline{\widehat{\boldsymbol{\Pi}}}\boldsymbol{Z}'\underline{\boldsymbol{X}}\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\boldsymbol{Z}'\boldsymbol{U} \\ &= \left(\underline{\widehat{\boldsymbol{\Pi}}}\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i'\right)\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\left(\sqrt{N}\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i U_i\right), \end{aligned} \tag{14}$$

where again $\underline{\widehat{\boldsymbol{\Pi}}} = \frac{1}{N}\sum_i \underline{\boldsymbol{X}}_i\underline{\boldsymbol{Z}}_i'\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i'\right)^{-1}$. In *iid* data, assuming finite $Var(\underline{\boldsymbol{Z}}_{ij}\underline{\boldsymbol{Z}}_{ik})$ and $Var(\underline{\boldsymbol{X}}_{ij}\underline{\boldsymbol{Z}}_{ik})$ for all (j,k) we know that $\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i' \xrightarrow{p} E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i']$ and $\frac{1}{N}\sum_i \underline{\boldsymbol{X}}_i\underline{\boldsymbol{Z}}_i' \xrightarrow{p} E[\underline{\boldsymbol{X}}_i\underline{\boldsymbol{Z}}_i']$ by the law of large numbers. We further know that $\underline{\widehat{\boldsymbol{\Pi}}} \xrightarrow{p} \underline{\boldsymbol{\Pi}} = E[\underline{\boldsymbol{X}}_i\underline{\boldsymbol{Z}}_i']E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i']^{-1}$ by the continuous mapping theorem. Moreover, so long as $Var(\underline{\boldsymbol{Z}}_{ik}U_i)$ is finite for all k we have by the central limit theorem that $\sqrt{N}\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i U_i \Rightarrow \mathrm{N}(0, Var(\underline{\boldsymbol{Z}}_i U_i))$ since, by definition, $E[\underline{\boldsymbol{Z}}_i U_i] = 0$. It follows by Slutsky's theorem that

$$\begin{aligned}\sqrt{N}\left(\underline{\widehat{\boldsymbol{\beta}}}-\underline{\boldsymbol{\beta}}\right) &\Rightarrow \mathrm{N}(0,\boldsymbol{\Sigma}) \\ \text{where } \boldsymbol{\Sigma} &= \left(\underline{\boldsymbol{\Pi}}E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']\right)^{-1}\underline{\boldsymbol{\Pi}}Var(\underline{\boldsymbol{Z}}_i U_i)\underline{\boldsymbol{\Pi}}'\left(\underline{\boldsymbol{\Pi}}E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']\right)^{-1}, \end{aligned} \tag{15}$$

similar to how we derived the asymptotic variance of OLS. Paralleling our Eicker-Hubert-White "robust" standard error calculation, we can moreover consistently estimate the asymptotic variance by

$$\widehat{\boldsymbol{\Sigma}} = \left(\underline{\widehat{\boldsymbol{\Pi}}}\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i'\right)\right)^{-1}\underline{\widehat{\boldsymbol{\Pi}}}\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i'\widehat{U}_i^2\right)\underline{\widehat{\boldsymbol{\Pi}}}'\left(\underline{\widehat{\boldsymbol{\Pi}}}\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i'\right)\right)^{-1}, \tag{16}$$

where $\widehat{U}_i = Y_i - \underline{\boldsymbol{X}}'\underline{\widehat{\boldsymbol{\beta}}}$ consistently estimate the IV residual U_i. Notice how this gives the OLS formula as a special case, where $\underline{\boldsymbol{X}}_i = \underline{\boldsymbol{Z}}_i$ and so $\underline{\widehat{\boldsymbol{\Pi}}} = \boldsymbol{I}$. As before we can extend this derivation and variance estimator to non-*iid* (e.g. clustered) data, and can derive a homoskedastic special case that applies when $Var(U_i \mid \underline{\boldsymbol{Z}}_i) = \sigma^2$. In this case $Var(\underline{\boldsymbol{Z}}_i U_i) = \sigma^2 E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i']$ so we can simplify

