Chapter 4

Instrumental Variables in Action: Sometimes You Get What You Need

> Anything that happens, happens.
> Anything that, in happening, causes something else to happen,
> causes something else to happen.
> Anything that, in happening,
> causes itself to happen again, happens again.
> It doesn't necessarily do it in chronological order, though.
>
> Douglas Adams, *Mostly Harmless* (1995)

Two things distinguish the discipline of Econometrics from our older sister field of Statistics. One is a lack of shyness about causality. Causal inference has always been the name of the game in applied econometrics. Statistician Paul Holland (1986) cautions that there can be "no causation without manipulation," a maxim that would seem to rule out causal inference from non-experimental data. Less thoughtful observers fall back on the truism that "correlation is not causality." Like most people who work with data for a living, we believe that correlation can sometimes provide pretty good evidence of a causal relation, even when the variable of interest has not been manipulated by a researcher or experimenter. [1]

The second thing that distinguishes us from most statisticians—and indeed most other social scientists—is an arsenal of statistical tools that grew out of early econometric research on the problem of how to estimate the parameters in a system of linear simultaneous equations. The most powerful weapon in this arsenal is the method of Instrumental Variables (IV), the subject of this chapter. As it turns out, IV does more than allow us to consistently estimate the parameters in a system of simultaneous equations, though it allows us

[1] Recent years have seen an increased willingness by statisticians to discuss statistical models for observational data in an explicitly causal framework; see, for example, Freedman's (2005) review.

to do that as well.

Studying agricultural markets in the 1920s, the father and son research team of Phillip and Sewall Wright were interested in a challenging problem of causal inference: how to estimate the slope of supply and demand curves when observed data on prices and quantities are determined by the intersection of these two curves. In other words, equilibrium prices and quantities—the only ones we get to observe—solve these two stochastic equations at the same time. Upon which curve, therefore, does the observed scatterplot of prices and quantities lie? The fact that population regression coefficients do not capture the slope of any one equation in a set of simultaneous equations had been understood by Phillip Wright for some time. The IV method, first laid out in Wright (1928), solves the statistical simultaneous equations problem by using variables that appear in one equation to shift this equation and trace out the other. The variables that do the shifting came to be known as *instrumental variables* (Reiersol, 1941).

In a separate line of inquiry, IV methods were pioneered to solve the problem of bias from measurement error in regression models[2]. One of the most important results in the statistical theory of linear models is that a regression coefficient is biased towards zero when the regressor of interest is measured with random errors (to see why, imagine the regressor contains only random error; then it will be uncorrelated with the dependent variable, and hence the regression of $\texttt{Y}_i$ on this variable will be zero). Instrumental variables methods can be used to eliminate this sort of bias.

Simultaneous equations models (SEMs) have been enormously important in the history of econometric thought. At the same time, few of today's most influential applied papers rely on an orthodox SEM framework, though the technical language used to discuss IV still comes from this framework. Today, we are more likely to find IV used to address measurement error problems than to estimate the parameters of an SEM. Undoubtedly, however, the most important contemporary use of IV is to solve the problem of omitted variables bias. IV solves the problem of missing or unknown control variables, much as a randomized trial obviates the need for extensive controls in a regression.[3]

4.1 IV and causality

We like to tell the IV story in two iterations, first in a restricted model with constant effects, then in a framework with unrestricted heterogeneous potential outcomes, in which case causal effects must also be heterogeneous. The introduction of heterogeneous effects enriches the interpretation of IV estimands, without changing the mechanics of the core statistical methods we are most likely to use in practice (typically, two-stage least squares). An initial focus on constant effects allows us to explain the mechanics of IV with a

[2] Key historical references here are Wald (1940) and Durbin (1954), both discussed below.

[3] See Angrist and Krueger (2001) for a brief exposition of the history and uses of IV; Stock and Trebbi (2003) for a detailed account of the birth of IV; and Morgan (1990) for an extended history of econometric ideas, including the simultaneous equations model.

minimum of fuss.

To motivate the constant-effects setup as a framework for the causal link between schooling and wages, suppose, as before, that potential outcomes can be written

$$\mathrm{Y}_{si} \equiv f_i(s),$$

and that

$$f_i(s) = \pi_0 + \pi_1 s + \eta_i, \tag{4.1.1}$$

as in the introduction to regression in Chapter 3. Also, as in the earlier discussion, imagine that there is a vector of control variables, A_i, called "ability", that gives a selection-on-observables story:

$$\eta_i = A_i'\gamma + v_i,$$

where γ is again a vector of population regression coefficients, so that v_i and A_i are uncorrelated by construction. For now, the variables A_i, are assumed to be the only reason why η_i and S_i are correlated, so that

$$E[\mathrm{S}_i v_i] = 0.$$

In other words if A_i were observed, we would be happy to include it in the regression of wages on schooling; thereby producing a long regression that can be written

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + A_i'\gamma + v_i. \tag{4.1.2}$$

Equation (4.1.2) is a version of the linear causal model, (3.2.9). The error term in this equation is the random part of potential outcomes, v_i, left over after controlling for A_i. This error term is uncorrelated with schooling by assumption. If this assumption turns out to be correct, the population regression of Y_i on S_i and A_i produces the coefficients in (4.1.2).

The problem we initially want to tackle is how to estimate the long-regression coefficient, ρ, when A_i is unobserved. Instrumental variables methods can be used to accomplish this when the researcher has access to a variable (the instrument, which we'll call Z_i), that is correlated with the causal variable of interest, S_i, but uncorrelated with any other determinants of the dependent variable. Here, the phrase "uncorrelated with any other determinants of the dependent variables" is like saying $Cov(\eta_i, \mathrm{Z}_i) = 0$, or, equivalently, Z_i is uncorrelated with both A_i and v_i. This statement is called an *exclusion restriction* since Z_i can be said to be excluded from the causal model of interest. The exclusion restriction is a version of the conditional independence assumption of the previous chapter, except that now it is the instrument which is independent of potential outcomes, instead of schooling itself (the "conditional" in conditional independence enters into

the discussion when we consider IV models with covariates).

Given the exclusion restriction, it follows from equation (4.1.2) that

$$\rho = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)} = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)}. \tag{4.1.3}$$

The second equality in (4.1.3) is useful because it's usually easier to think in terms of regression coefficients than in terms of covariances. The coefficient of interest, ρ, is the ratio of the population regression of Y_i on Z_i (the reduced form) to the population regression of S_i on Z_i (the first stage). The IV *estimator* is the sample analog of expression (4.1.3). Note that the IV *estimand* is predicated on the notion that the first stage is not zero, but this is something you can check in the data. As a rule, if the first stage is only marginally significantly different from zero, the resulting IV estimates are unlikely to be informative, a point we return to later.

It's worth recapping the assumptions needed for the ratio of covariances in (4.1.3) to equal the casual effect, ρ. First, the instrument must have a clear effect on S_i. This is the first stage. Second, the only reason for the relationship between Y_i and Z_i is the first-stage. For the moment, we're calling this second assumption the exclusion restriction, though as we'll see in the discussion of models with heterogeneous effects, this assumption really has two parts: the first is the statement that the instrument is as good as randomly assigned (i.e., independent of potential outcomes, conditional on covariates), while the second is that the instrument has no effect on outcomes other than through the first-stage channel.

So where can you find an instrumental variable? Good instruments come from institutional knowledge and your ideas about the processes determining the variable of interest. For example, the economic model of education suggests that educational attainment is determined by comparing the costs and benefits of alternative choices. Thus, one possible source of instruments for schooling is differences in costs due, say, to loan policies or other subsidies that vary independently of ability or earnings potential. A second source of variation in schooling is institutional constraints. A set of institutional constraints relevant for schooling are compulsory schooling laws. Angrist and Krueger (1991) exploit the variation induced by compulsory schooling in a paper that typifies the use of "natural experiments" to try to eliminate omitted variables bias

The starting point for the Angrist and Krueger (1991) quarter-of-birth strategy is the observation that most states required students to enter school in the calendar year in which they turn 6. School start age is therefore a function of date of birth. Specifically, those born late in the year are young for their grade. In states with a December 31st birthday cutoff, children born in the fourth quarter enter school shortly before they turn 6, while those born in the first quarter enter school at around age $6\frac{1}{2}$. Furthermore, because compulsory schooling laws typically require students to remain in school only until their 16th birthday, these groups of students will be in different grades or through a given grade to different degree, when they reach the legal dropout age. In essence, the combination of school start age policies and compulsory schooling laws

creates a natural experiment in which children are compelled to attend school for different lengths of time depending on their birthdays.

Angrist and Krueger looked at the relationship between educational attainment and quarter of birth using US census data. Panel A of Figure 4.1.1 (adapted from Angrist and Krueger, 2001) displays the education-quarter-of-birth pattern for men in the 1980 Census who were born in the 1930s. The figure clearly shows that men born earlier in the calendar year tend to have lower average schooling levels. Panel A of Figure 4.1.1 is a graphical representation of the first-stage. The first-stage in a general IV framework is the regression of the causal variable of interest on covariates and the instrument(s). The plot summarizes this regression because average schooling by year and quarter of birth is what you get for fitted values from a regression of schooling on a full set of year-of-birth and quarter-of-birth dummies.

Panel B of Figure 4.1.1 displays average earnings by quarter of birth for the same sample used to construct panel A. This panel illustrates what econometricians call the "reduced form" relationship between the instruments and the dependent variable. The reduced form is the regression of the dependent variable on any covariates in the model and the instrument(s). Panel B shows that older cohorts tend to have higher earnings, because earnings rise with work experience. The figure also shows that men born in early quarters almost always earned less, on average, than those born later in the year, even after adjusting for year of birth, which plays the role of an exogenous covariate in the Angrist and Krueger (1991) setup. Importantly, this reduced-form relation parallels the quarter-of-birth pattern in schooling, suggesting the two patterns are closely related. Because an individual's date of birth is probably unrelated to his or her innate ability, motivation, or family connections, it seems credible to assert that the only reason for the up-and-down quarter-of-birth pattern in earnings is indeed the up-and-down quarter-of-birth pattern in schooling. This is the critical assumption that drives the quarter-of-birth IV story.[4]

A mathematical representation of the story told by Figure 4.1.1 comes from the first-stage and reduced-form regression equations, spelled out below:

$$\mathrm{S}_i = \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i + \xi_{1i} \tag{4.1.4a}$$

$$\mathrm{Y}_i = \mathrm{X}_i'\pi_{20} + \pi_{21}\mathrm{Z}_i + \xi_{2i} \tag{4.1.4b}$$

The parameter π_{11} in equation (4.1.4a) captures the first-stage effect of Z_i on S_i, adjusting for covariates,

[4] Other explanations are possible, the most likely being some sort of family background effect associated with season of birth (see, e.g., Bound, Jaeger, and Baker, 1995). Weighing against the possibility of omitted family background effects is the fact that the quarter of birth pattern in average schooling is much more pronounced at the schooling levels most affected by compulsory attendance laws. Another possible concern is a pure age-at-entry effect which operates through channels other than highest grade completed (e.g., achievement). The causal effect of age-at-entry on learning is difficult, if not impossible, to separate from pure age effects, as noted in Chapter 1). A recent study by Elder and Lubotsky (2008) argues that the evolution of putative age-at-entry effects over time is more consistent with effects due to age differences *per se* than to a within-school learning advantage for older students.

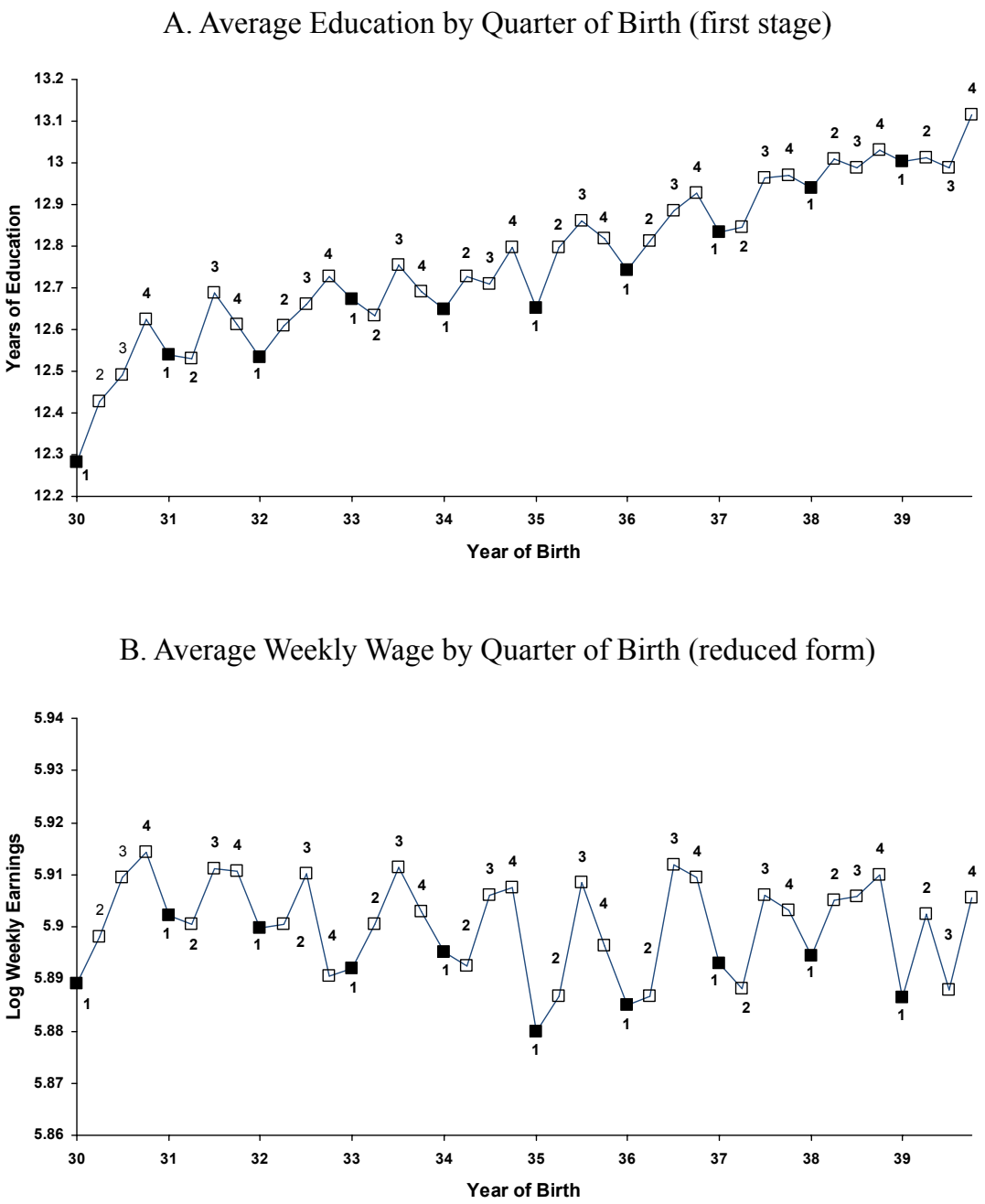

Figure 4.1.1: Graphical depiction of first stage and reduced form for IV estimates of the economic return to schooling using quarter of birth (from Angrist and Krueger 1991).

X_i. The parameter π_{21} in equation (4.1.4b) captures the reduced-form effect of Z_i on Y_i, adjusting for these same covariates. In the language of the SEM, the dependent variables in these two equations are said to be the *endogenous variables* (where they are determined jointly within the system) while the variables on the right-hand side are said to be the *exogenous variables* (determined outside the system). The instruments, Z_i, are a subset of the exogenous variables. The exogenous variables that are not instruments are said to be *exogenous covariates*. Although we're not estimating a traditional supply and demand system in this case, these SEM variable labels are still widely used in empirical practice.

The covariate-adjusted IV estimator is the sample analog of the ratio $\frac{\pi_{21}}{\pi_{11}}$. To see this, note that the denominators of the reduced-form and first-stage effects are the same. Hence, their ratio is

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{Cov(\mathrm{Y}_i, \tilde{z}_i)}{Cov(\mathrm{S}_i, \tilde{z}_i)}, \tag{4.1.5}$$

where $\tilde{z}_i$ is the residual from a regression of Z_i on the exogenous covariates, X_i. The right-hand side of (4.1.5) therefore swaps $\tilde{z}_i$ for Z_i in the general IV formula, (4.1.3). Econometricians call the sample analog of the left-hand side of equation (4.1.5) an Indirect Least Squares (ILS) estimator of ρ in the causal model with covariates,

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho \mathrm{S}_i + \eta_i, \tag{4.1.6}$$

where η_i is the compound error term, $A_i'\gamma + v_i$[5]. It's easy to use equation (4.1.6) to confirm directly that $Cov(\mathrm{Y}_i, \tilde{z}_i) = \rho Cov(\mathrm{S}_i, \tilde{z}_i)$ since $\tilde{z}_i$ is uncorrelated with X_i by construction and with η_i by assumption. In Angrist and Krueger (1991), the instrument, Z_i, is quarter of birth (or dummies indicating quarters of birth) and the covariates are dummies for year of birth, state of birth, and race.

4.1.1 Two-Stage Least Squares

The reduced-form equation, (4.1.4b), can be derived by substituting the first stage equation, (4.1.4a), into the causal relation of interest, (4.1.6), which is also called a "structural equation" in simultaneous equations language. We then have:

$$\begin{aligned} \mathrm{Y}_i &= \alpha' \mathrm{X}_i + \rho[\mathrm{X}_i' \pi_{10} + \pi_{11} \mathrm{Z}_i + \xi_{1i}] + \eta_i \\ &= \mathrm{X}_i'[\alpha + \rho\pi_{10}] + \rho\pi_{11}\mathrm{Z}_i + [\rho\xi_{1i} + \eta_i] \\ &= \mathrm{X}_i'\pi_{20} + \pi_{21}\mathrm{Z}_i + \xi_{2i}, \end{aligned} \tag{4.1.7}$$

[5] For a direct proof that (4.1.5) equals ρ in (4.1.6), use (4.1.6) to substitute for Y_i in $\frac{Cov(\mathrm{Y}_i, \tilde{z}_i)}{Cov(\mathrm{S}_i, \tilde{z}_i)}$.

where $\pi_{20} \equiv \alpha + \rho\pi_{10}$, $\pi_{21} \equiv \rho\pi_{11}$, and $\xi_{2i} \equiv \rho\xi_{1i} + \eta_i$ in equation (4.1.4b). Equation (4.1.7) again shows why $\rho = \frac{\pi_{21}}{\pi_{11}}$. Note also that a slight re-arrangement of (4.1.7) gives

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i] + \xi_{2i}, \tag{4.1.8}$$

where $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$ is the population fitted value from the first-stage regression of S_i on X_i and Z_i. Because Z_i and X_i are uncorrelated with the reduced-form error, ξ_{2i}, the coefficient on $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$ in the population regression of Y_i on X_i and $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$ equals ρ.

