see why the more robust 2SLS and Abadie estimators should not be preferred.[37]

4.6.4 *The Bias of 2SLS*★

It is a fortunate fact that the OLS estimator is not only consistent, it is also unbiased (as we briefly noted at the end of section 3.1.3). This means that in a sample of any size, the estimated OLS coefficient vector has a distribution that is centered on the population coefficient vector.[38] The 2SLS estimator, in contrast, is consistent, but biased. This means that the 2SLS estimator only promises to be close to the causal effect of interest in large samples. In small samples, 2SLS estimates can differ systematically from the target parameter.

For many years, applied researchers lived with the knowledge that 2SLS is biased without losing too much sleep. Neither of us heard much about the bias of 2SLS in our graduate econometrics classes. A series of papers in the early 1990s changed this, however. These papers show that 2SLS estimates can be highly misleading in cases relevant for empirical practice.[39]

The 2SLS estimator is most biased when the instruments are "weak," meaning the correlation with endogenous regressors is low, and when there are many overidentifying restrictions. When the instruments are both many and weak, the 2SLS estimator is biased toward the probability limit of the corresponding OLS estimate. In the worst-case scenario, when the instruments are so weak that there is no first stage in the population, the 2SLS sampling distribution is centered on the

[37] Angrist (2001) makes the same point using twins instruments and reports a similar pattern in a comparison of 2SLS, Abadie, and nonlinear structural estimates of models for hours worked.

[38] A more precise statement is that OLS is unbiased when either (1) the CEF is linear or (2) the regressors are nonstochastic, that is, fixed in repeated samples. In practice, these qualifications do not seem to matter much. As a rule, the sampling distribution of $\hat{\beta} = [\sum_i \mathrm{X}_i\mathrm{X}_i']^{-1} \sum_i \mathrm{X}_i\mathrm{Y}_i$, tends to be centered on the population analog, $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$, in samples of any size, whether or not the CEF is linear or the regressors are stochastic.

[39] Key references are Nelson and Startz (1990a,b), Buse (1992), Bekker (1994), and especially Bound, Jaeger, and Baker (1995).

probability limit of OLS. The theory behind this result is a little technical, but the basic idea is easy to see. The source of the bias in 2SLS estimates is the randomness in estimates of the first-stage fitted values. In practice, the first-stage estimates reflect some of the randomness in the endogenous variable, since the first-stage coefficients come from a regression of the endogenous variable on the instruments. If the population first stage is zero, then all randomness in the first stage is due to the endogenous variable. This randomness generates finite-sample correlation between first-stage fitted values and second-stage errors, since the endogenous variable is correlated with the second-stage errors (or else you wouldn't be instrumenting in the first place).

A more formal derivation of 2SLS bias goes like this. To streamline the discussion we use matrices and vectors and a simple constant-effects model (it's difficult to discuss bias in a heterogeneous effects world, since the target parameter may change as the number of instruments changes). Suppose you are interested in estimating the effect of a single endogenous regressor, stored in a vector x, on a dependent variable, stored in the vector y, with no other covariates. The causal model of interest can then be written

$$y = \beta x + \eta. \tag{4.6.17}$$

The $N \times Q$ matrix of instrumental variables is Z, with the associated first-stage equation

$$x = Z\pi + \xi. \tag{4.6.18}$$

OLS estimates of (4.6.17) are biased because η_i is correlated with ξ_i. The instruments Z_i are uncorrelated with ξ_i by construction and uncorrelated with η_i by assumption.

The 2SLS estimator is

$$\hat{\beta}_{2SLS} = (x'P_Z x)^{-1} x'P_Z y = \beta + (x'P_Z x)^{-1} x'P_Z \eta,$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix that produces fitted values from a regression of x on Z. Substituting for x in

$x'P_Z\eta$, we get

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= (x'P_Zx)^{-1}(\pi'Z' + \xi')P_Z\eta \\ &= (x'P_Zx)^{-1}\pi'Z'\eta + (x'P_Zx)^{-1}\xi'P_Z\eta. \quad (4.6.19)\end{aligned}$$

The bias in 2SLS comes from the nonzero expectation of terms on the right-hand side.

The expectation of (4.6.19) is hard to evaluate because the expectation operator does not pass through the inverse $(x'P_Zx)^{-1}$, a nonlinear function. It's possible to show, however, that the expectation of the ratios on the right-hand side of (4.6.19) can be closely approximated by the ratio of expectations. In other words,

$$E[\hat{\beta}_{2SLS} - \beta] \approx (E[x'P_Zx])^{-1}E[\pi'Z'\eta] + (E[x'P_Zx])^{-1}E[\xi'P_Z\eta].$$

This approximation is much better than the usual asymptotic approximation invoked in large-sample theory, so we think of it as giving us a good measure of the finite-sample behavior of the 2SLS estimator.[40] Furthermore, because $E[\pi'Z'\xi] = 0$ and $E[\pi'Z'\eta] = 0$, we have

$$E[\hat{\beta}_{2SLS} - \beta] \approx [E(\pi'Z'Z\pi) + E(\xi'P_Z\xi)]^{-1}E(\xi'P_Z\eta). \quad (4.6.20)$$

The approximate bias of 2SLS therefore comes from the fact that $E(\xi'P_Z\eta)$ is not zero unless η_i and ξ_i are uncorrelated. But correlation between η_i and ξ_i is what led us to use IV in the first place.