$$\begin{aligned}\boldsymbol{\Sigma} &= \left(\underline{\boldsymbol{\Pi}}E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']\right)^{-1}\underline{\boldsymbol{\Pi}}\sigma^2 E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{Z}}_i']\underline{\boldsymbol{\Pi}}'\left(\underline{\boldsymbol{\Pi}}E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']\right)^{-1}, \\ &= \left(\underline{\boldsymbol{\Pi}}E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']\right)^{-1}\sigma^2 \end{aligned} \tag{17}$$

and estimate this accordingly. As with OLS, the homoskedastic estimator is not "robust" to heteroskedasticity and is thus not to be used in most applications, though it can be useful for building intuition. A different special case is one in which we are just identified ($L = K$), in which case

$$\boldsymbol{\Sigma} = E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']^{-1}Var(\underline{\boldsymbol{Z}}_i U_i)E[\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']^{-1} \tag{18}$$

does not depend on $\underline{\boldsymbol{\Pi}}$.

As with OLS, it can be useful to refer to the "simple" IV case with one instrument, one endogenous variable, and controls. As you'd expect, we get by solving out the general 2SLS formula a slope coefficient estimate of

$$\widehat{\beta} = \frac{\widehat{Cov}(Z_i, Y_i)}{\widehat{Cov}(Z_i, X_i)} \tag{19}$$

and by working through the general algebra on the asymptotic variance you can show $\sqrt{N}(\widehat{\beta} - \beta) \Rightarrow \mathrm{N}(0, \Sigma)$ where

$$\Sigma = \frac{Var((Z_i - E[Z_i])U_i)}{Cov(Z_i, X_i)^2} \tag{20}$$

which we estimate by

$$\widehat{\Sigma} = \frac{\frac{1}{N}\sum_i (Z_i - \bar{Z})^2\widehat{U}_i^2}{\left(\frac{1}{N}\sum_i (Z_i - \bar{Z})(X_i - \bar{X})\right)^2}. \tag{21}$$

Again, it is useful to see how this nests the bivariate OLS estimator and variance we saw before, when $X_i = Z_i$.

Although it's temping to run 2SLS regressions in two steps (given the name) you shouldn't actually do this in practice. The coefficient you get from first regressing (say) X_i on Z_i and then regressing Y_i on the fitted values $\widehat{\Pi} X_i$ will be numerically equivalent to the 2SLS coefficient, but the standard errors that Stata or R give you from this second-stage regression will generally be incorrect. This is because the fitted values regression has an "extra" residual term:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + U_i \\ &= \alpha + \beta \Pi X_i + (\beta \eta_i + U_i) \end{aligned} \tag{22}$$

where η_i denotes the first-stage residual as before. Stata will compute a standard error on β that uses both $\beta\eta_i$ and U_i in the residual, but the correct IV residual is just U_i. Thus you should never run 2SLS regressions "by hand" but instead use built-in commands that uses the correct second-stage residual: in Stata, this is the *ivregress 2sls* command, but you can also use the user-written (and somewhat more flexible) *ivreg2* command. Don't forget when you do to specify heteroskedastic-robust or clustered standard errors, since again Stata will by default use the (non-robust) variance estimator. For estimating IV regressions with high dimensional fixed effects, such as those involve difference-in-difference-type arguments, you can use the *ivreghdfe* command.

Weak and Many Instruments

Recall that in our Chapter 5 discussion of OLS we touched on the issue of multicollinearity: effectively, that the standard errors we get on a given coefficient X_i can be large when the residual variation $\tilde{X}_i$ after partialling out some control vector $\boldsymbol{W}_i$ is small. I said then that this was not necessarily a "problem," per se, just a reflection of the fact that we didn't have enough (conditional) explanatory power in X_i to make precise inferences. So long as nothing else went wrong in the OLS regression (such as perfect collinearity), it would remain consistent and asymptotically normal despite moderate multicollinearity.