In practice, of course, we almost always work with data from samples. Given a random sample, the first-stage fitted values in the population are consistently estimated by

$$\hat{s}_i = \mathrm{X}_i'\hat{\pi}_{10} + \hat{\pi}_{11}\mathrm{Z}_i,$$

where $\hat{\pi}_{10}$ and $\hat{\pi}_{11}$ are OLS estimates from equation (4.1.4a). The coefficient on $\hat{s}_i$ in the regression of Y_i on X_i and $\hat{s}_i$ is called the Two-Stage Least Squares (2SLS) estimator of ρ. In other words, 2SLS estimates can be constructed by OLS estimation of the "second-stage equation,"

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\hat{s}_i + [\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)], \tag{4.1.9}$$

This is called 2SLS because it can be done in two steps, the first estimating $\hat{s}_i$ using equation (4.1.4a), and the second estimating equation (4.1.9). The resulting estimator is consistent for ρ because (a) first-stage estimates are consistent; and, (b) the covariates, X_i, and instruments, Z_i, are uncorrelated with both η_i and $(\mathrm{S}_i - \hat{s}_i)$.

The 2SLS name notwithstanding, we don't usually construct 2SLS estimates in two-steps. For one thing, the resulting standard errors are wrong, as we discuss later. Typically, we let specialized software routines (such as are available in SAS or Stata) do the calculation for us. This gets the standard errors right and helps to avoid other mistakes (see Section 4.6.1, below). Still, the fact that the 2SLS estimator can be computed by a sequence of OLS regressions is one way to remember why it works. Intuitively, conditional on covariates, 2SLS retains only the variation in S_i that is generated by quasi-experimental variation, i.e., generated by the instrument, Z_i.

2SLS is a many-splendored thing. For one, it is an instrumental variables estimator: the 2SLS estimate of ρ in (4.1.9) is the sample analog of $\frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{Cov(\mathrm{S}_i, \hat{s}_i^*)}$, where $\hat{s}_i^*$ is the residual from a regression of $\hat{s}_i$ on X_i. This follows from the multivariate regression anatomy formula and the fact that $Cov(\mathrm{S}_i, \hat{s}_i^*) = V(\hat{s}_i^*)$. It is also easy to show that, in a model with a single endogenous variable and a single instrument, the 2SLS estimator is the same as the corresponding ILS estimator.[6]

[6]Note that $\hat{s}_i^* = \tilde{z}_i\hat{\pi}_{11}$, where $\tilde{z}_i$ is the residual from a regression of Z_i on X_i, so that the 2SLS estimator is therefore the

The link between 2SLS and IV warrants a bit more elaboration in the multi-instrument case. Assuming each instrument captures the same causal effect (a strong assumption that is relaxed below), we might want to combine these alternative IV estimates into a single more precise estimate. In models with multiple instruments, 2SLS provides just such a linear combination by combining multiple instruments into a single instrument. Suppose, for example, we have three instrumental variables, z_{1i}, z_{2i}, and z_{3i}. In the Angrist and Krueger (1991) application, these are dummies for first, second, and third-quarter births. The first-stage equation then becomes

$$\mathrm{s}_i = \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{z}_{1i} + \pi_{12}\mathrm{z}_{2i} + \pi_{13}\mathrm{z}_{3i} + \xi_{1i}, \tag{4.1.10a}$$

while the 2SLS second stage is the same as (4.1.9), except that the fitted values are from (4.1.10a) instead of (4.1.4a). The IV interpretation of this 2SLS estimator is the same as before: the instrument is the residual from a regression of first-stage fitted values on covariates. The exclusion restriction in this case is the claim that all of the quarter of birth dummies in (4.1.10a) are uncorrelated with η_i in equation equation (4.1.6).

The results of 2SLS estimation of a schooling equation using three quarter-of-birth dummies, as well as other interactions, are shown in Table 4.1.1, which reports OLS and 2SLS estimates of models similar to those estimated by Angrist and Krueger (1991). Each column in the table contains OLS and 2SLS estimates of ρ from an equation like (4.1.6), estimated with different combinations of instruments and control variables. The OLS estimate in column 1 is from a regression of log wages with no control variables, while the OLS estimates in column 2 are from a model adding dummies for year of birth and state of birth as control variables. In both cases, the estimated return to schooling is around .075.

sample analog of $\frac{\left[\frac{Cov(\mathrm{y}_i,\tilde{z}_i)}{V(\tilde{z}_i)}\right]}{\hat{\pi}_{11}}$. But the sample analog of the numerator, $\frac{Cov(\mathrm{y}_i,\tilde{z}_i)}{V(\tilde{z}_i)}$, is the OLS estimate of π_{21} in the reduced form, (4.1.4b), while $\hat{\pi}_{11}$ is the OLS estimate of the first-stage effect, π_{11}, in (4.1.4a). Hence, 2SLS with a single instrument is ILS, i.e., the ratio of the reduced form-effect of the instrument to the corresponding first-stage effect where both the first-stage and reduced-form include covariates.

Table 4.1.1: 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Years of education	0.075 (0.0004)	0.072 (0.0004)	0.103 (0.024)	0.112 (0.021)	0.106 (0.026)	0.108 (0.019)	0.089 (0.016)	0.061 (0.031)
Covariates:								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year of birth dummies		✓			✓	✓	✓	✓
50 state of birth dummies		✓			✓	✓	✓	✓
Instruments:			dummy for QOB=1	dummy for QOB=1 or QOB=2	dummy for QOB=1	full set of QOB dummies	full set of QOB dummies int. with year of birth dummies	full set of QOB dummies int. with year of birth dummies

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses.

The first pair of IV estimates, reported in columns 3 and 4, are from models without controls. The instrument used to construct the estimates in column 1 is a single dummy for first quarter births, while the instruments used to construct the estimates in column 2 are a pair of dummies indicating first and second quarter births. The standard error estimates range from .10 – .11. The results from models including year of birth and state of birth dummies as control variables are similar, not surprisingly, since quarter of birth is not closely related to either of these controls. Overall, the 2SLS estimates are mostly a bit larger than the corresponding OLS estimates. This suggests that the observed associated between schooling and earnings is not driven by omitted variables like ability and family background.

Column 7 in Table 4.1.1 shows the results of adding interaction terms to the instrument list. In particular, each specification adds interaction with 9 dummies for year of birth (the sample includes cohorts born 1930-39), for a total of 30 excluded instruments. The first stage equation becomes

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_{1i} + \pi_{12}\mathrm{Z}_{2i} + \pi_{13}\mathrm{Z}_{3i} \\ &\quad + \sum_j (\mathrm{B}_{ij}\mathrm{Z}_{1i})\kappa_{1j} + \sum_j (\mathrm{B}_{ij}\mathrm{Z}_{2i})\kappa_{2j} + \sum_j (\mathrm{B}_{ij}\mathrm{Z}_{3i})\kappa_{3j} + \xi_{1i} \end{aligned} \tag{4.1.10b}$$

where B_{ij} is a dummy equal to one if individual i was born in year j for j equal to 1931 – 39. The coefficients $\kappa_{1j}, \kappa_{2j}, \kappa_{3j}$ are the corresponding year-of-birth interactions. These interaction terms capture differences in the relation between quarter-of-birth and schooling across cohorts. The rationale for adding these interaction terms is an increase in precision that comes from increasing the first-stage R^2, which goes up because the quarter of birth pattern in schooling differs across cohorts. In this example, the addition of interaction terms to the instrument list leads to a modest gain in precision; the standard error declines from .0194 to .0161.[7]

The last 2SLS model reported in Table 4.1.1 includes controls for linear and quadratic terms in age-in-quarters in the list of covariates, X_i. In other words, someone who was born in the first quarter of 1930 is recorded as being 50 years old on census day (April 1), 1980, while someone born in the fourth quarter is recorded as being 49.25 years old. This finely coded age variable, entered into the model with a linear and quadratic term, provides a partial control for the fact that small differences age may be an omitted variable that confounds the quarter-of-birth identification strategy. As long as the effects of age are similarly smooth, the quadratic age-in-quarters model will pick them up.

This variation in the 2SLS set-up illustrates the inter-play between identification and estimation. For the 2SLS procedure to work, there must be some variation in the first-stage fitted values conditional on whatever control variables (covariates) are included in the model. If the first-stage fitted values are a linear combination of the included covariates, then the 2SLS estimate simply does not exist. In equation (4.1.9) this

[7] This gain may not be without cost, as the use of many additional instruments opens up the possibility of increased bias, an issue discussed in Chapter 8, below.

is manifest by perfect multicollinearity. 2SLS estimates with quadratic age exist. But the variability "left over" in the first-stage fitted values is reduced when the covariates include variables like age in quarters, that are closely related to the instruments (quarter of birth dummies). Because this variability is the primary determinant of 2SLS standard errors, the estimate in column 8 is markedly less precise than that in column 7, though it is still close to the corresponding OLS estimate.

Recap of IV and 2SLS Lingo

As we've seen, the *endogenous variables* are the dependent variable and the independent variable(s) to be instrumented; in a simultaneous equations model, endogenous variables are determined by solving a system of stochastic linear equations. To *treat an independent variable as endogenous* is to instrument it, i.e., to replace it with fitted values in the second stage of a 2SLS procedure. The independent endogenous variable in the Angrist and Krueger (1991) study is schooling. The *exogenous variables* include the *exogenous covariates* that are not instrumented and the instruments themselves. In a simultaneous equations model, exogenous variables are determined outside the system. The exogenous covariates in the Angrist and Krueger (1991) study are dummies for year of birth and state of birth. We think of exogenous covariates as controls. 2SLS aficionados live in a world of mutually exclusive labels: in any empirical study involving instrumental variables, the random variables to be studied are either dependent variables, independent endogenous variables, instrumental variables, or exogenous covariates. Sometimes we shorten this to: dependent and endogenous variables, instruments and covariates (fudging the fact that the dependent variable is also endogenous in a traditional SEM).

4.1.2 The Wald Estimator

The simplest IV estimator uses a single binary (0-1) instrument to estimate a model with one endogenous regressor and no covariates. Without covariates, the causal regression model is

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + \eta_i, \tag{4.1.11}$$

where η_i and S_i may be correlated. Given the further simplification that Z_i is a dummy variable that equals 1 with probability p, we can easily show that

$$Cov(\mathrm{Y}_i, \mathrm{Z}_i) = \{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]\}p(1-p),$$

with an analogous formula for $Cov(\mathrm{S}_i,\mathrm{Z}_i)$. It therefore follows that

$$\rho = \frac{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]}{E[\mathrm{S}_i|\mathrm{Z}_i = 1] - E[\mathrm{S}_i|\mathrm{Z}_i = 0]}. \tag{4.1.12}$$

A direct route to this result uses (4.1.11) and the fact that $E[\eta_i|\mathrm{Z}_i] = 0$, so we have

$$E[\mathrm{Y}_i|\mathrm{Z}_i] = \alpha + \rho E[\mathrm{S}_i|\mathrm{Z}_i]. \tag{4.1.13}$$

Solving this equation for ρ produces (4.1.12).

Equation (4.1.12) is the population analog of the landmark Wald (1940) estimator for a bivariate regression with mismeasured regressors.[8] The *Wald estimator* is the sample analog of this expression. In our context, the Wald formula provides an appealingly transparent implementation of the IV strategy for the elimination of omitted variables bias. The principal claim that motivates IV estimation of causal effects is that the *only* reason for any relation between the dependent variable and the instrument is the effect of the instrument on the causal variable of interest. In the context of a binary instrument, it therefore seems natural to divide—or rescale—the reduced-form difference in means by the corresponding first-stage difference in means.

The Angrist and Krueger (1991) study using quarter of birth to estimate the economic returns to schooling shows the Wald estimator in action. Table 4.1.2 displays the ingredients behind a Wald estimate constructed using the 1980 census. The difference in earnings between men born in the first and second halves of the year is -.01349 (s.e.=.00337), while the corresponding difference in schooling is -.1514. The ratio of these two differences is a Wald estimate of the economic value of schooling in per-year terms. This comes out to be .0891 (s.e.=.021). Not surprisingly, this estimate is not too different from the 2SLS estimates in Table 4.1.1. The reason we should expect the Wald and 2SLS estimates to be similar is that they are both constructed from the same information: differences in earnings by season of birth.

The Angrist (1990) study of the effects of Vietnam-era military service on the earnings of veterans also shows the Wald estimator in action. In the 1960s and early 1970s, young men were at risk of being drafted for military service. Concerns about the fairness of US conscription policy led to the institution of a draft lottery in 1970 that was used to determine priority for conscription. A promising instrumental variable for Vietnam veteran status is therefore draft-eligibility, since this was determined by a lottery over birthdays. Specifically, in each year from 1970 to 1972, random sequence numbers (RSNs) were randomly assigned to each birth date in cohorts of 19-year-olds. Men with lottery numbers below an eligibility ceiling were eligible for the draft, while men with numbers above the ceiling could not be drafted. In practice, many draft-eligible men were still exempted from service for health or other reasons, while many men who were draft-exempt nevertheless volunteered for service. So veteran status was not completely determined by randomized draft-eligibility,

[8] As noted in the introduction to this chapter, measurement error in regressors tends to shrink regression coefficients towards zero. To eliminate this bias, Wald (1940) suggested that the data be divided in a manner independent of the measurement error, and the coefficient of interest estimated as a ratio of differences in means as in (4.1.12). Durbin (1954) showed that Wald's method of fitting straight lines is an IV estimator where the instrument is a dummy marking Wald's division of the data. Hausman (2001) provides an overview of econometric strategies for dealing with measurement error.

Table 4.1.2: Wald estimates of the returns to schooling using quarter of birth instruments

	(1) Born in the 1st or 2nd quarter of year	(2) Born in the 3rd or 4th quarter of year	(3) Difference (std. error) (1)-(2)
ln (weekly wage)	5.8916	5.9051	-0.01349 (0.00337)
Years of education	12.6881	12.8394	-0.1514 (0.0162)
Wald estimate of return to education			0.0891 (0.0210)
OLS estimate of return to education			0.0703 (0.0005)

Notes: Adapted from a re-analysis of Angrist and Krueger (1991) by Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930-39 birth cohorts in the 1980 Census 5 percent file. The sample size is 329,509.

but draft-eligibility provides a binary instrument highly correlated with Vietnam-era veteran status.

For white men who were at risk of being drafted in the 1970 draft lottery, draft-eligibility is clearly associated with lower earnings in years after the lottery. This is documented in Table 4.1.3, which reports the effect of randomized draft-eligibility status on average Social Security-taxable earnings in column 2. column 1 shows average annual earnings for purposes of comparison. For men born in 1950, there are significant negative effects of eligibility status on earnings in 1971, when these men were mostly just beginning their military service, and, perhaps more surprisingly, in 1981, ten years later. In contrast, there is no evidence of an association between draft-eligibility status and earnings in 1969, the year the lottery drawing for men born in 1950 was held but before anyone born in 1950 was actually drafted.

Because eligibility status was randomly assigned, the claim that the estimates in column 2 represent the effect of *draft-eligibility* on earnings seems uncontroversial. The information required to go from draft-eligibility effects to veteran-status effects is the denominator of the Wald estimator, which is the effect of draft-eligibility on the probability of serving in the military. This information is reported in column 3 of Table 4.1.3, which shows that draft-eligible men were almost 16 percentage points more likely to have served in the Vietnam era. The Wald estimate of the effect of military service on 1981 earnings, reported in column 4, amounts to about 15 percent of the mean. Effects were even larger in 1971 (in percentage terms), when affected soldiers were still in the army.

An important feature of the Wald/IV estimator is that the identifying assumptions are easy to assess and

Table 4.1.3: Wald estimates of the effects of military service on the earnings of white men born in 1950

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8	0.267	0.159	-2,741
		(210.5)		(0.040)	(1,324)
1971	3,338	-325.9			-2050
		(46.6)			(293)
1969	2,299	-2.0			
		(34.5)			

Notes: Adapted from Angrist (1990), Tables 2 and 3. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

interpret. Suppose D_i denotes Vietnam-era veteran status and Z_i indicates draft-eligibility. The fundamental claim justifying our interpretation of the Wald estimator as capturing the causal effect of D_i is that the only reason why $E[Y_i|Z_i]$ changes as Z_i changes is the variation in $E[D_i|Z_i]$. A simple check on this is to look for an association between Z_i and personal characteristics that should not be affected by D_i, for example, age, race, sex, or any other characteristic that was determined before D_i was determined. Another useful check is to look for an association between the instrument and outcomes in samples where there is no relationship between D_i and Z_i. If the only reason for draft-eligibility affects on earnings is veteran status, then draft-eligibility effects on earnings should be zero in samples where draft-eligibility status is unrelated to veteran status.

This idea is illustrated in the Angrist (1990) study of the draft lottery by looking at 1969 earnings, an estimate repeated in the last row of Table 4.1.3. It's comforting that the draft-eligibility treatment effect on 1969 earnings is zero since 1969 earnings predate the 1970 draft lottery. A second variation on this idea looks at the cohort of men born in 1953. Although there was a lottery drawing which assigned RSNs to the 1953 birth cohort in February of 1972, no one born in 1953 was actually drafted (the draft officially ended in July of 1973). The first-stage relationship between draft-eligibility and veteran status for men born in 1953 (defined using the 1952 lottery cutoff of 95) therefore shows only a small difference in the probability of serving by eligibility status. Importantly, there is also no significant relationship between earnings and draft-eligibility status for men born in 1953, a result that supports the claim that the only reason for draft-eligibility effects is military service.

We conclude the discussion of Wald estimators with a set of IV estimates of the effect of family size on mothers' employment and work. Like the schooling and military service studies, these estimates are used for illustration elsewhere in the book. The relationship between fertility and labor supply has long been of interest to labor economists, while the case for omitted variables bias in this context is clear: mothers with weak labor force attachment or low earnings potential may be more likely to have children than mothers with strong labor force attachment or high earnings potential. This makes the observed association between family size and employment hard to interpret since mothers who have big families may have worked less anyway. Angrist and Evans (1998) solve this omitted-variables problem using two instrumental variables, both of which lend themselves to Wald-type estimation strategies.

The first Wald estimator uses multiple births, an identification strategy for the effects of family size pioneered by Rosenzweig and Wolpin (1980). The twins instrument in Angrist and Evans (1998) is a dummy for a multiple third birth in a sample of mothers with at least two children. The twins first-stage is .625, an estimate reported in column 3 of Table 4.1.4. This means that 37.5 percent of mothers with two or more children would have had a third birth anyway; a multiple third birth increases this proportion to 1. The twins instrument rests on the idea that the occurrence of a multiple birth is essentially random, unrelated to potential outcomes or demographic characteristics.

The second Wald estimator in Table 4.1.4 uses sibling sex composition, an instrument motivated by the fact that American parents with two children are much more likely to have a third child if the first two are same-sex than if the sex-composition is mixed. This is illustrated in column 5 of Table 4.1.4, which shows that parents of same-sex sibling birth are 6.7 percentage points more likely to have a third birth (the probability of a third birth among parents with a mixed-sex sibship is .38). The same-sex instrument is based on the claim that sibling sex composition is essentially random and affects family labor supply solely by increasing fertility.

Twins and sex-composition instruments both suggest that the birth of a third child has a large effect on employment rates and on weeks and hours worked. Wald estimates using twins instruments show a precisely-estimate employment reduction of about .08, while weeks worked fall by 3.8 and hours per week fall by 3.4. These results, which appear in column 4 of Table 4.1.4, are smaller in absolute value than the corresponding OLS estimates reported in column 2. This suggests the latter are exaggerated by selection bias. Interestingly, the Wald estimates constructed using a same-sex dummy, reported in column 6, are larger than the twins estimates. The juxtaposition of twins and sex-composition instruments in Table 4.1.4 suggests that different instruments need not generate similar estimates of causal effects even if both are valid. We expand on this important point in Section 4.4. For now, however, we stick with a constant-effects framework.