Further manipulation of (4.6.20) generates an expression that is especially useful:

$$E[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma_\xi^2}\left[\frac{E(\pi'Z'Z\pi)/Q}{\sigma_\xi^2} + 1\right]^{-1}$$

[40]See Bekker (1994) and Angrist and Krueger (1995). This is also called a group-asymptotic approximation because it can be derived from an an asymptotic sequence that lets the number of instruments go to infinity at the same time as the number of observations goes to infinity, keeping the number of observations per instrument (group) constant.

(see the appendix for a derivation). The term $(1/\sigma_\xi^2)E(\pi'Z'Z\pi)/Q$ is the F-statistic for the joint significance of all regressors in the first stage regression.[41] Call this statistic F, so that we can write

$$E[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma_\xi^2} \frac{1}{F+1}. \tag{4.6.21}$$

From this we see that as the first stage F-statistic gets small, the bias of 2SLS approaches $\sigma_{\eta\xi}/\sigma_\xi^2$. The bias of the OLS estimator is $\sigma_{\eta\xi}/\sigma_x^2$, which also equals $\sigma_{\eta\xi}/\sigma_\xi^2$ if $\pi = 0$. Thus, we have shown that 2SLS is centered on the same point as OLS when the first stage is zero. More generally, we can say 2SLS estimates are "biased toward OLS estimates" when there isn't much of a first stage. On the other hand, the bias of 2SLS vanishes when F gets large, as should happen in large samples when $\pi \neq 0$.

When the instruments are weak, the F-statistic varies inversely with the number of instruments. To see why, consider adding useless instruments to your 2SLS model, that is, instruments with no effect on the first-stage R^2. The model sum of squares, $E(\pi'Z'Z\pi)$, and the residual variance, σ_ξ^2, will both stay the same while Q goes up. The F-statistic becomes smaller as a result. From this we learn that the addition of many weak instruments increases bias.

Intuitively, the bias in 2SLS is a consequence of the fact that the first-stage is estimated. If the first stage coefficients were known, we could use $\hat{x}_{pop} = Z\pi$ for the first-stage fitted values. These fitted values are uncorrelated with the second-stage error. In practice, however, we use $\hat{x} = P_Z x = Z\pi + P_Z\xi$, which differs from $\hat{x}_{pop}$ by the term $P_Z\xi$. The bias in 2SLS arises from the fact that $P_Z\xi$ is correlated with η, so some of

[41] Sort of; the actual F-statistic is $(1/\hat{\sigma}_\xi^2)\hat{\pi}'Z'Z\hat{\pi}/Q$, where hats denote estimates. $(1/\sigma_\xi^2)E(\pi'Z'Z\pi)/Q$ is therefore sometimes called the population F-statistic since it's the F-statistic we'd get in an infinitely large sample. In practice, the distinction between population and sample F matters little in this context. Some econometricians prefer to multiply the first-stage F by the number of instruments when summarizing instrument strength. This product is called the "concentration parameter."

the correlation between errors in the first and second stages seeps into our 2SLS estimates through the sampling variability in $\hat{\pi}$. Asymptotically, this correlation disappears, but real life does not play out in asymptopia.

Formula (4.6.21) shows that, other things equal, the bias in 2SLS is an increasing function of the number of instruments, so bias is least in the just-identified case when the number of instruments is as low as it can get. In fact, just-identified 2SLS (say, the simple Wald estimator) is approximately *unbiased*. This is hard to show formally because just-identified 2SLS has no moments (i.e., the sampling distribution has fat tails). Nevertheless, even with weak instruments, just-identified 2SLS is approximately centered where it should be. We therefore say that just-identified 2SLS is median-unbiased. This is not to say that you can happily use weak instruments in just-identified models. With a weak instrument, just-identified estimates tend to be too imprecise to be useful.

The limited information maximum likelihood (LIML) estimator is approximately median-unbiased for overidentified constant effects models, and therefore provides an attractive alternative to just-identified estimation using one instrument at a time (see, e.g., Davidson and MacKinnon, 1993, and Mariano, 2001). LIML has the advantage of having the same asymptotic distribution as 2SLS (under constant effects) while providing a finite-sample bias reduction. A number of other estimators also reduce the bias in overidentified 2SLS models. But an extensive Monte Carlo study by Flores-Lagunes (2007) suggests that LIML does at least as well as the alternatives in a wide range of circumstances (in terms of bias, mean absolute error, and the empirical rejection rates for t-tests). Another advantage of LIML is that many statistical packages compute it, while other estimators typically require some programming.[42]

[42]LIML is available in SAS and in STATA 10. With weak instruments, LIML standard errors are not quite right, but Bekker (1994) gives a simple fix for this. Why is LIML unbiased? Expression (4.6.21) shows that the approximate bias of 2SLS is proportional to the bias of OLS. From this we conclude that there is a linear combination of OLS and 2SLS that is approximately