It turns out there is an analogous, but more pernicious problem in 2SLS called *weak instruments*. In the case of a single instrument and treatment, the weak instrument problem arises when $Cov(\tilde{Z}_i, \tilde{X}_i)$ is small where again $\tilde{Z}_i$ and $\tilde{X}_i$ denote the residuals from regressing Z_i or X_i respectively on the included control vector. We can see from the previous section (and an extension of the Frisch-Waugh-Lovell theorem) that a small $Cov(\tilde{Z}_i, \tilde{X}_i)$ will generally lead to a large asymptotic variance of the 2SLS estimator—just as how a small $Var(\tilde{X}_i)$ led to a large asymptotic variance of OLS in Chapter 5. But it turns out that in 2SLS the problem is worse: a small $Cov(\tilde{Z}_i, \tilde{X}_i)$ can cause 2SLS to be *biased.*

The claim that "weak instruments" (formally, a small $\Pi = Cov(Z_i, X_i)/Var(Z_i)$ in the simple IV case, where we drop the controls for notational convenience) can cause bias might come as a surprise initially. After all, didn't we just prove above that 2SLS is consistent under weak conditions? The way to reconcile these two things is to recall why we use asymptotics in the first place. In reality we never have an increasing sample size N and wait until we've collected "enough" observations to conduct inference. In any given study we only have the data we have, and we use asymptotics to approximate the distribution of our estimator in this (finite) sample. The point is that when Π is small we may not be able to estimate it well in the sample, and in this case the 2SLS estimator which uses an estimate of Π to scale the reduced-form regression may be badly behaved. So-called *weak instrument asymptotics* are designed to capture this scenario and approximate the bias that results from such bad behavior, in contrast to the *strong instrument asymptotics* used above to characterize the asymptotic distribution of 2SLS when we can estimate Π arbitrarily well.

To formalize the weak IV problem, we thus need an alternative asymptotic "sequence" which allows us to isolate the problem of not being able to estimate a small Π reliably. Such a sequence will generally involve $\Pi \to 0$ as $N \to \infty$, such that even in a large sample size we have difficulty precisely estimating Π. It is important to recognize that the "assumption" of $\Pi \to 0$ in such sequences is just a mathematical trick to

achieve a good finite-sample approximation to the data at hand. Just as we never actually see an increase in our sample size N, we never actually have a decrease in the first-stage coefficient Π. Instead, we use this mathematical sequence to figure out how our estimator is behaving when N is moderately large and Π is moderately small.

Concretely, let's consider a simple 2SLS regression of Y_i on X_i instrumenting by Z_i where in the population regression the slope coefficient is $\beta = 0$. Let's further normalize the sample variance of Z_i to one to keep things simple. The 2SLS estimator of β is then

$$\widehat{\beta} = \frac{\widehat{Cov}(Z_i, Y_i)}{\widehat{Cov}(Z_i, X_i)} = \frac{\widehat{Cov}(Z_i, U_i)}{\Pi + \widehat{Cov}(Z_i, \eta_i)} = \frac{\sqrt{N}\widehat{Cov}(Z_i, U_i)}{\sqrt{N}\Pi + \sqrt{N}\widehat{Cov}(Z_i, \eta_i)}, \tag{23}$$

where in the first equality we substitute in the population second stage of $Y_i = \alpha + \beta X_i + U_i$ (recalling $\beta = 0$) and the population first stage of $X_i = \gamma + \Pi Z_i + \eta_i$, and in the second equality we multiply the top and bottom of the expression by $\sqrt{N}$. By the central limit theorem, under our usual conditions, $\sqrt{N}\widehat{Cov}(Z_i, U_i)$ and $\sqrt{N}\widehat{Cov}(Z_i, \eta_i)$ will have approximately normal distributions in large samples N that are centered around zero. Under "strong" instrument asymptotics (where Π is a fixed number) $\sqrt{N}\Pi \to \infty$ as $N \to \infty$ such that the denominator tends to infinity and $\widehat{\beta}$ tends to zero as we'd expect. But what if we consider a model where Π is "small relative to sampling variation"? In particular let's model $\Pi = \pi/\sqrt{N}$ for some fixed π. This ensures $\Pi \to 0$ as $N \to \infty$, and per equation (23) we now have $\widehat{\beta}$ converging not to zero but to the ratio of two normal random variables! This distribution has a particular name: a *Cauchy* distribution. It turns out it's a gnarly one, with very thick tails and no finite moments; interestingly, you can show that in this case the Cauchy distribution that (24) becomes is centered around the OLS estimand, $Cov(X_i, Y_i)/Var(X_i)$. In the course slides, I illustrate this distribution and "bias" towards OLS.