Table 4.1.4: Wald estimates of labor supply effects

			IV Estimates using:			
			Twins		Sex-composition	
Dependent variable	Mean (1)	OLS (2)	First stage (3)	Wald estimates (4)	First stage (5)	Wald estimates (6)
Employment	0.528	-0.167	0.625	-0.083	0.067	-0.135
		(0.002)	(0.011)	(0.017)	(0.002)	(0.029)
Weeks worked	19.0	-8.05	"	-3.83	"	-6.23
		(0.09)	"	(0.758)	"	(1.29)
Hours/week	16.7	-6.02	"	-3.39	"	-5.54
		(0.08)	"	(0.637)	"	(1.08)

Note: The table reports OLS and Wald estimates of the effects of a third birth on labor supply using twins and sex-composition instruments. Data are from the Angrist and Evans (1998) extract including married women aged 21-35 with at least two children in the 1980 Census. OLS models include controls for mother's age, age at first birth, dummies for the sex of first and second births, and dummies for race.

4.1.3 Grouped Data and 2SLS

The Wald estimator is the mother of all instrumental variables estimators because more complicated 2SLS estimators can typically be constructed from an underlying set of Wald estimators. The link between Wald and 2SLS is grouped-data: 2SLS using dummy instruments is the same thing as GLS on a set of group means. GLS in turn can be understood as a linear combination of all the Wald estimators that can be constructed from pairs of means. The generality of this link might appear to be limited by the presumption that the instruments at hand are dummies. Not all instrumental variables are dummies, or even discrete, but this is not really important. For one thing, many credible instruments can be thought of as defining categories, such as quarter of birth. Moreover, instrumental variables that appear more continuous (such as draft lottery numbers, which range from 1-365) can usually be grouped without much loss of information (for example, a single dummy for draft-eligibility status, or dummies for groups of 25 lottery numbers).[9]

To explain the Wald/grouping/2SLS nexus more fully, we stick with the draft-lottery study. Earlier we noted that draft-eligibility is a promising instrument for Vietnam-era veteran status. The draft-eligibility ceilings were RSN 195 for men born in 1950, RSN 125 for men born in 1951, and RSN 95 for men born in 1952. In practice, however, there is a richer link between draft lottery numbers (which we'll call R_i, short for RSN) and veteran status (D_i) than draft-eligibility status alone. Although men with numbers above the eligibility ceiling were not drafted, the ceiling was unknown in advance. Some men therefore volunteered in the hope of serving under better terms and gaining some control over the timing of their service. The pressure to become a draft-induced volunteer was high for men with low lottery numbers, but low for men with high numbers. As a result, there is variation in $P[\text{D}_i = 1|\text{R}_i]$ even for values strictly above or below the draft-eligibility cutoff. For example, men born in 1950 with lottery numbers 200 – 225 were more likely to serve than those with lottery numbers 226 – 250, though ultimately no one in either group was drafted.

The Wald estimator using draft-eligibility as an instrument for men born in 1950 compares the earnings of men with $\text{R}_i < 195$ to the earnings of men with $\text{R}_i > 195$. But the previous discussion suggests the possibility of many more comparisons, for example men with $\text{R}_i \leq 25$ vs. men with $\text{R}_i \in [26 - 50]$; men with $\text{R}_i \in [51 - 75]$ vs. men with $\text{R}_i \in [76 - 100]$, and so on, until these 25-number intervals are exhausted. We might also make the intervals finer, comparing, say, men in 5-number or single-number intervals instead of 25-number intervals. The result of this expansion in the set of comparisons is a set of Wald estimators. These sets are complete in that the intervals partition the support of the underlying instrument, while the individual estimators are linearly independent in the sense that their numerators are linearly independent. Finally, each of these Wald estimators consistently estimates the same causal effect, assumed here to be constant, as long as R_i is independent of potential outcomes and correlated with veteran status (i.e., the Wald denominators are not zero).

[9] An exception is the classical measurement error model, where both the variable to be instrument and the instrument are assumed to be continuous. Here, we have in mind IV scenarios involving omitted variables bias.

The possibility of constructing multiple Wald estimators for the same causal effect naturally raises the question of what to do with all of them. We would like to come up with a single estimate that somehow combines the information in the individual Wald estimates efficiently. As it turns out, the most efficient linear combination of a full set of linearly independent Wald estimates is produced by fitting a line through the group means used to construct these estimates.

The grouped data estimator can be motivated directly as follows. As in (4.1.11), we work with a bivariate constant-effects model, which in this case can be written

$$\text{Y}_i = \alpha + \rho \text{D}_i + \eta_i, \tag{4.1.14}$$

where $\rho = \text{Y}_{1i} - \text{Y}_{0i}$ is the causal effect of interest and $\text{Y}_{0i} = \alpha + \eta_i$. Because R_i was randomly assigned and lottery numbers are assumed to have no effect on earnings other than through veteran status, $E[\eta_i|\text{R}_i] = 0$. It therefore follows that

$$E[\text{Y}_i|\text{R}_i] = \alpha + \rho P[\text{D}_i = 1|\text{R}_i], \tag{4.1.15}$$

since $P[\text{D}_i = 1|\text{R}_i] = E[\text{D}_i|\text{R}_i]$. In other words, the slope of the line connecting average earnings given lottery number with the average probability of service by lottery number is equal to the effect of military service, ρ. This is in spite of the fact that the regression Y_i on D_i—in this case, the difference in means by veteran status—almost certainly differs from ρ since Y_{0i} and D_i are likely to be correlated.

Equation (4.1.15) suggests an estimation strategy based on fitting a line to the sample analog of $E[\text{Y}_i|\text{R}_i]$ and $P[\text{D}_i = 1|\text{R}_i]$. Suppose that R_i takes on values $j = 1, ..., \text{J}$. In principle, j might run from 1 to 365, but in Angrist (1990), lottery-number information was aggregated to 69 five-number intervals, plus a 70th for numbers 346-365. We can therefore think of R_i as running from 1 to 70. Let $\bar{y}_j$ and $\hat{p}_j$ denote estimates of $E[\text{Y}_i|\text{R}_i = j]$ and $P[\text{D}_i = 1|\text{R}_i = j]$, while $\bar{\eta}_j$ denotes the average error in (4.1.14). Because sample moments converge to population moments it follows that OLS estimates of ρ in the grouped equation

$$\bar{y}_j = \alpha + \rho \hat{p}_j + \bar{\eta}_j \tag{4.1.16}$$

are consistent. In practice, however, GLS may be preferable since a grouped equation is heteroskedastic with a known variance structure. The efficient GLS estimator for grouped data in a constant-effects linear model is weighted least squares, weighted by the variance of $\bar{\eta}_j$ (see, e.g., Prais and Aitchison, 1954 or Wooldridge, 2006). Assuming the microdata residual is homoskedastic with variance σ^2_η, this variance is $\frac{\sigma^2_\eta}{n_j}$, where n_j is the group size.

The GLS (or weighted least squares) estimator of ρ in equation (4.1.16) is especially important in this context for two reasons. First, the GLS slope estimate constructed from J grouped observations is an asymptotically efficient linear combination of any full set of J−1 linearly independent Wald estimators

(Angrist, 1991). This can be seen without any mathematics: GLS and any linear combination of pairwise Wald estimators are both linear combinations of the grouped dependent variable. Moreover, GLS is the asymptotically efficient linear estimator for grouped data. Therefore we can conclude that there is no better (i.e., asymptotically more efficient) linear combination of Wald estimators than GLS (again, a maintained assumption here is that ρ is constant). The formula for constructing the GLS estimator from a full set of linearly independent Wald estimators appears in Angrist (1988).

Second, just as each Wald estimator is also an IV estimator, the GLS (weighted least squares) estimator of equation (4.1.16) is also 2SLS. The instruments in this case are a full set of dummies to indicate each lottery-number cell. To see why, define the set of dummy instruments $Z_i \equiv \{r_{ji} = 1[\text{R}_i = j]; j = 1, ...J-1\}$. Now, consider the first stage regression of D_i on Z_i plus a constant. Since this first stage is saturated, the fitted values will be the sample conditional means, $\hat{p}_j$, repeated n_j times for each j. The second stage slope estimate is therefore exactly the same as weighted least squares estimation of the grouped equation, (4.1.16), weighted by the cell size, n_j.

The connection between grouped-data and 2SLS is of both conceptual and practical importance. On the conceptual side, any 2SLS estimator using a set of dummy instruments can be understood as a linear combination of all the Wald estimators generated by these instruments one at a time. The Wald estimator in turn provides a simple framework used later in this chapter to interpret IV estimates in the much more realistic world of heterogeneous potential outcomes.

Although not all instruments are inherently discrete and therefore immediately amenable to a Wald or grouped-data interpretation, many are. Examples include the draft lottery number, quarter of birth, twins, and sibling-sex composition instruments we've already discussed. See also the recent studies by Bennedsen, *et al.,* 2007, and Ananat and Michaels, 2008, both of which use dummies for male first births as instruments. Moreover, instruments that have a continuous flavor can often be fruitfully turned into discrete variables. For example, Angrist, Graddy and Imbens (2000) group continuous weather-based instruments into 3 dummy variables, *stormy*, *mixed*, and *clear*, which they then use to estimate the demand fish. This dummy-variable parameterization seems to capture the main features of the relationship between weather conditions and the price of fish.[10]

On the practical side, the grouped-data equivalent of 2SLS gives us a simple tool that can be used to explain and evaluate any IV strategy. In the case of the draft lottery, for example, the grouped model embodies the assumption that the only reason average earnings vary with lottery numbers is the variation in probability of service across lottery-number groups. If the underlying causal relation is linear with constant effects, then equation (4.1.16) should fit the group means well, something we can assess by inspection and, as discussed in the next section, with the machinery of formal statistical inference.

[10] Continuous instruments recoded as dummies can be seen as providing a parsimonious non-parametric model for the underlying first-stage relation, $E[\text{D}_i|\text{Z}_i]$. In homoskedastic models with constant coefficients, the asymptotically efficient instrument is $E[\text{D}_i|\text{Z}_i]$ (Newey, 1990).

Sometimes labor economists refer to grouped-data plots for discrete instruments as Visual Instrumental Variables (VIV).[11] An example appears in Angrist (1990), reproduced here as Figure 4.1.2. This figure shows the relationship between average earnings in 5-number RSN cells and the probability of service in these cells, for the 1981-84 earnings of white men born 1950-53. The slope of the line through these points is an IV estimate of the earnings loss due to military service, in this case about $2,400, not very different from the Wald estimates discussed earlier but with a lower standard error (in this case, about $800).

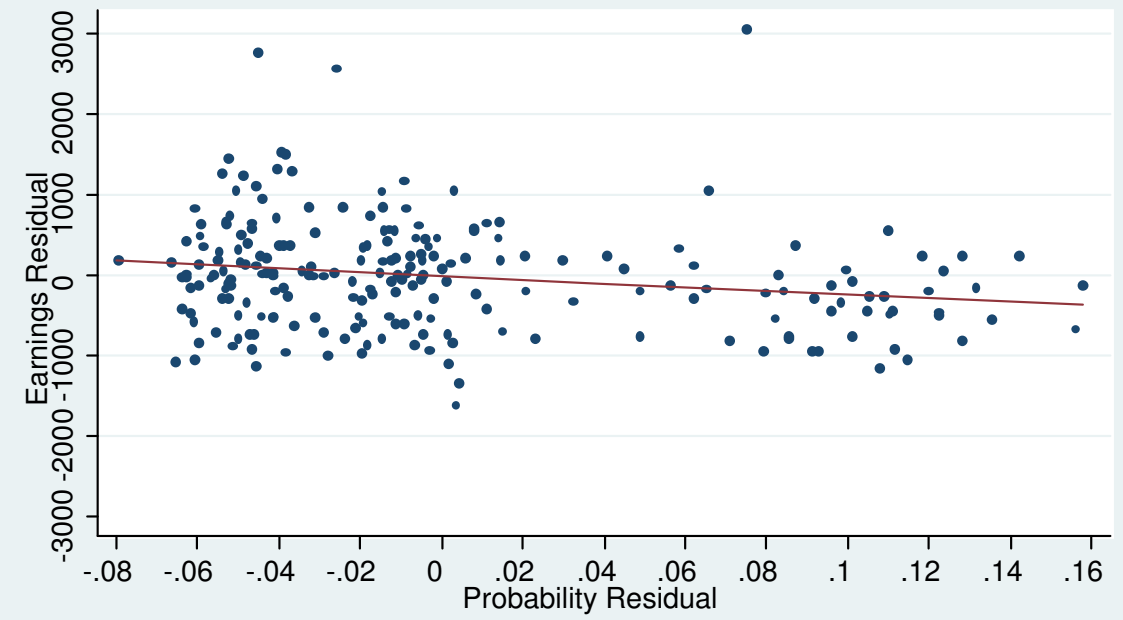

Figure 4.1.2: The relationship between average earnings and the probability of military service (from Angrist 1990). This is a VIV plot of average 1981-84 earnings by cohort and groups of five consecutive draft lottery numbers against conditional probabilities of veteran status in the same cells. The sample includes white men born 1950-53. Plotted points consist of average residuals (over four years of earnings) from regressions on period and cohort effects. The slope of the least-squares regression line drawn through the points is -2,384, with a standard error of 778.

4.2 Asymptotic 2SLS Inference

4.2.1 The Limiting Distribution of the 2SLS Coefficient Vector

We can derive the limiting distribution of the 2SLS coefficient vector using an argument similar to that used in Section 3.1.3 for OLS. In this case, let $V_i \equiv \begin{bmatrix} \mathrm{X}_i' & \hat{s}_i \end{bmatrix}'$ denote the vector of regressors in the 2SLS second

[11] See, e.g., the preface to Borjas (2005).

stage, equation (4.1.9). The 2SLS estimator can then be written

$$\hat{\Gamma}_{2SLS} \equiv \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \mathrm{Y}_i,$$

where $\Gamma \equiv \begin{bmatrix} \alpha' & \rho \end{bmatrix}'$ is the corresponding coefficient vector. Note that

$$\begin{aligned} \hat{\Gamma}_{2SLS} &= \Gamma + \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i[\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)] \\ &= \Gamma + \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \eta_i \end{aligned} \tag{4.2.1}$$

where the second equality comes from the fact that the first-stage residuals, $(\mathrm{S}_i - \hat{s}_i)$, are orthogonal to V_i in the sample. The limiting distribution of the 2SLS coefficient vector is therefore the limiting distribution of $\left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \eta_i$. This quantity is a little harder to work with than the corresponding OLS quantity, because the regressors in this case involve estimated fitted values, $\hat{s}_i$. A Slutsky-type argument shows, however, that we get the same limiting distribution replacing estimated fitted values with the corresponding population fitted values (i.e., replacing $\hat{s}_i$ with $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$). It therefore follows that $\hat{\Gamma}_{2SLS}$ has an asymptotically normal distribution, with probability limit Γ, and a covariance matrix estimated consistently by $\left[\sum_i V_i V_i'\right]^{-1} \left[\sum_i V_i V_i' \eta_i^2\right] \left[\sum_i V_i V_i'\right]^{-1}$. This is a sandwich formula like the one for OLS standard errors (White, 1982). As with OLS, if η_i is conditionally homoskedastic given covariates and instruments, the consistent covariance matrix estimator simplifies to $\left[\sum_i V_i V_i'\right]^{-1} \sigma_\eta^2$.

There is little new here, but there is one tricky point. It seems natural to construct 2SLS estimates manually by first estimating the first stage (4.1.4a) and then plugging the fitted values into equation (4.1.9) and estimating this by OLS. That's fine as far as the coefficient estimates go, but the resulting standard errors will be incorrect. Conventional regression software does not know that you are trying to construct a 2SLS estimate. The residual variance estimator that goes into the standard formulas will therefore be incorrect. When constructing standard errors, the software will estimate the residual variance of the equation you estimate by OLS in the second stage:

$$\mathrm{Y}_i - [\alpha'\mathrm{X}_i + \rho\hat{s}_i] = [\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)],$$

replacing the coefficients with the corresponding estimates. The correct residual variance estimator, however, uses the original endogenous regressor to construct residuals and not the first-stage fitted values, $\hat{s}_i$. In other words, the residual you want is $\mathrm{Y}_i - [\alpha'\mathrm{X}_i + \rho\mathrm{S}_i] = \eta_i$, so as to consistently estimate σ_η^2, and not $\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)$. Although this problem is easy to fix (you can construct the appropriate residual variance estimator in a separate calculation), software designed for 2SLS gets this right automatically, and may help

you avoid other common 2SLS mistakes.

4.2.2 Over-identification and the 2SLS Minimand★

Constant-effects models with more instruments than endogenous regressors are said to be over-identified. Because there are more instruments than needed to identify the parameters of interest, these models impose a set of restrictions that can be evaluated as part of a process of specification testing. This process amounts to asking whether the line plotted in a VIV-type picture fits the relevant conditional means tightly enough given the precision with which the means are estimated. The details behind this useful idea are easiest to spell out using matrix notation and a traditional linear model.

Let $Z_i \equiv \left[\begin{array}{cccc} \mathrm{X}_i' & \mathrm{Z}_{1i} & ... & \mathrm{Z}_{\mathrm{Q}i} \end{array} \right]'$ denote the vector formed by concatenating the exogenous covariates and Q instrumental variables and let $W_i \equiv \left[\begin{array}{cc} \mathrm{X}_i' & \mathrm{S}_i \end{array} \right]'$ denote the vector formed by concatenating the covariates and the single endogenous variable of interest. In the quarter-of-birth paper, for example, the covariates are year-of-birth and state-of-birth dummies, the instruments are quarter-of-birth dummies, and the endogenous variable is schooling. The coefficient vector is still $\Gamma \equiv [\alpha', \rho]'$, as in the previous subsection. The residuals for the causal model can be defined as a function of Γ using

$$\eta_i(\Gamma) \equiv \mathrm{Y}_i - \Gamma' W_i = \mathrm{Y}_i - [\alpha' \mathrm{X}_i + \rho \mathrm{S}_i] .$$

This residual is assumed to be uncorrelated with the instrument vector, Z_i. In other words, η_i satisfies the orthogonality condition,

$$E[\mathrm{Z}_i \eta_i(\Gamma)] = 0. \tag{4.2.2}$$

In any sample, however, this equation will not hold exactly because there are more moment conditions than there are elements of Γ.[12] The sample analog of (4.2.2) is the sum over i,

$$\frac{1}{N} \sum Z_i \eta_i(\Gamma) \equiv m_N(\Gamma). \tag{4.2.3}$$

2SLS can be understood as a generalized method of moments (GMM) estimator that chooses a value for Γ by making the sample analog of (4.2.2) as close to zero as possible.

By the central limit theorem, the sample moment vector $\sqrt{N} m_N(\Gamma)$ has an asymptotic covariance matrix equal to $E[Z_i Z_i' \eta_i(\Gamma)^2]$, a matrix we'll call Λ. Although somewhat intimidating at first blush, this is just a matrix of 4th moments, as in the sandwich formula used to construct robust standard errors, (3.1.7). As shown by Hansen (1982), the optimal GMM estimator based on (4.2.2) minimizes a quadratic form in the sample moment vector, $m_N(\hat{g})$, where $\hat{g}$ is a candidate estimator of Γ.[13] The optimal weighting matrix in

[12] With a single endogenous variable and more than one instrument, Γ is $[\mathrm{K}+1] \times 1$, while Z_i is $[\mathrm{K}+\mathrm{Q}] \times 1$ for $\mathrm{Q} > 1$. Hence the resulting linear system cannot be solved unless there is a linear dependency that makes some of the instruments redundant.

[13] "Quadratic form" is matrix language for a weighted sum of squares. Suppose v is an $N \times 1$ vector and M is an $N \times N$

the middle of the GMM quadratic form is Λ^{-1}. In practice, of course, Λ, is unknown and must be estimated. A feasible version of the GMM procedure uses a consistent estimator of Λ in the weighting matrix. Since the estimator using known and estimated Λ have the same limiting distribution, we'll ignore this distinction for now. The quadratic form to be minimized can therefore be written,

$$J_N(\hat{g}) \equiv N m_N(\hat{g})' \Lambda^{-1} m_N(\hat{g}), \tag{4.2.4}$$

where the N-term out front comes from $\sqrt{N}$ normalization of the sample moments. As shown immediately below, when the residuals are conditionally homoskedastic, the minimizer of $J_N(\hat{g})$ is the 2SLS estimator. Without homoskedasticity, the GMM estimator that minimizes (4.2.4) is White's (1982) Two-Stage IV (a generalization of 2SLS) so that it makes sense to call $J_N(\hat{g})$ the "2SLS minimand".

Here are some of the details behind the GMM interpretation of 2SLS[14]. Conditional homoskedasticity means that

$$E[Z_i Z_i' \eta_i (\Gamma)^2] = E[Z_i Z_i'] \sigma_\eta^2.$$

Substituting for Λ^{-1} and using Z,Y and W to denote sample data vectors and matrices, the quadratic form to be minimized becomes

$$J_N(\hat{g}) = (N\sigma_\eta^2)^{-1} \times (\mathrm{Y} - W\hat{g})' Z E[Z_i Z_i']^{-1} Z' (\mathrm{Y} - W\hat{g}). \tag{4.2.5}$$

Finally, substituting the sample cross-product matrix $\left[\frac{Z'Z}{N}\right]$ for $E[Z_i Z_i']$, we have

$$\hat{J}_N(\hat{g}) = (1/\sigma_\eta^2) \times (\mathrm{Y} - W\hat{g})' P_Z (\mathrm{Y} - W\hat{g}),$$

where $P_Z = Z(Z'Z)^{-1}Z$. From here, we get the solution

$$\hat{g} = \hat{\Gamma}_{2SLS} = [W' P_Z W]^{-1} W' P_Z \mathrm{Y}.$$

Since the projection operator, P_Z, produces fitted values, and P_Z is an idempotent matrix, this can be seen to be the OLS estimator of the second-stage equation, (4.1.9), written in matrix notation. More generally, even without homoskedasticity we can obtain a feasible efficient 2SLS-type estimator by minimizing (4.2.4) and using a consistent estimator of $E[Z_i Z_i' \eta_i(\hat{g})^2]$ to form $\hat{J}_N(\hat{g})$. Typically, we'd use the empirical fourth moments, $\sum Z_i Z_i' \hat{\eta}_i^2$, where $\hat{\eta}_i$ is the regular 2SLS residual computed without worrying about heteroskedasticity (see, White, 1982, for distribution theory and other details).

---

matrix. A quadratic form in v is $v'Mv$. If M is a $N \times N$ diagonal matrix with diagonal elements m_i, then $v'Mv = \sum_i m_i v_i^2$.

[14] Much more detailed explanations can be found in Newey (1985), Newey and West (1987), and the original Hansen (1982) GMM paper.

The over-identification test statistic is given by the minimized 2SLS minimand. Intuitively, this statistic tells us whether the sample moment vector, $m_N(\hat{g})$, is close enough to zero for the assumption that $E[Z_i\eta_i] = 0$ to be plausible. In particular, under the null hypothesis that the residuals and instruments are indeed orthogonal, the minimized $J_N(\hat{g})$ has a $\chi^2(\mathrm{Q}-1)$ distribution. We can therefore compare the empirical value of the 2SLS minimand with chi-square tables in a formal testing procedure for $H_0 : E[Z_i\eta_i] = 0$.

For reasons that will soon become apparent, we're not often interested in over-identification *per se*. Our main interest is in the 2SLS minimand when the instruments are a full set of mutually exclusive dummy variables, as for the Wald estimators and grouped-data estimation strategies discussed above. In this important special case, the 2SLS becomes weighted least squares of a grouped equation like (4.1.16), while the 2SLS minimand is the relevant weighted sum of squares being minimized. To see this, note that projection on a full set of mutually exclusive dummy variables for an instrument that takes on J values produces an $N \times 1$ vector of fitted values equal to the J conditional means at each value of the instrument (included covariates are counted as instruments), each one of these n_j times, where n_j is the group size and $\sum n_j = N$. The cross product matrix $[Z'Z]$ in this case is a J×J diagonal matrix with elements n_j. Simplifying, we then have

$$\hat{J}_N(\hat{g}) = (1/\sigma_\eta^2) \times \sum_j n_j(\bar{y}_j - \hat{g}'\bar{W}_j)^2, \tag{4.2.6}$$

where $\bar{W}_j$ is the sample mean of the rows of matrix W in group j. Thus, $\hat{J}_N(\hat{g})$ is the GLS weighted least squares minimand for estimation of the grouped regression: $\bar{y}_j$ on $\bar{W}_j$. With a little bit more work (here we skip the details), we can similarly show that the efficient Two-Step IV procedure without homoskedasticity minimizes

$$\hat{J}_N(\hat{g}) = \sum_j \left(\frac{n_j}{\sigma_j^2}\right)(\bar{y}_j - \hat{g}'\bar{W}_j)^2, \tag{4.2.7}$$

where σ_j^2 is the variance of η_i in group j. Estimation using (4.2.7) is feasible because we can estimate σ_j^2 in a first-step, say, using inefficient-but-still-consistent 2SLS that ignores heteroskedasticity. Efficient two-step IV estimators are constructed in Angrist (1990, 1991).

The GLS structure of the 2SLS minimand allows us to see the over-identification test statistic for dummy instruments as a simple measure of the goodness of fit of the line connecting $\bar{y}_j$ and $\bar{W}_j$. In other words, this is the chi-square goodness of fit statistic for the line in a VIV plot like figure 4.1.2. The chi-square degrees of freedom parameter is given by the difference between the number of values taken on by the instrument and the number of parameters being estimated[15].

Like the various paths leading to the 2SLS estimator, there are many roads to the test-statistic, (4.2.7), as well. Here are two further paths that are worth knowing. First, the test-statistic based on the general GMM minimand for IV, whether the instruments are group dummies or not, is the same as the over-

[15] If, for example, the instrument takes on three values, one of which is assigned to the constant, and the model includes a constant and a single the endogenous variable only, the test statistic has 1 degree of freedom.

identification test statistic discussed in many widely-used econometric references on simultaneous equations models. For example, this statistic features in Hausman's (1983) chapter on simultaneous equations in the *Handbook of Econometrics*, which also proposes a simple computational procedure: for homoskedastic models, the minimized 2SLS minimand is the sample size times the R^2 from a regression of the 2SLS residuals on the instruments (and the included exogenous covariates). The formula for this is $N\left[\frac{\hat{\eta}'P_Z\hat{\eta}}{\hat{\eta}'\hat{\eta}}\right]$, where $\hat{\eta} = \text{Y} - W\hat{\Gamma}_{2SLS}$ is the vector of 2SLS residuals.

Second, it's worth emphasizing that the essence of over-identification can be said to be "more than one way to skin the same econometric cat." In other words, given more than one instrument for the same causal relation, we might consider constructing simple IV estimators one at a time and comparing them. This comparison checks over-identification directly: If each just-identified estimator is consistent, the distance between them should be small relative to sampling variance, and should shrink as the sample size and hence the precision of these estimates increases. In fact, we might consider formally testing whether all possible just-identified estimators are the same. The resulting test statistic is said to generate a Wald[16] test of this null, while the test-statistic based on the 2SLS minimand is said to be a Lagrange Multiplier (LM) test because it can be related to the score vector in a maximum likelihood version of the IV setup.

In the grouped-data version of IV, the Wald test amounts to a test of equality for the set of all possible linearly independent Wald estimators. If, for example, lottery numbers are divided into 4 groups based on various cohorts eligibility cutoffs (RSN 1-95, 96-125, 126-195, and the rest), then 3 linearly independent Wald estimators can be constructed. Alternatively, the efficient grouped-data estimator can be constructed by running GLS on these four conditional means. Four groups means there are 3 possible Wald estimators and 2 non-redundant equality restrictions on these three; hence, the relevant *Wald statistic* has 2 degrees of freedom. On the other hand, 4 groups means three instruments and a constant available to estimate a model with 2 parameters (the constant and the causal effect of military service). So the 2SLS minimand generates an over-identification test statistic with $4 - 2 = 2$ degrees of freedom. And, in fact, provided you use the same method of estimating the weighting matrix in the relevant quadratic forms, these two test statistics not only test the same thing, they are numerically equivalent. This makes sense since we have already seen that 2SLS is the efficient linear combination of Wald estimators.[17]

Finally, a caveat regarding over-identification tests in practice: In our experience, the "over-ID statistic" is often of little value in applied work. Because $J_N(\hat{g})$ measures variance-normalized goodness of-fit, the over-ID test-statistic tends to be low when the underlying estimates are imprecise. Since IV estimates are very often imprecise, we cannot take much satisfaction from the fact that one estimate is within sampling variance of another even if the individual estimates appear precise enough to be informative. On the other

[16] The Wald estimator and Wald test are named after the same statistician, Abraham Wald, but the latter reference is Wald (1943).

[17] The fact that Wald and LM testing procedures for the same null are equivalent in linear models was established by Newey and West (1987). Angrist (1991) gives a formal statement of the argument in this paragraph.

hand, in cases where the underlying IV estimates are quite precise, the fact that the over-ID statistic rejects need not point to an identification failure. Rather, this may be evidence of treatment effect heterogeneity, a possibility we discuss further below. On the conceptual side, however, an understanding of the anatomy of the 2SLS minimand is invaluable, for it once again highlights the important link between grouped data and IV. This link takes the mystery out of estimation and testing with instrumental variables and forces us to confront the raw moments that are the foundation for causal inference.

4.3 Two-Sample IV and Split-Sample IV★

The GMM interpretation of 2SLS highlights the fact that the IV estimator can be constructed from sample moments alone, with no micro data. Returning to the sample moment condition, (4.2.3), and re-arranging slightly produces a regression-like equation involving second moments:

$$\frac{Z'\mathrm{Y}}{N} = \frac{Z'W}{N}\Gamma + \frac{Z'\eta}{N} \tag{4.3.1}$$

GLS estimates of Γ in (4.3.1) are consistent because $E\left[\frac{Z'\mathrm{Y}}{N}\right] = E\left[\frac{Z'W}{N}\right]\Gamma$.

The 2SLS minimand can be thought of as GLS applied to equation (4.3.1), after multiplying by $\sqrt{N}$ to keep the residual from disappearing as the sample size gets large. In other words, 2SLS minimizes a quadratic form in the residuals from (4.3.1) with a (possibly non-diagonal) weighting matrix.[18] An important insight that comes from writing the 2SLS problem in this way is that we do not need the individual observations in our sample to estimate (4.3.1). Just as with the OLS coefficient vector, which can be constructed from the sample conditional mean function, IV estimators can also be constructed from sample moments. The moments needed for IV are $\frac{Z'\mathrm{Y}}{N}$ and $\frac{Z'W}{N}$. The dependent variable, $\frac{Z'\mathrm{Y}}{N}$, is a vector of dimension $[\mathrm{K}+\mathrm{Q}] \times 1$. The regressor matrix, $\frac{Z'W}{N}$, is of dimension $[\mathrm{K}+\mathrm{Q}] \times [\mathrm{K}+1]$. The second-moment equation cannot be solved exactly unless $\mathrm{Q}= 1$ so it makes sense to make the fit as good as possible by minimizing a quadratic form in the residuals. The most efficient weighting matrix for this purpose is the asymptotic covariance matrix of $\frac{Z'\eta}{\sqrt{N}}$. This again produces the 2SLS minimand, $\hat{J}_N(\hat{g})$.

A related insight is the fact that the moment matrices on the left and right hand side of the equals sign in equation (4.3.1) need not come from the same data sets provided these data sets are drawn from the same population. This observation leads to the two-sample instrumental variables (TSIV) estimator used by Angrist (1990) and developed formally in Angrist and Krueger (1992)[19]. Briefly, let Z_1 and Y_1 denote

[18] A quadratic form is the matrix-weighted product, $x'Ax$, where x is a random vector of, say, dimension K and A is a $\mathrm{K}\times\mathrm{K}$ matrix of constants.

[19] Applications of TSIV include Bjorklund and Jantti (1997), Jappelli, Pischke, and Souleles (1998), Currie and Yelowitz (2000), and Dee and Evans (2003). In a recent paper, Inoue and Solon (2005) compare the asymptotic distributions of alternative TSIV estimators, and introduce a maximum likelihood (LIML-type) version of TSIV. They also correct a mistake in the distribution theory in Angrist and Krueger (1995), discussed further, below.

the instrument/covariate matrix and dependent variable vector in data set 1 of size N_1 and let Z_2 and W_2 denote the instrument /covariate matrix and endogenous variable/covariate matrix in data set 2 of size N_2. Assuming $plim\left(\frac{Z_2'W_2}{N_2}\right) = plim\left(\frac{Z_1'W_1}{N_1}\right)$, GLS estimates of the two-sample moment equation

$$\frac{Z_1'\mathrm{Y}_1}{N_1} = \frac{Z_2'W_2}{N_2}\Gamma + \left\{\left[\frac{Z_1'W_1}{N_1} - \frac{Z_2'W_2}{N_2}\right]\Gamma + \frac{Z_1'\eta_1}{N_1}\right\}$$

are also consistent for Γ. The limiting distribution of this estimator is obtained by normalizing by $\sqrt{N_1}$ and assuming $plim\left(\frac{N_2}{N_1}\right)$ is a constant.

The utility of TSIV comes from the fact that it widens the scope for IV estimation to situations where observations on dependent variables, instruments, and the endogenous variable of interest are hard to find in a single sample. It may be easier to find one data set that has information on outcomes and instruments, with which the reduced form can be estimated, and another data set which has information on endogenous variables and instruments, with which the first stage can be estimated. For example, in Angrist (1990), administrative records from the Social Security Administration (SSA) provide information on the dependent variable (annual earnings) and the instruments (draft lottery numbers coded from dates of birth, as well as covariates for race and year of birth). The SSA, however, does not track participants' veteran status. This information was taken from military records, which also contain dates of birth that can used to code lottery numbers. Angrist (1990) used these military records to construct $\frac{Z_2'W_2}{N_2}$, the first-stage correlation between lottery numbers and veteran status conditional on race and year of birth, while the SSA data were used to construct $\frac{Z_1'\mathrm{Y}_1}{N_1}$.

Two further simplifications make TSIV especially easy to use. First, as noted previously, when the instruments consist of a full set of mutually exclusive dummy variables, as in Angrist (1990) and Angrist and Krueger (1992), the second moment equation, (4.3.1), simplifies to a model for conditional means. In particular, the 2SLS minimand for the two-sample problem becomes

$$\hat{J}_N(\hat{g}) = \sum_j \omega_j \left(\bar{y}_{1j} - \hat{g}'\bar{W}_{2j}\right)^2, \tag{4.3.2}$$

where $\bar{y}_{1j}$ is the mean of the dependent variable at instrument/covariate value j in one sample, $\bar{W}_{2j}$ is the mean of endogenous variables and covariates at instrument/covariate value j in a second sample, and ω_j is an appropriate weight. This amounts to weighted least squares estimation of the VIV equation, except that the dependent and independent variables do not come from the same sample. Again, Angrist (1990) and Angrist and Krueger (1992) provide illustrations. The optimal weights for asymptotically efficient TSIV are given by variance of $\bar{y}_{1j} - \hat{g}'\bar{W}_{2j}$. This variance is affected by the fact that moments come from different samples, as are the TSIV standard errors, which are easy to compute in the dummy-instrument case since the estimator is equivalent to weighted least squares.

Second, Angrist and Krueger (1995) introduced a computationally attractive TSIV-type estimator that requires no matrix manipulation and can be implemented with ordinary regression software. This estimator, called Split-Sample IV (SSIV), works as follows.[20] The first-stage estimates in data set two are given by $(Z_2'Z_2)^{-1}Z_2'W_2$. These fitted values can be carried over to data set 1 by constructing the *cross-sample fitted value*, $\hat{W}_{12} \equiv Z_1(Z_2'Z_2)^{-1}Z_2'W_2$. The SSIV second stage is a regression of Y_1 on $\hat{W}_{12}$. The correct limiting distribution for this estimator is derived in Inoue and Solon (2005), who show that the limiting distribution presented in Angrist and Krueger (1992) requires the assumption that $Z_1'Z_1 = Z_2'Z_2$ (as would be true if the marginal distribution of the instruments and covariates is fixed in repeated samples). It's worth noting, however, that the limiting distributions of SSIV and 2SLS are the same when the coefficient on the endogenous variable is zero. The standard errors for this special case are simple to construct and probably provide a reasonably good approximation to the general case.[21]

4.4 IV with Heterogeneous Potential Outcomes

The discussion of IV up to this point postulates a constant causal effect. In the case of a dummy variable like veteran status, this means $\mathrm{Y}_{1i}-\mathrm{Y}_{0i} = \rho$ for all i, while with a multi-valued treatment like schooling, this means $Y_{si} - Y_{s-1,i} = \rho$ for all s and all i. Both are highly stylized views of the world, especially the multi-valued case which imposes linearity as well as homogeneity. To focus on one thing at a time in a heterogeneous-effects model, we start with a zero-one causal variable. In this context, we'd like to allow for treatment-effect heterogeneity, in other words, a distribution of causal effects across individuals.

Why is treatment-effect heterogeneity important? The answer lies in the distinction between the two types of validity that characterize a research design. *Internal validity* is the question of whether a given design successfully uncovers causal effects for the population being studied. A randomized clinical trial or, for that matter, a good IV study, has a strong claim to internal validity. *External validity* is the predictive value of the study's findings in a different context. For example, if the study population in a randomized trial is especially likely to benefit from treatment, the resulting estimates may have little external validity. Likewise,

[20] Angrist and Krueger called this estimator SSIV because they were concerned with a scenario where a single data set is deliberately split in two. As discussed in Section (4.6.4), the resulting estimator may have less bias than conventional 2SLS. Inoue and Solon (2005) refer to the estimator Angrist and Krueger (1995) called SSIV as Two-sample 2SLS or TS2SLS.