What do we learn from this example? Again, we never literally have a scenario where $\sqrt{N}\Pi \to \pi$ as $N \to \infty$. But we may have a scenario where the first stage coefficient Π is small and thus difficult to estimate even though the sample size N is large. What this example tells us is that in those scenarios our IV coefficient may be biased, and perhaps badly so, in the direction of OLS. In practice, to guard against this possibility, is it thus important to worry about instrument "strength" – you want to show that the relationship between the instruments and the endogenous variables is large enough that the kind of bias illustrated in the above example is negligible. Typically, such strength is gauged by the first stage F-statistic*:* i.e., the Wald test statistic for the joint null hypothesis that all of the elements of $\mathbf{\Pi}$ are zero. This is straightforward enough to do by hand, though most built-in IV procedures have an option to report first-stage F-statistics automatically (and *ivreg2* does it without asking).

You might be wondering: how big of a first-stage F-statistic do I need to not worry about weak IV? The conventional advice, due to a now fairly old paper by Staiger and Stock (1997), is a F-statistic cutoff of 10. This comes from a particular setting and series of monte carlo simulations on weak instrument bias in a homoskedastic setting, and is now a bit out of date.[2] Still, it remains in many cases the practical "rule of thumb."

What to do if your first-stage F-statistic is below 10? There aren't many simple solutions, besides changing your estimator or estimation procedure. You could of course try to find a stronger instrument (though this is clearly easier said than done!). You might also think about what controls you're including and whether including more or fewer may help with instrument strength. Just as in the OLS/multicollinearity case, including controls that are correlated with Z_i tend to reduce first-stage strength (all else equal) though this can increase the plausibility of instrument validity to the extent they account for systematic relationships between the instruments and observables. There are of course more complicated ways to address weak instruments bias without changing the instrument or control set, which you'll encounter in future classes.

A conceptually related (but distinct) issue with 2SLS estimation is the problem of *many instruments*, some (but not all) of which are weak. This issue was first pointed out by Bound et al. (1995) in the context

[2]For more recent analyses, see Montiel-Olea and Pflugger (2013) and Andrews, Stock, and Sun (2019).

of Angrist and Krueger (1991), who in some specifications used a ton of instruments based on interactions of quarter of birth and other observables (specifically an individual's year and state of birth). Bound et al. (1995) showed by simulating a bunch of "placebo" instruments that had no relationship to the years-of-schooling treatment that they could achieve very similar 2SLS estimates as were printed in Angrist and Krueger (1991), which turned out to be in turn very close to the OLS estimates from regressing earnings on years of schooling. They then showed formally that this is not a surprise: 2SLS with many random instruments tends to also be biased towards OLS in finite samples.

To gain some intuition on the many instruments problem, consider an extreme case in which every observation in the data gets its own instrument Z_{ii}. In this scenario with $L = N$ instruments, the first stage regression of X_i on $Z_{i1}, \ldots, Z_{iN}$ will generally fit perfectly; in a sense this regression is "saturated" in individual indicators. Consequently, the first stage fitted values $\boldsymbol{Z}_i'\widehat{\Pi}$ will coincide with the endogenous treatment X_i such that 2SLS coincides with OLS numerically. Such "overfitting" of X_i will generally occur whenever we have a large number of instruments which by chance track the endogenous variation in X_i, and this will generally lead to bias (in finite samples) towards the OLS estimand.

The many-instruments problem will again manifest with a low (i.e. below 10) first-stage F-statistic, giving more justification to always check this statistic when running an IV regression. In most cases many instruments is less of a concern than weak instruments, since we typically do not have a large number of independent Z_{ij} arising naturally and we now (since Bound et al. (1995), anyway) know not to interact an instrument with controls to generate many instruments without cause. But there are some empirical settings where many instruments are common, and the solution in such cases is again not always clear. As with many topics in this class, you should expect a more sophisticated and thorough treatment of the problem and possible solutions in your future econometrics courses.