[21] This shortcut formula uses the standard errors from the manual SSIV second stage. The correct asymptotic covariance matrix formula, from Inoue and Solon (2005), is

$$\{B[(\sigma_{11} + \kappa\Gamma'\Sigma_{22}\Gamma)A]^{-1}B\}^{-1}$$

where B=$plim\left(\frac{Z_2'W_2}{N_2}\right) = plim\left(\frac{Z_1'W_1}{N_1}\right)$, $A = plim\left(\frac{Z_1'Z_1}{N_1}\right) = plim\left(\frac{Z_2Z_2}{N_2}\right)$, $plim\left(\frac{N_2}{N_1}\right) = \kappa$, σ_{11} is the variance of the reduced-form residual in data set 1, and Σ_{22} is the variance of the first-stage residual in data set 2. In principle, these pieces are easy enough to calculate. Other approaches to SSIV inference include those of Dee and Evans (2003), who calculate standard errors for just-identified models using the delta-method, and Bjorklund and Jantti (1997), who use a bootstrap.

draft-lottery estimates of the effects of conscription for service in the Vietnam era need not be a good measure of the consequences of voluntary military service. An econometric framework with heterogeneous treatment effects helps us to assess both the internal and external validity of IV estimates.[22]

4.4.1 Local Average Treatment Effects

In an IV framework, the engine that drives causal inference is the instrument, Z_i, but the variable of interest is still D_i. This feature of the IV setup leads us to adopt a generalized potential-outcomes concept, indexed against both instruments and treatment status. Let $\text{Y}_i(d, z)$ denote the potential outcome of individual i were this person to have treatment status $\text{D}_i = d$ and instrument value $\text{Z}_i = z$. This tells us, for example, what the earnings of i would be given alternative combinations of veteran status and draft-eligibility status. The causal effect of veteran status given i's realized draft-eligibility status is $\text{Y}_i(1,\text{Z}_i)-\text{Y}_i(0,\text{Z}_i)$, while the causal effect of draft-eligibility status given i's veteran status is $\text{Y}_i(\text{D}_i, 1)-\text{Y}_i(\text{D}_i, 0)$.

We can think of instrumental variables as initiating a causal chain where the instrument, Z_i, affects the variable of interest, D_i, which in turn affects outcomes, Y_i. To make this precise, we need notation to express the idea that the instrument has a causal effect on D_i. Let D_{1i} be i's treatment status when $\text{Z}_i = 1$, while D_{0i} is i's treatment status when $\text{Z}_i = 0$. Observed treatment status is therefore

$$\text{D}_i = \text{D}_{0i} + (\text{D}_{1i} - \text{D}_{0i})\text{Z}_i = \pi_0 + \pi_{1i}\text{Z}_i + \xi_i. \tag{4.4.1}$$

In random-coefficients notation, $\pi_0 \equiv E[\text{D}_{0i}]$ and $\pi_{1i} \equiv (\text{D}_{1i}-\text{D}_{0i})$, so π_{1i} is the heterogeneous causal effect of the instrument on D_i. As with potential outcomes, only one of the potential treatment assignments, D_{1i} and D_{0i}, is ever observed for any one person. In the draft lottery example, D_{0i} tells us whether i would serve in the military if he draws a high (draft-ineligible) lottery number, while D_{1i} tells us whether i would serve if he draws a low (draft-eligible) lottery number. We get to see one or the other of these potential assignments depending on Z_i. The average causal effect of Z_i on D_i is $E[\pi_{1i}]$.

The first assumption in the heterogeneous framework is that the instrument is as good as randomly assigned: it is independent of the vector of potential outcomes and potential treatment assignments. Formally, this can be written

$$[\{\text{Y}_i(d, z); \forall\ d, z\}, \text{D}_{1i}, \text{D}_{0i}] \amalg \text{Z}_i, \tag{4.4.2}$$

Independence is sufficient for a causal interpretation of the *reduced form*, i.e., the regression of Y_i on Z_i.

[22] The distinction between internal and external validity is relatively new to applied econometrics but has a long history in social science. See, for example, the chapter-length discussion in Shadish, Cook, and Campbell (2002), the successor to a classic text on research methods by Campbell and Stanley (1963).

Specifically,

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{Z}_i=1\right]-E\left[\mathrm{Y}_i|\mathrm{Z}_i=0\right] &= E\left[\mathrm{Y}_i(\mathrm{D}_{1i},1)|\mathrm{Z}_i=1\right]-E\left[\mathrm{Y}_i(\mathrm{D}_{0i},0)|\mathrm{Z}_i=0\right] \\ &= E\left[\mathrm{Y}_i(\mathrm{D}_{1i},1)-\mathrm{Y}_i(\mathrm{D}_{0i},0)\right], \end{aligned}$$

the causal effect of the instrument on Y_i. Independence also means that

$$\begin{aligned} E\left[\mathrm{D}_i|\mathrm{Z}_i=1\right]-E\left[\mathrm{D}_i|\mathrm{Z}_i=0\right] &= E\left[\mathrm{D}_{1i}|\mathrm{Z}_i=1\right]-E\left[\mathrm{D}_{0i}|\mathrm{Z}_i=0\right] \\ &= E\left[\mathrm{D}_{1i}-\mathrm{D}_{0i}\right], \end{aligned}$$

in other words, the first-stage from our earlier discussion of 2SLS captures the causal effect of Z_i on D_i.

The second key assumption in the heterogeneous-outcomes framework is the presumption that $\mathrm{Y}_i(d,z)$ is only a function of d.[23] To be specific, while draft-eligibility clearly affects veteran status, an individual's potential earnings *as a veteran* are assumed to be unchanged by draft-eligibility status; while potential earnings *as a nonveteran* are similarly unaffected. In general, the claim that an instrument operates through a single known causal channel is called an *exclusion restriction*. In a linear model with constant effects, the exclusion restriction is expressed by the omission of the instrument from the causal equation of interest, or, equivalently, $E[\mathrm{Z}_i\eta_i]=0$ in equation (4.1.14). It's worth noting that the traditional error-term notation used for simultaneous equations models doesn't lend itself to a clear distinction between independence and exclusion. We need Z_i and η_i to be uncorrelated in this equation, but the reasoning that lies behind this assumption is unclear until we consider both the independence and exclusion restrictions.

The exclusion restriction fails for draft-lottery instruments if men with low draft lottery numbers were affected in some way other than through an increased likelihood of service. For example, Angrist and Krueger (1992) looked for an association between draft lottery numbers and schooling. Their idea was that educational draft deferments would have led men with low lottery numbers to stay in college longer than they would have otherwise desired. If so, draft lottery numbers are correlated with earnings for at least two reasons: an increased likelihood of military service and an increased likelihood of college attendance. The fact that the lottery number is randomly assigned (and therefore satisfies the independence assumption) does not make this possibility less likely. The exclusion restriction is distinct from the claim that the instrument is (as good as) randomly assigned. Rather, it is a claim about a unique channel for causal effects of the instrument.[24]

Using the exclusion restriction, we can define potential outcomes indexed solely against treatment status

[23] Hirano, Imbens, Rubin and Zhou (2000) note that the exclusion restriction that $\mathrm{Y}_i(d,z)$ *equals* $\mathrm{Y}_i(d,z')$ can be weakened to require only that the distributions of $\mathrm{Y}_i(d,z)$ and $\mathrm{Y}_i(d,z')$ be the same.

[24] As it turns out, there is not much of a relationship between schooling and lottery numbers in the Angrist and Krueger (1992) data, probably because educational deferments were phased out during the lottery period.

using the single-index (Y_{1i},Y_{0i}) notation we have been using all along. In particular,

$$\begin{aligned} Y_{1i} &\equiv Y_i(1,1) = Y_i(1,0); \\ Y_{0i} &\equiv Y_i(0,1) = Y_i(0,0). \end{aligned} \tag{4.4.3}$$

The observed outcome, Y_i, can therefore be written in terms of potential outcomes as:

$$\begin{aligned} Y_i &= Y_i(0,Z_i) + [Y_i(1,Z_i) - Y_i(0,Z_i)]D_i \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i. \end{aligned} \tag{4.4.4}$$

A random-coefficients notation for this is

$$Y_i = \alpha_0 + \rho_i D_i + \eta_i,$$

a compact version of (4.4.4) with $\alpha_0 \equiv E[Y_{0i}]$ and $\rho_i \equiv Y_{1i} - Y_{0i}$.

A final assumption needed for heterogeneous IV models is that either $\pi_{1i} \geq 0$ for all i or $\pi_{1i} \leq 0$ for all i. This *monotonicity* assumption, introduced by Imbens and Angrist (1994), means that while the instrument may have no effect on some people, all of those who are affected are affected in the same way. In other words, either $D_{1i} \geq D_{0i}$ or $D_{1i} \leq D_{0i}$ for all i. In what follows, we assume monotonicity holds with $D_{1i} \geq D_{0i}$. In the draft-lottery example, this means that although draft-eligibility may have had no effect on the probability of military service for some men, there is no one who was actually kept out of the military by being draft-eligible. Without monotonicity, instrumental variables estimators are not guaranteed to estimate a weighted average of the underlying individual causal effects, $Y_{1i} - Y_{0i}$.

Given the exclusion restriction, the independence of instruments and potential outcomes, the existence of a first stage, and monotonicity, the Wald estimand can be interpreted as the effect of veteran status on those whose treatment status can be changed by the instrument. This parameter is called the local average treatment effect ((LATE); Imbens and Angrist, 1994). Here is a formal statement:

Theorem 4.4.1 *THE LATE THEOREM. Suppose*

(A1, Independence) $\{Y_i(D_{1i},1), Y_{0i}(D_{0i},0), D_{1i}, D_{0i}\} \amalg Z_i$*;*

(A2, Exclusion) $Y_i(d,0) = Y_i(d,1) \equiv Y_{di}$ *for* $d = 0, 1$*;*

(A3, First-stage), $E[D_{1i} - D_{0i}] \neq 0$

(A4, Monotonicity) $D_{1i} - D_{0i} \geq 0 \forall i$*, or vice versa;*

Then

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}] = E[\rho_i|\pi_{1i} > 0].$$

Proof. Use the exclusion restriction to write $E[Y_i|Z_i = 1] = E[Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 1]$, which equals

$E[\text{Y}_{0i} + (\text{Y}_{1i} - \text{Y}_{0i})\text{D}_{1i}]$ by independence. Likewise $E[\text{Y}_i|\text{Z}_i = 0] = E[\text{Y}_{0i} + (\text{Y}_{1i} - \text{Y}_{0i})\text{D}_{0i}]$, so the numerator of the Wald estimator is $E[(\text{Y}_{1i} - \text{Y}_{0i})(\text{D}_{1i} - \text{D}_{0i})]$. Monotonicity means $\text{D}_{1i} - \text{D}_{0i}$ equals one or zero, so

$$E[(\text{Y}_{1i} - \text{Y}_{0i})(\text{D}_{1i} - \text{D}_{0i})] = E[\text{Y}_{1i} - \text{Y}_{0i}|\text{D}_{1i} > \text{D}_{0i}]P[\text{D}_{1i} > \text{D}_{0i}].$$

A similar argument shows

$$E[\text{D}_i|\text{Z}_i = 1] - E[\text{D}_i|\text{Z}_i = 0] = E[\text{D}_{1i} - \text{D}_{0i}] = P[\text{D}_{1i} > \text{D}_{0i}].$$

■

This theorem says that an instrument which is as good as randomly assigned, affects the outcome through a single known channel, has a first-stage, and affects the causal channel of interest only in one direction, can be used to estimate the average causal effect on the affected group. Thus, IV estimates of effects of military service using the draft lottery estimate the effect of military service on men who served because they were draft-eligible, but would not otherwise have served. This obviously excludes volunteers and men who were exempted from military service for medical reasons, but it includes men for whom draft policy was binding.

How useful is LATE? No theorem answers this question, but it's always worth discussing. Part of the interest in the effects of Vietnam-era service revolves around the question of whether veterans (especially, conscripts) were adequately compensated for their service. Internally valid draft lottery estimates answer this question. Draft lottery estimates of the effects of Vietnam-era conscription may also be relevant for discussions of any future conscription policy. On the other hand, while draft lottery instruments produce internally valid estimates of the causal effect of Vietnam-era conscription, the external validity - i.e., the predictive value of these estimates for military service in other times and places - is not directly addressed by the IV framework. There is nothing in IV formulas to explain *why* Vietnam-era service affects earnings; for that, you need a theory.[25]

You might wonder why we need monotonicity for the LATE theorem, an assumption that plays no role in the traditional simultaneous-equations framework with constant effects. A failure of monotonicity means the instrument pushes some people into treatment while pushing others out. Angrist, Imbens, and Rubin (1996) call the latter group *defiers*. Defiers complicate the link between LATE and the reduced form. To see why, go back to the step in the proof of the LATE theorem which shows the reduced form is

$$E[\text{Y}_i|\text{Z}_i = 1] - E[\text{Y}_i|\text{Z}_i = 0] = E[(\text{Y}_{1i} - \text{Y}_{0i})(\text{D}_{1i} - \text{D}_{0i})].$$

[25] Angrist (1990) interprets draft lottery estimates as the penalty for lost labor market experience. This suggests draft lottery estimates should have external validity for the effects of conscription in other periods, a conjecture born out by the results for WWII draftees in Angrist and Krueger (1994).

Without monotonicity, this is equal to

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} > \mathrm{D}_{0i}]P[\mathrm{D}_{1i} > \mathrm{D}_{0i}] - E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} < \mathrm{D}_{0i}]P[\mathrm{D}_{1i} < \mathrm{D}_{0i}].$$

We might therefore have a scenario where treatment effects are positive for everyone yet the reduced form is zero because effects on compliers are canceled out by effects on defiers. This doesn't come up in a constant-effects model because the reduced form is always the constant effect times the first stage regardless of whether the first stage includes defiant behavior.[26]

A deeper understanding of LATE can be had by linking it to a workhorse of contemporary econometrics, the latent-index model for "dummy endogenous variables" like assignment to treatment. These models describe individual choices as determined by a comparison of partly observed and partly unknown ("latent") utilities and costs (see, e.g., Heckman, 1978). Typically, these unobservables are thought of as being related to outcomes, in which case the treatment variable is said to be endogenous (though it is not really endogenous in a simultanenous-equations sense). For example (ignoring covariates), we might model veteran status as

$$\mathrm{D}_i = \begin{cases} 1 & \text{if } \gamma_0 + \gamma_1 \mathrm{Z}_i > v_i \\ 0 & \text{otherwise} \end{cases},$$

where v_i is a random factor involving unobserved costs and benefits of military service assumed to be independent of Z_i. This latent-index model characterizes potential treatment assignments as:

$$\mathrm{D}_{0i} = 1[\gamma_0 > v_i] \text{ and } \mathrm{D}_{1i} = 1[\gamma_0 + \gamma_1 > v_i].$$

Note that in this model, monotonicity is automatically satisfied since γ_1 is a constant. Assuming $\gamma_1 > 0$, LATE can be written

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} > \mathrm{D}_{0i}] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\gamma_0 + \gamma_1 > v_i > \gamma_0],$$

which is a function of the latent first-stage parameters, γ_0 and γ_1, as well as the joint distribution of $\mathrm{Y}_{1i} - \mathrm{Y}_{0i}$ and v_i. This is not, in general, the same as the population average treatment effect, $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}]$, or the

[26] With a constant effect, ρ,

$$\begin{aligned} E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} \quad &> \quad \mathrm{D}_{0i}]P[\mathrm{D}_{1i} > \mathrm{D}_{0i}] \\ -E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} \quad &< \quad \mathrm{D}_{0i}]P[\mathrm{D}_{1i} < \mathrm{D}_{0i}]. \\ &= \quad \rho\{P[\mathrm{D}_{1i} > \mathrm{D}_{0i}] - P[\mathrm{D}_{1i} < \mathrm{D}_{0i}]\} \\ &= \quad \rho\{E[\mathrm{D}_{1i} - \mathrm{D}_{0i}]\}. \end{aligned}$$

So a zero reduced form effect means either the first stage is zero or $\rho = 0$.

effect on the treated, $E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1]$. We explore the distinction between different average causal effects in Section 4.4.2.

4.4.2 The Compliant Subpopulation

The LATE framework partitions any population with an instrument into a set of three instrument-dependent subgroups, defined by the manner in which members of the population react to the instrument:

Definition 4.4.1 *Compliers. The subpopulation with* $\mathrm{D}_{1i}=1$ *and* $\mathrm{D}_{0i}=0$.

Always-takers. The subpopulation with $\mathrm{D}_{1i}=\mathrm{D}_{0i}=1$.

Never-takers. The subpopulation with $\mathrm{D}_{1i}=\mathrm{D}_{0i}=0$.

LATE is the effect of treatment on the population of compliers. The term "compliers" comes from an analogy with randomized trials where some experimental subjects comply with the randomly assigned treatment protocol (e.g., take their medicine) but some do not, while some control subjects obtain access to the experimental treatment even though they were not supposed to. Those who don't take their medicine when randomly assigned to do so are never-takers while those who take the medicine even when put into the control group are always-takers. Without adding further assumptions (e.g., constant causal effects), LATE is not informative about effects on never-takers and always-takers because, by definition, treatment status for these two groups is unchanged by the instrument (random assignment). The analogy between IV and a randomized trial with partial compliance is more than allegorical - IV solves the problem of causal inference in a randomized trial with partial compliance. This important point merits a separate subsection, below.

Before turning to this important special case, we make a few general points. First, the average causal effect on compliers is not usually the same as the average treatment effect on the treated. From the simple fact that $\mathrm{D}_i=\mathrm{D}_{0i}+(\mathrm{D}_{1i}-\mathrm{D}_{0i})\mathrm{Z}_i$, we learn that the treated population consists of two non-overlapping groups. By monotonicity, we cannot have both $\mathrm{D}_{0i}=1$ and $\mathrm{D}_{1i}-\mathrm{D}_{0i}=1$ since $\mathrm{D}_{0i}=1$ implies $\mathrm{D}_{1i}=1$. The treated therefore have *either* $\mathrm{D}_{0i}=1$ or $\mathrm{D}_{1i}-\mathrm{D}_{0i}=1$ and $\mathrm{Z}_i=1$, and hence D_i can be written as the sum of two mutually-exclusive dummies, D_{i0} and $(\mathrm{D}_{1i}-\mathrm{D}_{0i})\mathrm{Z}_i$. The treated consist of either always-takers or compliers with the instrument switched on. Since the instrument is as good as randomly assigned, compliers with the instrument switched on are representative of all compliers. From here we get

$$
\begin{aligned}
&\underbrace{E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1\right]}_{\text{effect on the treated}} \qquad (4.4.5)\\
&\quad= E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{0i}=1]P[\mathrm{D}_{0i}=1|\mathrm{D}_i=1]\\
&\qquad +E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i},\mathrm{Z}_i=1\right]P\left[\mathrm{D}_{1i}>\mathrm{D}_{0i},\mathrm{Z}_i=1|\mathrm{D}_i=1\right]\\
&\quad= \underbrace{E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{0i}=1]}_{\text{effect on always-takers}}P[\mathrm{D}_{0i}=1|\mathrm{D}_i=1]\\
&\qquad +\underbrace{E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]}_{\text{effect on compliers}}P\left[\mathrm{D}_{1i}>\mathrm{D}_{0i},\mathrm{Z}_i=1|\mathrm{D}_i=1\right]
\end{aligned}
$$

Since $P[\text{D}_{0i}=1|\text{D}_i=1]$ and $P[\text{D}_{1i}>\text{D}_{0i},\text{Z}_i=1|\text{D}_i=1]$ add up to one, this means that the effect of treatment on the treated is a weighted average of effects on always-takers and compliers.