Overidentification Tests

Our final IV topic concerns the scenario where you are lucky enough to have more instruments than endogenous variables, $L = dim(\boldsymbol{Z}_i) > dim(\boldsymbol{X}_i) = K$, and are interested in testing the identifying assumption of $E[\boldsymbol{Z}_i\varepsilon_i] = 0$. Recall that we said in this *overidentified* setting any full-rank combination of the instruments $\boldsymbol{M}\boldsymbol{Z}_i$ can be used for identification and estimation, though we typically restricted ourselves to $\boldsymbol{M} = \Pi$ matrix coming from the first-stage regression. Intuitively, the fact that we have multiple possible ways to estimate $\boldsymbol{\beta}$ in this case gives us grounds to test the validity of our instruments: if we use different $\boldsymbol{M}$ combinations and get (statistically) different answers, then we know that $E[\boldsymbol{Z}_i\varepsilon_i] = 0$ does not hold: at least some (and perhaps all) of the instruments in $\boldsymbol{Z}_i$ are invalid.

Before formalizing this logic, it is important to be clear on what such *overidentification tests* can and can't be useful for. At first glance, they appear to answer what I've previously said is a fundamentally untestable assumption of instrument validity. Alas, they do not quite tell us what instruments are and are not valid: when we reject the joint null of $E[\boldsymbol{Z}_i\varepsilon_i] = 0$ we don't know which elements of $\boldsymbol{Z}_i$ are correlated with ε_i,and in fact all could be. Furthermore, as it turns out the overidentification tests we are about to derive are not especially powerful when the instruments do not induce sufficiently different variation in $\boldsymbol{X}_i$ in the first stage—which is more common than you might think. Still, overidentification tests can be useful diagnostics or summaries of the underlying IV variation and, perhaps more importantly, they are often deployed in papers you might read.

The most common overidentification test statistic is due to an econometrician named Lars Hansen, and can be written in matrix form as

$$\widehat{J} = (\boldsymbol{Y} - \underline{\boldsymbol{X}}\widehat{\underline{\boldsymbol{\beta}}})'\underline{\boldsymbol{Z}}\widehat{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{Z}}'(\boldsymbol{Y} - \underline{\boldsymbol{X}}\widehat{\underline{\boldsymbol{\beta}}}), \tag{24}$$

where $\widehat{\underline{\boldsymbol{\beta}}}$ is an IV estimate, $\widehat{\boldsymbol{\Sigma}}$ consistently estimates the asymptotic variance of $\sqrt{N}\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\varepsilon_i\right)$, and we collect observations of the outcome, endogenous variables, instruments, and controls in $\boldsymbol{Y}$, $\underline{\boldsymbol{X}}$, and $\underline{\boldsymbol{Z}}$ as you would expect. Hansen showed that this test statistic has, under the null of $E[\boldsymbol{Z}_i\varepsilon_i] = 0$, an asymptotic $\chi^2(L-J)$ distribution and this can be used to test the joint null of instrument validity. To build intuition for

this, note that under this null $\widehat{\boldsymbol{\Sigma}}^{-1/2}\frac{1}{\sqrt{N}}\underline{\boldsymbol{Z}}'(\boldsymbol{Y}-\underline{\boldsymbol{X}}\underline{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\Sigma}}^{-1/2}\sqrt{N}\left(\frac{1}{N}\sum_i \underline{\boldsymbol{Z}}_i\varepsilon_i\right)$ is asymptotically normal with mean zero and variance $\boldsymbol{I}$. Thus, as we saw before with the Wald statistic, the inner product of this vector given in (25) is asymptotically chi-squared with degrees of freedom of $dim(\boldsymbol{Z}_i)+dim(\boldsymbol{W}_i) = L+K$. It turns out that using an estimate of $\underline{\boldsymbol{\beta}}$ in place of the true coefficient in (25) "uses up" $J+K$ degrees of freedom in this calculation, such that the feasible version has degrees of freedom of $L+K-(J-K) = L-J$. Intuitively, this means $\widehat{T} \neq 0$ (such that the test has power) only when $L > J$ (i.e., when we are overidentified). In Stata, the *ivreg2* command automatically computes the Hansen "J-statistic" whenever this is the case.