Likewise, LATE is not the average causal effect of treatment on the non-treated, $E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=0]$. In the draft-lottery example, the average effect on the non-treated is the average causal effect of military service on the population of non-veterans from the Vietnam-era cohorts. The average effect of treatment on the non-treated is a weighted average of effects on never-takers and compliers. In particular,

$$\begin{aligned} & \underbrace{E\left[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=0\right]}_{\text{effect on the non-treated}} \\ = \; & \underbrace{E\left[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_{1i}=0\right]}_{\text{effect on never-takers}} P\left[\text{D}_{1i}=0|\text{D}_i=0\right] \\ & + \underbrace{E\left[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_{1i}>\text{D}_{0i}\right]}_{\text{effect on compliers}} P\left[\text{D}_{1i}>\text{D}_{0i},\text{Z}_i=0|\text{D}_i=0\right], \end{aligned} \tag{4.4.6}$$

where we use the fact that, by monotonicity, those with $\text{D}_{1i}=0$ must be never-takers.

Finally, averaging (4.4.5) and (4.4.6) using

$$E[\text{Y}_{1i}-\text{Y}_{0i}]=E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=1]P[\text{D}_i=1]+E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=0]P[\text{D}_i=0]$$

shows the overall population average treatment effect to be a weighted average of effects on compliers, always-takers, and never-takers. Of course, this is a conclusion we could have reached directly given monotonicity and the definition at the beginning of this subsection.

Because an instrumental variable is not directly informative about effects on always-takers and never-takers, instruments do not usually capture the average causal effect on all of the treated or on all of the non-treated. There are important exceptions to this rule, however: instrumental variables that allow no always-takers or no never-takers. Although this scenario is not typical, it is an important special case. One example is the twins instrument for fertility, used by Rosenzweig and Wolpin (1980), Bronars and Grogger (1994), Angrist and Evans (1998), and Angrist, Lavy, and Schlosser (2006). Another is Oreopoulos' (2006) recent study using changes in compulsory attendance laws as instruments for schooling in Britain.

To see how this special case works with twins instruments, let T_i be a dummy variable indicating multiple second births. Angrist and Evans (1998) used this instrument to estimate the causal effect of having three children on earnings in the population of women with at least two children. The third child is especially interesting because reduced fertility for American wives in the 1960s and 1970s meant a switch from three children to two. Multiple second births provide quasi-experimental variation on this margin. Let Y_{0i} denote potential earnings if a woman has only two children while Y_{1i} denotes her potential earnings if she has three, an event indicated by D_i. Assuming that T_i is randomly assigned, i.e., that fertility increases by at most one child in response to a multiple birth, and that multiple births affect outcomes only by increasing fertility,

LATE using the twins instrument, T_i, is also $E[Y_{1i}-Y_{0i}|D_i = 0]$, the average causal effect on women who are not treated (i.e., have two children only). This is because all women who have a multiple second birth end up with three children, i.e., there are no never-takers in response to the twins instrument.

Oreopoulos (2006) also uses IV to estimate an average causal effect of treatment on the non-treated. His study estimates the economic returns to schooling using an increase in the British compulsory attendance age from 14 to 15. Compliance with the Britain's new compulsory attendance law was near perfect, though many teens would previously have dropped out of school at age 14. The causal effect of interest in this case is the earnings premium for an additional year of high-school. Finishing this year can be thought of as the treatment. Since everybody in Oreopoulos' British sample finishes the additional year when compulsory schooling laws are made stricter, Oreopoulos' IV strategy captures the average causal effect of obtaining one more year of high school on all those who leave school at 14. This turns on the fact that British teens are remarkably law-abiding people - Oreopoulos' IV strategy wouldn't estimate the effect of treatment on the non-treated in, say, Israel, where teenagers get more leeway when it comes to compulsory school attendance. Israeli econometricians using changes in compulsory attendance laws as instruments must therefore make do with LATE.

4.4.3 IV in Randomized Trials

The language of the LATE framework is based on an *analogy* between IV and randomized trials. But some instruments really come from randomized trials. If the instrument is a randomly assigned offer of treatment, then LATE is the effect of treatment on those who comply with the offer but are not treated otherwise. An especially important case is when the instrument is generated by a randomized trial with one-sided non-compliance. In many randomized trials, participation is voluntary among those randomly assigned to receive treatment. On the other hand, no one in the control group has access to the experimental intervention. Since the group that receives (i.e., complies with) the assigned treatment is a self-selected subset of those offered treatment, a comparison between those actually treated and the control group is misleading. The selection bias in this case is almost always positive: those who take their medicine in a randomized trial tend to be healthier; those who take advantage of randomly assigned economic interventions like training programs tend to earn more anyway.

IV using the randomly assigned treatment intended as an instrumental variable for treatment received solves this sort of compliance problem. Moreover, LATE is the effect of treatment on the treated in this case. Suppose the instrument, Z_i, is a dummy variable indicating random assignment to a treatment group, while D_i is a dummy indicating whether treatment was actually received. In practice, because of non-compliance, D_i is not equal to Z_i. An example is the randomized evaluation of the JTPA training program, where only 60 percent of those assigned to be trained received training, while roughly 2 percent of those assigned to the control group received training anyway (Bloom, *et al.*, 1997). Non-compliance in the JTPA arose from lack

of interest among participants and the failure of program operators to encourage participation. Since the compliance problem in this case is largely confined to the treatment group, LATE using random assignment, Z_i, as an instrument for treatment received, D_i, is the effect of treatment on the treated.

This use of IV to solve the compliance problems is illustrated in Table 4.4.1, which presents results from the JTPA experiment. The outcome variable of primary interest in the JTPA experiment is total earnings in the 30-month period after random assignment. Columns 1-2 of the table show the difference in earnings between those who were trained and those who were not (the estimates in column 2 are from a regression model that adjusts for a number of individual characteristics measured at the beginning of the experiment. The contrast reported in columns 1-2 is on the order of $4,000 for men and $2,200 for women, in both cases a large treatment effect that amounts to about 20 percent of average earnings. But these estimates are misleading because they compare individuals according to D_i, the actual treatment received. Since individuals assigned to the treatment group were free to decline (and 40% did so), this comparison throws away the random assignment unless the decision to accept treatment is itself independent of potential outcomes. This seems unlikely.

Table 4.4.1: Results from the JTPA experiment: OLS and IV estimates of training impacts

	Comparisons by Training Status		Comparisons by Assignment Status		Instrumental Variable Estimates	
	Without Covariates	With Covariates	Without Covariates	With Covariates	Without Covariates	With Covariates
	(1)	(2)	(3)	(4)	(5)	(6)
A. Men	3,970	3,754	1,117	970	1,825	1,593
	(555)	(536)	(569)	(546)	(928)	(895)
B. Women	2,133	2,215	1,243	1,139	1,942	1,780
	(345)	(334)	(359)	(341)	(560)	(532)

Notes: The table reports OLS, reduced-form, and IV estimates of the effect of subsidized training on earnings in the JTPA experiment. Columns (1) and (2) show differences in earnings by training status; columns (3) and (4) show differences by random-assignment status. Columns (5) and (6) report the result of using random-assignment status as an instrument for training. The covariates used in columns (2), (5) and (6) are *High school or GED, Black, Hispanic, Married, Worked less than 13 weeks in past year, AFDC* (for women), plus indicators for the service strategy recommended, age group and second follow-up survey. Robust standard errors are shown in parenthesis.

Columns 3 and 4 of Table 4.4.1. compare individuals according to whether they were *offered* treatment. In other words, this comparison is based on randomly assigned Z_i. In the language of clinical trials, the contrast in columns 3-4 is known as the *intention-to-treat (ITT) effect*. The intention-to-treat effects in the table are on the order \$1,200 (somewhat less with covariates). Since Z_i was randomly assigned, the ITT effect have a causal interpretation: they tell us the causal effect of the offer of treatment, building in the fact that many of those offered will decline. For this reason, the ITT effect is too small relative to the average causal effect on those who were in fact treated. Columns 5 and 6 put the pieces together and give us the most interesting effect: intention-to-treat divided by the difference in compliance rates between treatment and control groups as originally assigned (about .6). These figures, roughly \$1,800, estimate the effect of treatment on the treated.

How do we know the that ITT-divided-by-compliance is the effect of treatment on the treated? We can recognize ITT as the reduced-form effect of the randomly assigned offer of treatment, our instrument in this case. The compliance rate is the first stage associated with this instrument, and the Wald estimand, as always, is the reduced-form divided by the first-stage. In general this equals LATE, but because we have (almost) no always-takers, the treated population consists (almost) entirely of compliers. The IV estimates in column 5 and 6 of Table 4.4.1 are therefore consistent estimates of the effect of treatment on the treated.

This conclusion is important enough that it warrants an alternative derivation. To the best of our knowledge the first person to point out that the IV formula can be used to estimate the effect of treatment on the treated in a randomized trial with one-sided non-compliance was Howard Bloom (1984). Here is Bloom's result with a simple direct proof.

Theorem 4.4.2 *THE BLOOM RESULT. Suppose the assumptions of the LATE theorem hold, and* $E[D_i|Z_i = 0] = 0$. *Then*

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1]} = E[Y_{1i} - Y_{0i}|D_i = 1].$$

Proof. $E[Y_i|Z_i = 1] = E[Y_{i0} + (Y_{1i} - Y_{0i})D_i|Z_i = 1]$, while $E[Y_i|Z_i = 0] = E[Y_{i0}|Z_i = 0]$ because $E[D_i|Z_i = 0] = 0$. Therefore

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[(Y_{1i} - Y_{0i})D_i|Z_i = 1]$$

by independence. But

$$E[(Y_{1i} - Y_{0i})D_i|Z_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1, Z_i = 1]P[D_i = 1|Z_i = 1]$$

while $E[D_i|Z_i = 0] = 0$ means $D_i = 1$ implies $Z_i = 1$. Hence, $E[Y_{1i} - Y_{0i}|D_i = 1, Z_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1]$

■

In addition to telling us how to analyze randomized trials with non-compliance, the LATE framework

opens the door to cleverly-designed randomized experiments in settings where it's impossible or unethical to compel treatment compliance. A famous example from the field of Criminology is the Minneapolis Domestic Violence Experiment (MDVE). The MDVE was a pioneering effort to determine the best police response to domestic violence (Sherman and Berk, 1984). In general, police use a number of strategies when on a domestic violence call. These include referral to counseling, separation orders, and arrest. A vigorous debate swirls around the question of whether a hard-line response - arrest and at least temporary incarceration - is productive, especially in view of the fact that domestic assault charges are frequently dropped.

As a result of this debate, the city of Minneapolis authorized a randomized trial where the police response to a domestic disturbance was determined in part by random assignment. The research design used randomly shuffled color-coded charge sheets telling the responding officers to arrest some perpetrators while referring others to counseling or separating the parties. In practice, however, the police were free to overrule the random assignment. For example, an especially dangerous or drunk offender was arrested no matter what. As a result, the actual response often deviated from the randomly assigned response, though the two are highly correlated.

Most published analyses of the MDVE data recognize this compliance problem and focus on ITT effects, i.e., an analysis using the original random assignment and not the treatment actually delivered. But the MDVE data can also be used to get the average causal effect on compliers, in this case those who were arrested because they were randomly assigned to be but would not have been arrested otherwise. The MDVE is analyzed in this spirit in Angrist (2006). Because everyone in the MDVE who was assigned to be arrested was in fact arrested, there are no never-takers. This is an interesting twist and the flip-side of the Bloom scenario: here, we have $\text{D}_{1i} = 1$ for everybody. Consequently, LATE is the effect of treatment on the non-treated, i.e.,

$$E[\text{Y}_{1i} - \text{Y}_{0i} | \text{D}_{1i} > \text{D}_{0i}] = E[\text{Y}_{1i} - \text{Y}_{0i} | \text{D}_i = 0],$$

where D_i indicates arrest. The IV estimates using MDVE data show that arrest reduces repeat offenses sharply, in this case, among the subpopulation that was not arrested.[27]

4.4.4 Counting and Characterizing Compliers

We've seen that, except in special cases, each instrumental variable identifies a unique causal parameter, one specific to the subpopulation of compliers for that instrument. Different valid instruments for the same causal relation therefore estimate different things, at least in principle (an important exception being

[27] Another application of IV to data from a randomized trial is Krueger (1999). This study uses randomly assigned class size as an instrument for actual class size with data from the Tennessee STAR experiment. For students in first grade and higher, actual class size differs from randomly assigned class size in the STAR experiment because parents and teachers move students around in years after the experiment began. Krueger 1999 also illustrates 2SLS applied to a model with variable treatment intensity, as discussed in section 4.5.3.

instruments that allow for perfect compliance on one side or the other). Although different IV estimates are "weighted-up" by 2SLS to produce a single average causal effect, over-identification testing of the sort discussed in Section 4.2.2, where multiple instruments are validated according to whether or not they estimate the same thing, is out the window in a fully heterogeneous world.

Differences in compliant sub-populations might explain variability in treatment effects from one instrument to another. We would therefore like to learn as much as we can about the compliers for different instruments. Moreover, if the compliant subpopulation is similar to other populations of interest, the case for extrapolating estimated causal effects to these other populations is stronger. In this spirit, Acemoglu and Angrist (2000) argue that quarter-of-birth instruments and state compulsory attendance laws (the minimum schooling required before leaving school in your state of birth when you were 14) affect essentially the same group of people and for the same reasons. We therefore expect IV estimates of the returns to schooling from these two sets of instruments to be similar. We might also expect the quarter of birth estimates to predict the impact of contemporary proposals to strengthen compulsory attendance laws.

On the other hand, if the compliant subpopulations associated with two or more instruments are very different, yet the IV estimates they generate are similar, we might be prepared to adopt homogeneous effects as a working hypothesis. This revives the over-identification idea, but puts it at the service of external validity.[28] This reasoning is illustrated by the study of the effects of family size on children's education by Angrist, Lavy, and Schlosser (2006). The Angrist, Lavy, and Schlosser study is motivated by the observation that children from larger families typically end up with less education than those from smaller families. A long-standing concern in research on fertility is whether the observed negative correlation between larger families and worse outcomes is causal. As it turns out, IV estimates of the effect of family size using a number of different instruments, each with very different compliant subpopulations, all generate results showing no effect of family size. Angrist, Lavy, and Schlosser (2006) argue that their results point to a common treatment of zero for just about everybody in the Israeli population they study.

We have already seen that the size of a complier group is easy to measure. This is just the Wald first-stage, since, given monotonicity, we have

$$\begin{aligned} P[\mathrm{D}_{1i}>\mathrm{D}_{0i}] &= E[\mathrm{D}_{1i}-\mathrm{D}_{0i}] \\ &= E[\mathrm{D}_{1i}]-E[\mathrm{D}_{0i}] \\ &= E[\mathrm{D}_i|\mathrm{Z}_i=1]-E[\mathrm{D}_i|\mathrm{Z}_i=0]. \end{aligned}$$

We can also tell what proportion of the treated are compliers since, for compliers, treatment status is

[28] In fact, maintaining the hypothesis that all instruments in an over-identified model are valid, the traditional over-identification test statistic becomes a formal test for treatment-effect heterogeneity.

completely determined by Z_i. Start with the definition of conditional probability:

$$\begin{aligned} P[\text{D}_{1i} \quad > \quad \text{D}_{0i} | \text{D}_i{=}1] &= \frac{P[\text{D}_i{=}1 | \text{D}_{1i}{>}\text{D}_{0i}] P[\text{D}_{1i}{>}\text{D}_{0i}]}{P[\text{D}_i{=}1]} \\ &= \quad \frac{P[\text{Z}_i{=}1](E[\text{D}_i|\text{Z}_i{=}1] - E[\text{D}_i|\text{Z}_i{=}0])}{P[\text{D}_i{=}1]}. \end{aligned} \tag{4.4.7}$$

The second equality uses the fact that $P[\text{D}_i{=}1|\text{D}_{1i} > \text{D}_{0i}] = P[\text{Z}_i{=}1|\text{D}_{1i} > \text{D}_{0i}]$ and that $P[\text{Z}_i{=}1|\text{D}_{1i} > \text{D}_{0i}]$ $= P[\text{Z}_i{=}1]$ by Independence. In other words, the proportion of the treated who are compliers is given by the first stage, times the probability the instrument is switched on, divided by the proportion treated.

Formula (4.4.7) is illustrated here by calculating the proportion of veterans who are draft-lottery compliers. The ingredients are reported in Table 4.4.2. For example, for white men born in 1950, the first stage is .159, the probability of draft-eligibility is $\frac{195}{366}$, and the marginal probability of treatment is .267. From these statistics, we compute that the compliant subpopulation is .32 of the veteran population in this group. The proportion of veterans who were draft-lottery compliers falls to 20 percent for non-white men born in 1950. This is not surprising since the draft-lottery first stage is considerably weaker for non-whites. The last column of the table reports the proportion of *nonveterans* who would have served if they had been draft-eligible. This ranges from 3 percent of non-whites to 10 percent of whites, reflecting the fact that most non-veterans were deferred, ineligible, or unqualified for military service.

Table 4.4.2: Probabilities of compliance in instrumental variables studies

Source	Endogenous Variable (D)	Instrument (Z)	Sample	$P[D=1]$	1st Stage, $P[D_1 > D_0]$	$P[Z=1]$	$P[D_1 > D_0 \mid D=1]$	$P[D_1 > D_0 \mid D=0]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Angrist (1990)	Veteran Status	Draft eligibility	White men born in 1950	0.267	0.159	0.534	0.318	0.101
			Non-white men born in 1950	0.163	0.060	0.534	0.197	0.033
Angrist and Evans (1998)	More than 2 children	Twins at second birth	Married women aged 21-35 with two or more children in 1980	0.381	0.603	0.008	0.013	0.966
		First two children are of the same sex	Married women aged 21-35 with two or more children in 1980	0.381	0.060	0.506	0.080	0.048
Angrist and Krueger (1991)	High school graduate	Third or fourth quarter birth	Men born between 1930 and 1939	0.770	0.016	0.509	0.011	0.034
Acemoglu and Angrist (2000)	High school graduate	State requires 11 or more years of school attendance	White men aged 40-49	0.617	0.037	0.300	0.018	0.068

Notes: The table shows an analysis of the absolute and relative size of the complier population for a number of instrumental variables. The first-stage, reported in column 6, gives the absolute size of the complier group. Columns 8 and 9 show the size of the complier population relative to the treated and untreated populations.

The effect of *compulsory* military service is the parameter of primary interest in the Angrist (1990) study, so the fact that draft-eligibility compliers are a minority of veterans is not really a limitation of this study. Even in the Vietnam era, most soldiers were volunteers, a little-appreciated fact about Vietnam-era veterans. The LATE interpretation of IV estimates using the draft lottery highlights the fact that other identification strategies are needed to estimate effects of military service on volunteers (some of these are implemented in Angrist, 1998).

The remaining rows in Table 4.4.2 document the size of the compliant subpopulation for the twins and sibling-sex composition instruments used by Angrist and Evans (1998) to estimate the effects of childbearing and for the quarter of birth instruments and compulsory attendance laws used by Angrist and Krueger (1991) and Acemoglu and Angrist (2000) to estimates the returns to schooling. In each of these studies, the compliant subpopulation is a small fraction of the treated group. For example, less than 2 percent of those who graduated from high school did so because of compulsory attendance laws or by virtue of having been born in a late quarter.

The question of whether a small compliant subpopulation is a cause for worry is context-specific. In some cases, it seems fair to say, "you get what you need." With many policy interventions, for example, it is a marginal group that is of primary interest, a point emphasized in McClellan's (1994) landmark IV study of the effects of surgery on heart attack patients. McClellan uses the relative distance to cardiac care facilities to construct instruments for whether an elderly heart-attack patient is treated with a surgical intervention. Most patients get the same treatment either way, but for some, the case for major surgery is marginal. In such cases, providers or patients opt for a less invasive strategy if the nearest surgical facility is far away. McClellan finds little benefit from surgical procedures for this marginal group. Similarly, an increase in the compulsory attendance age to age 18 is clearly irrelevant for the vast majority of American high school students, but it will affect a few who would otherwise drop out. IV estimates suggest the economic returns to schooling for this marginal group are substantial.

The last column of Table 4.4.2 illustrates the special feature of twins instruments alluded to at the end of the previous subsection. As before, let $\text{D}_i = 0$ for women with two children in a sample of women with at least two children, while $\text{D}_i = 1$ indicates women who have more than two. Because there are no never-takers in response to the event of a multiple birth, i.e., all mothers who have twins at second birth end up with (at least) three children, the probability of compliance among those with $\text{D}_i = 0$ is virtually one (the table shows an entry of .97). LATE is therefore the effect on the non-treated, $E[\text{Y}_{1i} - \text{Y}_{0i} | \text{D}_i = 0]$, in this case.

Unlike the size of the complier group, information on the *characteristics* of compliers seems like a tall order because the compliers cannot be individually identified. Because we can't see both D_{1i} and D_{0i} for each individual, we can't just list those with $\text{D}_{1i} > \text{D}_{0i}$ and then calculate the distribution of characteristics for this group. Nevertheless, it's easy to describe the distribution of complier characteristics. To simplify,

we focus here on characteristics - like race or degree completion - that can be described by dummy variables. In this case, everything we need to know can be learned from variation in the first stage across covariate groups.

Let x_{1i} be a Bernoulli-distributed characteristic, say a dummy indicating college graduates. Are sex-composition compliers more or less likely to be college graduates than other women with two children? This question is answered by the following calculation:

$$\frac{P[x_{1i}=1|\mathrm{D}_{1i}>\mathrm{D}_{0i}]}{P[x_{1i}=1]}=\frac{P[\mathrm{D}_{1i}>\mathrm{D}_{0i}|x_{1i}=1]}{P[\mathrm{D}_{1i}>\mathrm{D}_{0i}]}=\frac{E[\mathrm{D}_i|\mathrm{Z}_i=1,x_{1i}=1]-E[\mathrm{D}_i|\mathrm{Z}_i=0,x_{1i}=1]}{E[\mathrm{D}_i|\mathrm{Z}_i=1]-E[\mathrm{D}_i|\mathrm{Z}_i=0]}. \tag{4.4.8}$$

In other words, the relative likelihood a complier is a college graduate is given by the ratio of the first stage for college graduates to the overall first stage.[29]

This calculation is illustrated in Table 4.4.3**,** which reports compliers' characteristics ratios for age at first birth, nonwhite race, and degree completion using twins and same-sex instruments. The table was constructed from the Angrist and Evans (1998) 1980 census extract. Twins compliers are much more likely to be over 30 than the average mother in the sample, reflecting the fact that younger women who had a multiple birth were likely to go on to have additional children anyway. Twins compliers are also more educated than the average mother, while sex-composition compliers are less educated. This helps to explain the smaller 2SLS estimates generated by twins instruments (reported here in Table 4.1.4), since Angrist and Evans (1998) show that the labor supply consequences of childbearing decline with mother's schooling.

[29] A general method for constructing the mean or other features of the distribution of covariates for compliers uses Abadie's (2003) kappa-weighting scheme. For example,

$$E[\mathrm{X}_i|\mathrm{D}_{1i}>\mathrm{D}_{0i}]=\frac{E[\kappa_i\mathrm{X}_i]}{E[\kappa_i]},$$

where

$$\kappa_i=1-\frac{\mathrm{D}_i(1-\mathrm{Z}_i)}{1-P(\mathrm{Z}_i=1|\mathrm{X}_i)}-\frac{(1-\mathrm{D}_i)\mathrm{Z}_i}{P(\mathrm{Z}_i=1\mathrm{X}_i)}.$$

This works because the weighting function, κ_i, "finds compliers," in a sense discussed in Section (4.5.2), below.

Table 4.4.3: Complier-characteristics ratios for twins and sex-composition instruments

Variable	$E[x]$	Twins at second birth		First two children are same sex	
		$E[x\|D_1 > D_0]$	$P[x\|D_1 > D_0]/P[X]$	$E[x\|D_1 > D_0]$	$P[x\|D_1 > D_0]/P[X]$
	(1)	(2)	(3)	(6)	(5)
Age 30 or older at first birth	0.00291	0.00404	1.39	0.00233	0.995
			(0.0201)		(0.374)
Black or hispanic	0.125	0.103	0.822	0.102	0.814
			(0.00421)		(0.0775)
High school graduate	0.822	0.861	1.048	0.815	0.998
			(0.000772)		(0.0140)
College graduate	0.132	0.151	1.14	0.0904	0.704
			(0.00376)		(0.0692)

Notes: The table reports an analysis of complier characteristics for twins and sex-composition instruments. The ratios in columns 3 and 5 give the relative likelihood compliers have the characteristic indicated in each row. Data are from the 1980 Census 5% sample, including married mothers age 21-35 with at least two children, as in Angrist and Evans (1998). The sample size is 254,654 for all columns.

4.5 Generalizing LATE

The LATE theorem applies to a stripped-down causal model where a single dummy instrument is used to estimate the impact of a dummy treatment with no covariates. We can generalize this in three important ways: multiple instruments (e.g., a set of quarter-of-birth dummies), models with covariates (e.g., controls for year of birth), and models with variable and continuous treatment intensity (e.g., years of schooling). In all three cases, the IV estimand is a weighted average of causal effects for instrument-specific compliers. The econometric tool remains 2SLS and the interpretation remains fundamentally similar to the basic LATE result, with a few bells and whistles. 2SLS with multiple instruments produces a causal effect that averages IV estimands using the instruments one at a time; 2SLS with covariates produces an average of covariate-specific LATEs; 2SLS with variable or continuous treatment intensity produces a weighted average derivative along the length of a possibly nonlinear causal response function.

4.5.1 LATE with Multiple Instruments

The multiple-instruments extension is easy to see. This is essentially the same as a result we discussed in the grouped-data context. Consider a pair of dummy instruments, Z_{1i} and Z_{2i}. Without loss of generality, assume these dummies are mutually exclusive (if not, then we can work with a mutually exclusive set of three dummies, $\text{Z}_{1i}(1-\text{Z}_{2i})$,$\text{Z}_{2i}(1-\text{Z}_{1i})$, and $\text{Z}_{1i}\text{Z}_{2i}$). The two dummies can be used to construct Wald estimators. Again, without loss of generality assume monotonicity is satisfied for each with a positive first stage (if not, we can recode the dummies so this is true). Both therefore estimate a version of $E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_{1i} > \text{D}_{0i}]$, though the population with $\text{D}_{1i} > \text{D}_{0i}$ differs for Z_{1i} and Z_{2i}.

Instead of Wald estimators, we can use Z_{1i} and Z_{2i} together in a 2SLS procedure. Since these two dummies and a constant exhaust the information in the instrument set, this 2SLS procedure is the same as grouped-data estimation using conditional means defined given Z_{1i} and Z_{2i} (whether or not the instruments are correlated). As in Angrist (1991), the resulting grouped-data estimator is a linear combination of the underlying Wald estimators. In other words, it is a linear combination of the instrument-specific LATEs using the instruments one at a time (in fact, it is the efficient linear combination in a traditional homoskedastic linear constant-effects model).

This argument is not quite complete since we haven't shown that the linear combination of LATEs produced by 2SLS is also a weighted average (i.e., the weights are non-negative and sum to one). The relevant weighting formulas appear in Imbens and Angrist (1994) and Angrist and Imbens (1995). The formulas are a little messy, so here we lay out a simple version based on the two-instrument example. The example shows that 2SLS using Z_{1i} and Z_{2i} together is a weighted average of IV estimates using Z_{1i} and Z_{2i} one at a time. Let

$$\rho_j = \frac{Cov(\text{Y}_i, \text{Z}_{ji})}{Cov(\text{D}_i, \text{Z}_{ji})}; j = 1, 2$$

denote the two IV estimands using Z_{1i} and Z_{2i}.

The (population) first stage fitted values for 2SLS are $\hat{\mathrm{D}}_i = \pi_{11}\mathrm{Z}_{1i} + \pi_{12}\mathrm{Z}_{2i}$. By virtue of the IV interpretation of 2SLS, the 2SLS estimand is

$$\begin{aligned}
\rho_{2SLS} &= \frac{Cov(\mathrm{Y}_i, \hat{\mathrm{D}}_i)}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)} = \frac{\pi_{11}Cov(\mathrm{Y}_i, \mathrm{Z}_{1i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)} + \frac{\pi_{12}Cov(\mathrm{Y}_i, \mathrm{Z}_{2i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)} \\
&= \left[\frac{\pi_{11}Cov(\mathrm{D}_i, \mathrm{Z}_{1i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)}\right]\left[\frac{Cov(\mathrm{Y}_i, \mathrm{Z}_{1i})}{Cov(\mathrm{D}_i, \mathrm{Z}_{1i})}\right] + \left[\frac{\pi_{21}Cov(\mathrm{D}_i, \mathrm{Z}_{2i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)}\right]\left[\frac{Cov(\mathrm{Y}_i, \mathrm{Z}_{2i})}{Cov(\mathrm{D}_i, \mathrm{Z}_{2i})}\right] \\
&= \psi\rho_1 + (1-\psi)\rho_2,
\end{aligned}$$

where

$$\psi = \frac{\pi_{11}Cov(\mathrm{D}_i, \mathrm{Z}_{1i})}{\pi_{11}Cov(\mathrm{D}_i, \mathrm{Z}_{1i}) + \pi_{21}Cov(\mathrm{D}_i, \mathrm{Z}_{2i})}$$

is a number between zero and one that depends on the relative strength of each instrument in the first stage. Thus, we have shown that 2SLS is a weighted average of causal effects for instrument-specific compliant subpopulations. Suppose, for example, that Z_{1i} denotes twins births and Z_{2i} indicates same-sex sibships in families with two or more children, both instruments for family size as in Angrist and Evans (1998). A multiple second birth increases the likelihood of having a third child by about .6 while a same-sex sibling pair increases the likelihood of a third birth by about .07. When these two instruments are used together, the resulting 2SLS estimates are a weighted average of the Wald estimates produced by using the instruments one at a time.[30]

4.5.2 Covariates in the Heterogeneous-effects Model

You might be wondering where the covariates have gone. After all, covariates played a starring role in our earlier discussion of regression and matching. Yet the LATE theorem does not involve covariates. This stems from the fact that when we see instrumental variables as a type of (natural or man-made) randomized trial, covariates take a back seat. If, after all, the instrument is randomly assigned, it is likely to be independent of covariates. Not all instruments have this property, however. As with covariates in the regression models in the previous chapter, the main reason why covariates are included in causal analyses using instrumental variables is that the conditional independence and exclusion restrictions underlying IV estimation may be more likely to be valid after conditioning on covariates. Even randomly assigned instruments, like draft-eligibility status, may be valid only after conditioning on covariates. In the case of draft-eligibility, older cohorts were more likely to be draft-eligible because the cutoffs were higher. Because there are year-of-birth (or age) differences in earnings, draft-eligibility status is a valid instrument only after conditioning on year of birth.

[30] Using twins instruments alone, the IV estimate of the effect of a third child on female labor force participation is -.084 (s.e.=.017). The corresponding samesex estimate is -.138 (s.e.=.029). Using both instruments produces a 2SLS estimate of -.098 (.015). The 2SLS weight in this case is .74 for twins, .26 for samesex, due to the much stronger twins first stage.

More formally, IV estimation with covariates may be justified by a *conditional* independence assumption

$$\{\mathrm{Y}_{1i}, \mathrm{Y}_{0i}, \mathrm{D}_{1i}, \mathrm{D}_{0i}\} \amalg \mathrm{Z}_i | \mathrm{X}_i \tag{4.5.1}$$

In other words, we think of the instrumental variables as being "as good as randomly assigned," conditional on covariates, X_i (here we are implicitly maintaining the exclusion restriction as well). A second reason for incorporating covariates is that conditioning on covariates may reduce some of the variability in the dependent variable. This leads to more precise 2SLS estimates under constant conditional effects.

The simplest causal model with covariates is the constant-effects model, with functional form restrictions as follows:

$$\begin{aligned} E[\mathrm{Y}_{0i}|\mathrm{X}_i] &= \mathrm{X}_i'\alpha^* \text{ for a } \mathrm{K} \times 1 \text{ vector of coefficients, } \alpha^*; \\ \mathrm{Y}_{1i} - \mathrm{Y}_{0i} &= \rho. \end{aligned}$$

In combination with (4.5.1), this motivates 2SLS estimation of an equation like (4.1.6) as discussed in Section 4.1.

A straightforward generalization of the constant-effects model allows

$$\mathrm{Y}_{1i} - \mathrm{Y}_{0i} = \rho(\mathrm{X}_i),$$

where $\rho(\mathrm{X}_i)$ is a deterministic function of X_i. This model can be estimated by adding interactions between Z_i and X_i to the first stage and (the same) interactions between D_i and X_i to the second stage. There are now multiple endogenous variables and hence multiple first-stage equations. These can be written

$$\begin{aligned} \mathrm{D}_i &= \mathrm{X}_i'\pi_{00} + \pi_{01}\mathrm{Z}_i + \mathrm{Z}_i\mathrm{X}_i'\pi_{02} + \xi_{0i} \\ \mathrm{D}_i\mathrm{X}_i &= \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i + \mathrm{Z}_i\mathrm{X}_i'\pi_{12} + \xi_{1i} \end{aligned}$$

The second stage equation in this case is

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho_0\mathrm{D}_i + \mathrm{D}_i\mathrm{X}_i'\rho_1 + \eta_i,$$

so $\rho(\mathrm{X}_i) = \rho_0 + \rho_1'\mathrm{X}_i$. Alternately, a nonparametric version of $\rho(\mathrm{X}_i)$ can be estimated by 2SLS in subsamples stratified on X_i.

The heterogeneous-effects model underlying the LATE theorem also allows for identification based on conditional independence as in (4.5.1), though the estimand is a little more complicated. For each value of

X_i, we define covariate- specific LATE,

$$\lambda(X_i) \equiv E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}, X_i].$$

The "saturate and weight" approach to estimation with covariates is spelled out in the following theorem (from Angrist and Imbens, 1995).

Theorem 4.5.1 *SATURATE AND WEIGHT. Suppose the assumptions of the LATE theorem hold conditional on* X_i. *That is,*

(CA1, Independence) $\{Y_i(D_{1i}, 1), Y_{0i}(D_{0i}, 0), D_{1i}, D_{0i}\} \amalg Z_i | X_i$;

(CA2, Exclusion) $P[Y_i(d, 0) = Y_i(d, 1)|X_i] = 1$ *for* $d = 0, 1$;

(CA3, First-stage), $E[D_{1i} - D_{0i}|X_i] \neq 0$

We also assume monotonicity (A4) holds as before. Consider the 2SLS estimand based on the first stage equation

$$D_i = \pi_X + \pi_{1X} Z_i + \xi_{1i} \tag{4.5.3}$$

and the second stage equation

$$Y_i = \alpha_X + \rho_c D_i + \eta_i$$

where π_X *and* α_X *denote saturated models for covariates (a full set of dummies for all values of* X_i*) and* π'_{1X} *denotes a separate first-stage effect of* Z_i *for every value of* X_i. *Then* $\rho_c = E[\omega(X_i)\lambda(X_i)]$ *where*

$$\begin{aligned} \omega(X_i) &= \frac{V\{E[D_i|X_i, Z_i]|X_i\}}{E[V\{E[D_i|X_i, Z_i]|X_i\}]} \\ &= \frac{E\{P[D_i = 1|X_i, Z_i](1 - P[D_i = 1|X_i, Z_i])|X_i\}}{E[E[D_i|X_i, Z_i](1 - P[D_i = 1|X_i, Z_i])]}. \end{aligned} \tag{4.5.4}$$

.

This theorem says that 2SLS with a fully saturated first stage and a saturated model for covariates in the second stage produces a weighted average of covariate-specific LATEs. The weights are proportional to the average conditional variance of the population first-stage fitted value, $E[D_i|X_i, Z_i]$, at each value of X_i.[31] The theorem comes from he fact that the first stage coincides with $E[D_i|X_i, Z_i]$ when (4.5.3) is saturated (i.e., the first-stage regression recovers the CEF).

In practice, we may not want to work with a model with a first-stage parameter for each value of the covariates. First, there is the risk of bias, as we discuss at the end of this chapter, and second, a big pile of

[31] Note that the variability in $E[D_i|X_i, Z_i]$ conditional on X_i comes from Z_i. So the weighting formula gives more weight to covariate values where the instrument creates more variation in fitted values. The first line of the weight formula, (4.5.4), holds for any endogenous variable in a 2SLS setup. The second is a consequence of the fact that here the endogenous variable is a dummy.

individually-imprecise first-stage estimates is not pretty to look at. It seems reasonable to imagine that models with fewer parameters, say a restricted first stage imposing a constant π_{1X}, nevertheless approximates some kind of covariate-averaged LATE. This turns out to be true, but the argument is surprisingly indirect. The vision of 2SLS as providing a MMSE error approximation to an underlying causal relation was developed by Abadie (2003).

The Abadie approach begins by defining the object of interest to be $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}]$, the CEF for Y_i given treatment status and covariates, for compliers. An important feature of this CEF is that when the conditions of the LATE theorem hold conditional on X_i, it has a causal interpretation. In other words, *for compliers*, treatment-control contrasts conditional on X_i are equal to conditional-on-X_i LATEs:

$$\begin{aligned} & E\left[\mathrm{Y}_i|\mathrm{D}_i=1,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]-E\left[\mathrm{Y}_i|\mathrm{D}_i=0,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \;& E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \end{aligned}$$

This follows immediately from the fact that, given (4.5.1), potential outcomes are independent of D_i given X_i and $\mathrm{D}_{1i}>\mathrm{D}_{0i}$.[32] The upshot is that we can imagine running a regression of Y_i on D_i and X_i in the complier population. Although this regression might not give us the CEF of interest (unless it is linear or the model is saturated), it will, as always, provide the MMSE approximation to it. So a regression of Y_i on D_i and X_i in the complier population approximates $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}]$ just like OLS approximates $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]$. Alas, we do not know who the compliers are, so we cannot sample them. Nevertheless, they can be found, in the following sense:

Theorem 4.5.2 *ABADIE KAPPA. Suppose the assumptions of the LATE theorem hold conditional on covariates,* X_i. *Let* $g(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)$ *be any measurable function of* $(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)$ *with finite expectation. Define*

$$\kappa_i = 1-\frac{\mathrm{D}_i(1-\mathrm{Z}_i)}{1-P(\mathrm{Z}_i=1|\mathrm{X}_i)}-\frac{(1-\mathrm{D}_i)\mathrm{Z}_i}{P(\mathrm{Z}_i=1|\mathrm{X}_i)}.$$

Then

$$E[g(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)|\mathrm{D}_{1i}>\mathrm{D}_{0i}]=\frac{E[\kappa_i g(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)]}{E[\kappa_i]}.$$

[32] For compliers,

$$\begin{aligned} & P\left[\mathrm{D}_i=1|\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\},\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \;& P\left[\mathrm{Z}_i=1|\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\},\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]. \end{aligned}$$

And by conditional independence,

$$\begin{aligned} & P\left[\mathrm{Z}_i=1|\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\},\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \;& P\left[\mathrm{Z}_i=1|\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]. \end{aligned}$$

This can be proved by direct calculation using the fact that, given the assumptions of the LATE theorem, any expectation is a weighted average of means for always-takers, never-takers, and compliers. By monotonicity, those with $\mathrm{D}_i(1-\mathrm{Z}_i) = 1$ are always-takers because they have $\mathrm{D}_{0i} = 1$, while those with $(1-\mathrm{D}_i)\mathrm{Z}_i = 1$ are never-takers because they have $\mathrm{D}_{1i} = 0$. Hence, the compliers are the left-out group.

The Abadie theorem has a number of important implications; for example, it crops up again in the discussion of quantile treatment effects. Here, we use it to approximate $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$ by linear regression. Specifically, let α_a and β_a solve

$$(\alpha_a, \beta_a) = \arg\min_{a,b} E\{(E[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i, \mathrm{D}_{1i} > \mathrm{D}_{0i}] - a\mathrm{D}_i - \mathrm{X}_i'b)^2|\mathrm{D}_{1i} > \mathrm{D}_{0i}\}.$$

In other words, $\alpha_a\mathrm{D}_i + \mathrm{X}_i'\beta_a$ gives the MMSE approximation to $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$, or fits it exactly if it's linear. A consequence of Abadie's theorem is that this approximating function can be obtained by solving

$$(\alpha_a, \beta_a) = \arg\min_{a,b} E\{\kappa_i(\mathrm{Y}_i - a\mathrm{D}_i - \mathrm{X}_i'b)^2\}, \tag{4.5.5}$$

the kappa-weighted least-squares minimand.[33]

Abadie proposes an estimation strategy (and develops distribution theory) for a procedure which involves first-step estimation of κ_i using parametric or semiparametric models for the function, $p(\mathrm{X}_i) = P(\mathrm{Z}_i = 1|\mathrm{X}_i)$. The estimates from the first step are then plugged into the sample analog of (4.5.5) in the second step. Not surprisingly, when the only covariate is a constant, Abadie's procedure simplifies to the Wald estimator. More surprisingly, minimization of (4.5.5) produces the traditional 2SLS estimator as long as a linear model is used for $p(\mathrm{X}_i)$ in the construction of κ_i. In other words, if $P(\mathrm{Z}_i = 1|\mathrm{X}_i) = \mathrm{X}_i'\pi$ is used when constructing an estimate of κ_i, the Abadie estimand is 2SLS. Thus, we can conclude that whenever $p(\mathrm{X}_i)$ can be fit or closely approximated by a linear model, it makes sense to view 2SLS as an approximation to the complier causal response function, $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$. On the other hand, α_a is not, in general, the 2SLS estimand and β_a is not, in general, the vector of covariate effects produced by 2SLS. Still, the equivalence to 2SLS for linear $P(\mathrm{Z}_i = 1|\mathrm{X}_i)$ leads us to think that Abadie's method and 2SLS are likely to produce similar estimates in most applications, with the further implication that we can think of 2SLS as approximating $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$.

The Angrist (2001) re-analysis of Angrist and Evans (1998) is an example where estimates based on (4.5.5) are indistinguishable from 2SLS estimates. Using twins instruments to estimate the effect of a third child on female labor supply generates a 2SLS estimate of -.088 (s.e.=.017), while the corresponding Abadie estimate is -.089 (s.e.=.017). Similarly, 2SLS and Abadie estimates of the effect on hours worked

[33] The class of approximating functions needn't be linear. Instead of $a\mathrm{D}_i + \mathrm{X}_i'b$, it might make sense to use a nonlinear function like an exponential (if the dependent variable is non-negative) or probit (if the dependent variable is zero-one). We return to this point at the end of this chapter. As noted in Section (4.4.4), the kappa-weighting sceme can be used to characterize covariate distributions for compliers as well as to estimate outcome distributions.

are identical at -3.55 (s.e.=.617). This is not a strike against Abadie's procedure. Rather, it supports the notion, which we hold dear, that 2SLS approximates the causal relation of interest.[34]

4.5.3 Average Causal Response with Variable Treatment Intensity★

An important difference between the causal effects of a dummy variable and a variable that takes on the values {0, 1, 2, . . .} is that in the first case, there is only one causal effect for any one person, while in the latter there are many: the effect of going from 0 to 1, the effect of going from 1 to 2, and so on. The potential-outcomes notation we used for schooling recognizes this. Here it is again: let

$$Y_{si} \equiv f_i(s),$$

denote the potential (or latent) earnings that person i would receive after obtaining s years of education. Note that the function $f_i(s)$ has an "i" subscript on it while s does not. The function $f_i(s)$ tells us what i would earn for *any* value of schooling, s, and not just for the realized value, S_i. In other words, $f_i(s)$ answers causal "what if" questions for multinomial S_i.

Suppose that S_i takes on values in the set $\{0, 1, ..., \bar{s}\}$. Then there are $\bar{s}$ unit causal effects, $Y_{si} - Y_{s-1,i}$. A linear causal model assumes these are the same for all s and for all i, obviously unrealistic assumptions. But we need not take these assumptions literally. Rather, 2SLS provides a computational device that generates a weighted average of unit causal effects, with a weighting function we can estimate and study, so as to learn where the action is coming from with a particular instrument. This weighting function tells us how the compliers are distributed over the range of S_i. It tells us, for example, that the returns to schooling estimated using quarter of birth or compulsory schooling laws come from shifts in the distribution of high school grades. Other instruments, like the distance instruments used by Card (1995), act elsewhere on the schooling distribution and therefore capture a different sort of return.

To flesh this out, assume that a single binary instrument, Z_i, a dummy for having been born in a state with restrictive compulsory school laws, is to be used to estimate the returns to schooling (as in Acemoglu and Angrist, 2000). Also, let s_{1i} denote the schooling i would get if $z_i = 1$, and let s_{0i} denote the schooling i would get if $z_i = 0$. The theorem below, from Angrist and Imbens (1995), offers an interpretation of the Wald estimand with variable treatment intensity in this case. Note that here we combine the independence and exclusion restrictions by simply stating that potential outcomes indexed by s are independent of the instruments.

Theorem 4.5.3 *AVERAGE CAUSAL RESPONSE. Suppose*

[34] Abadie (2003) gives formulas for standard errors and Alberto Abadie has posted software to compute them. The bootstrap provides a simple alternative, which we used to construct standard errors for the Abadie estimates mentioned in this paragraph.

(ACR1, Independence and Exclusion) $\{Y_{0i}, Y_{1i}, ..., Y_{\bar{s}i}; s_{0i}, s_{1i}\} \amalg \mathrm{Z}_i$*;*

(ACR2, First-stage), $E[s_{1i} - s_{0i}] \neq 0$

(ACR3, Monotonicity) $s_{1i} - s_{0i} \geq 0 \forall i$*, or vice versa; assume the first*

Then

$$\frac{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]}{E[\mathrm{S}_i|\mathrm{Z}_i = 1] - E[\mathrm{S}_i|\mathrm{Z}_i = 0]} = \sum_{s=1}^{\bar{s}} \omega_s E[Y_{si} - Y_{s-1,i}|s_{1i} \geq s > s_{0i}]$$

where

$$\omega_s = \frac{P[s_{1i} \geq s > s_{0i}]}{\sum_{j=1}^{\bar{s}} P[s_{1i} \geq j > s_{0i}]}$$

The weights ω_s *are non-negative and sum to one.*

The average causal response (ACR) theorem says that the Wald estimator with variable treatment intensity is a weighted average of the *unit causal response* along the length of the potentially nonlinear causal relation described by $f_i(s)$. The unit causal response, $E[Y_{si} - Y_{s-1,i}|s_{1i} \geq s > s_{0i}]$, is the average difference in potential outcomes for *compliers at point* s, i.e., individuals driven by the instrument from a treatment intensity less than s to at least s. For example, the quarter of birth instruments used by Angrist and Krueger (1991) push some people from 11^{th} grade to finishing 12^{th} or higher, and others from 10^{th} grade to finishing 11^{th} or higher. The Wald estimator using quarter of birth instruments combines all of these effects into a single average causal response.

The relative size of the group of compliers at point s is $P[s_{1i} \geq s > s_{0i}]$. By monotonicity, this must be non-negative and is given by the difference in the CDF of S_i at point s. To see this, note that

$$\begin{aligned} P\left[s_{1i} \geq s > s_{0i}\right] &= P\left[s_{1i} \geq s\right] - P\left[s_{0i} \geq s\right] \\ &= P\left[s_{0i} < s\right] - P\left[s_{1i} < s\right], \end{aligned}$$

which is non-negative since monotonicity requires $s_{1i} \geq s_{0i}$. Moreover,

$$P[s_{0i} < s] - P[s_{1i} < s] = P[\mathrm{S}_i < s|\mathrm{Z}_i = 0] - P[\mathrm{S}_i < s|\mathrm{Z}_i = 1]$$

by Independence. Finally, note that because the mean of a non-negative random variable is one minus the CDF, we have,

$$\begin{aligned} & E\left[\mathrm{S}_i|\mathrm{Z}_i = 1\right] - E\left[\mathrm{S}_i|\mathrm{Z}_i = 0\right] \\ = & \sum_{j=1}^{\bar{s}} (P\left[\mathrm{S}_i < j|\mathrm{Z}_i = 1\right] - P\left[\mathrm{S}_i < j|\mathrm{Z}_i = 0\right]) = \sum_{j=1}^{\bar{s}} P\left[s_{1i} \geq j > s_{0i}\right] \end{aligned}$$

Thus, the ACR weighting function can be consistently estimated by comparing the CDFs of the endogenous variables (treatment intensity) with the instrument switched off and on. The weighting function is normalized

by the first-stage.

The ACR theorem helps us understand what we are learning from a 2SLS estimate. For example, instrumental variables derived from compulsory attendance and child labor laws capture the causal effect of increases in schooling in the 6-12 grade range, but not from post-secondary schooling. This is illustrated in Figure 4.5.1, taken from Acemoglu and Angrist (2000).

The figure plots differences in the probability that educational attainment is at or exceeds the grade level on the X-axis (i.e., one minus the CDF). The differences are between men exposed to different child labor laws and compulsory schooling laws in the a sample of white men aged 40-49 drawn from the 1960, 1970, and 1980 censuses. The instruments are coded as the number of years of schooling required either to work (Panel A) or leave school (Panel B) in the year the respondent was aged 14. Men exposed to the least restrictive laws are the reference group. Each instrument (e.g., a dummy for 7 years of schooling required before work is allowed) can be used to construct a Wald estimator by making comparisons with the reference group.

Panel A of Figure 4.5.1 shows that men exposed to more restrictive child labor laws were 1-6 percentage points more likely to complete grades 8-12. The intensity of the shift depends on whether the laws required 7, 8, or 9-plus years of schooling before work was allowed. But in all cases, the CDF differences decline at lower grades, and drop off sharply after grade 12. Panel B shows a similar pattern for compulsory attendance laws, though the effects are a little smaller and the action here is at somewhat higher grades, consistent with the fact that compulsory attendance laws are typically binding in higher grades than child labor laws.

Before wrapping up our discussion of LATE generalizations, it's worth noting that most of the elements we have covered work in combination. For example, models with multiple instruments and variable treatment intensity generate a weighted average of the ACR for each instrument. Likewise, the saturate and weight theorem applies to models with variable treatment intensity. On the other hand, we do not yet have an extension of Abadie's Kappa for models with variable treatment intensity. A final important extension is to the scenario where the causal variable of interest is continuous and we can therefore think of the causal response function as having derivatives.

So Long and Thanks for all the Fish

Suppose that as with the schooling problem, we imagine counterfactuals as being generated by an underlying functional relation. In this case, however, the causal variable of interest can take on any non-negative value and the functional relation is assumed to have a derivative. An example where this makes sense is a demand curve, the quantity demanded as a function of price. In particular, let $q_i(p)$ denote the quantity demanded in market i at hypothetical price p. This is a potential outcome, like $f_i(s)$, except that instead of individuals the unit of observation is a time or a location or both. For example, Angrist, Graddy, and Imbens (2000) estimate the elasticity of quantity demanded at the Fulton wholesale fish market in New York City. The

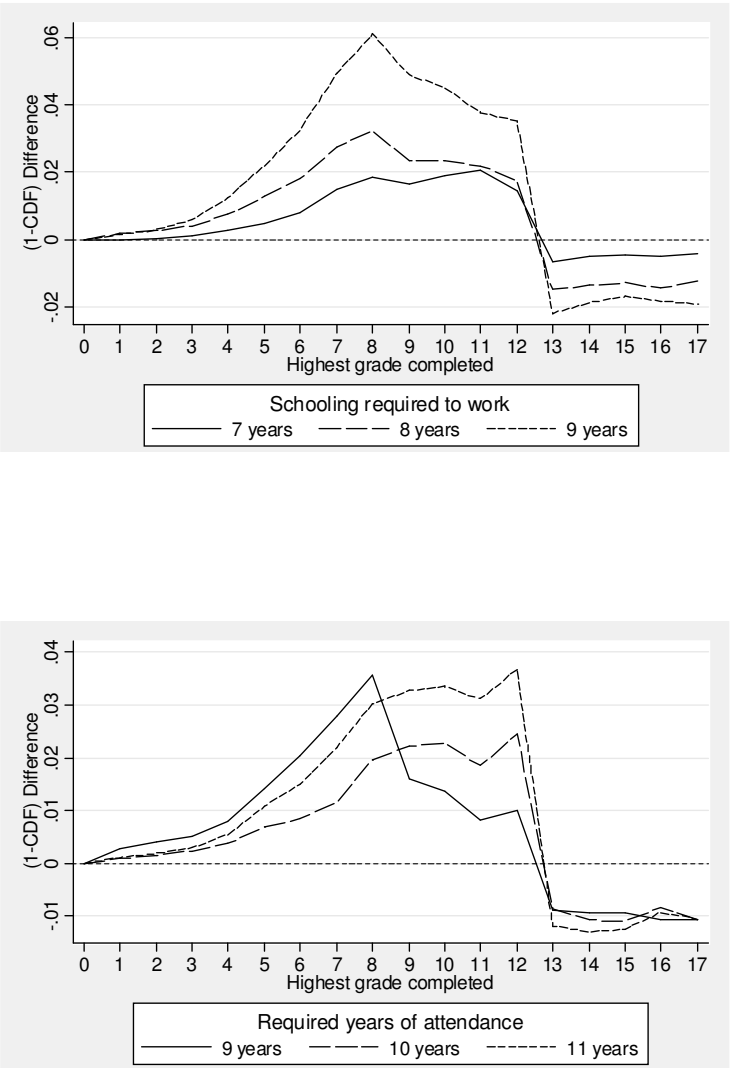

Figure 4.5.1: The effect of compulsory schooling instruments on the probability of schooling (from Acemoglu and Angrist 2000). The figures show the difference in the probability of schooling at or exceeding the grade level on the x-axis. The reference group is 6 or fewer years of required schooling in the top panel, and 8 or fewer years in the bottom panel. The top panel shows the CDF difference by severity of child labor laws. The bottom panel shows the CDF difference by severity of compulsory attendace laws.

slope of this demand curve is $q_i'(p)$; if quantity and price are measured in logs, this is an elasticity.

The instruments in Angrist, Graddy, and Imbens (2000) are derived from data on weather conditions off the coast of Long Island, not too far from major commercial fishing grounds. Stormy weather makes it hard to catch fish, driving up the price, and reducing quantity demanded. Angrist, Graddy, and Imbens use dummy variables such as $stormy_i$, a dummy indicating periods with high wind and waves to estimate the demand for fish. The data consist of daily observations on wholesale purchases of Whiting, a cheap fish used for fish cakes and things like that.

The Wald estimator using the $stormy_i$ instrument can be represented as

$$\frac{E[q_i|stormy_i = 1] - E[q_i|stormy_i = 0]}{E[p_i|stormy_i = 1] - E[p_i|stormy_i = 0]} \tag{4.5.6}$$

$$= \frac{\int E[q_i'(t)|\ p_{1i} \geq t > p_{0i}]P[p_{1i} \geq t > p_{0i}]dt}{\int P[p_{1i} \geq t > p_{0i}]dt}, \tag{4.5.7}$$

where p_i is the price in market (day) i and p_{1i} and p_{0i} are potential prices indexed by $stormy_i$. This is a weighted average derivative with weighting function $P[p_{1i} \geq t > p_{0i}] = P[p_i \leq t|z_i = 0] - P[p_i \leq t|z_i = 1]$ at price t. In other words, IV estimation using $stormy_i$ produces an average of the derivative $q_i'(t)$, with weight given to each possible price (indexed by t) in proportion to the instrument-induced change in the cumulative distribution function (CDF) of prices at that point. This is the same sort of averaging as in the ACR theorem except that now the underlying causal response is a derivative instead of a one-unit difference.

The average causal response formula, (4.5.6), comes from the fact that

$$E[q_i|stormy_i = 1] - E[q_i|stormy_i = 0] = E\int_{p_{0i}}^{p_{1i}} q_i'(t)dt, \tag{4.5.8}$$

by the fundamental theorem of calculus. Two interesting special cases fall neatly out of equation (4.5.8). The first is when the causal response function is linear, i.e., $q_i(p) = \alpha_{0i} + \alpha_{1i}p$, for some random coefficients, α_{0i} and α_{1i}. Then, we have

$$\frac{E[q_i|stormy_i = 1] - E[q_i|stormy_i = 0]}{E[p_i|stormy_i = 1] - E[p_i|stormy_i = 0]} = \frac{E[\alpha_{1i}(p_{1i} - p_{0i})]}{E[p_{1i} - p_{0i}]}, \tag{4.5.9}$$

a weighted average of the random coefficient, α_{1i}. The weights are proportional to the price change induced by the weather in market i.

The second special case is when we can write quantity demanded as

$$q_i(p) = Q(p) + \eta_i, \tag{4.5.10}$$

where $Q(p)$ is a non-stochastic function and η_i is an additive random error. By this we mean $q_i'(p) = Q'(p)$

every day or in every market. In this case, the average causal response function becomes

$$\int Q'(t)\omega(t)dt, \text{where } \omega(t) = \frac{P[p_{1i} \geq t > p_{0i}]}{\int P[p_{1i} \geq r > p_{0i}]dr}.$$

These special cases highlight the two types of averaging wrapped up in the ACR theorem and its continuous corollary, (4.5.6). First, there is averaging *across* markets, with weights proportional to the first-stage impact on prices in each market. Markets where prices are highly sensitive to the weather contribute the most. Second, there is averaging *along* the length of the causal response function in a given market. IV recovers the average derivative over a range of prices where the CDF of prices shifts most sharply.

4.6 IV Details

4.6.1 2SLS Mistakes

2SLS estimates are easy to compute, especially since software like SAS and Stata will do it for you. Occasionally, however, you might be tempted to do it yourself just to see if it really works. Or you may be stranded on the planet Krikkit with all of your software licenses expired (Krikkit is encased in a slo-time envelope, so it will take you a long time to get licenses renewed). "Manual 2SLS" is for just such emergencies. In the Manual 2SLS procedure, you estimate the first stage yourself (which in any case, you should be looking at), and plug the fitted values into the second stage equation, which is then estimated by OLS. Returning to the system at the beginning of this chapter, the first and second stages are

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_i'\pi_{10} + \pi_{11}'\mathrm{Z}_i + \xi_{1i} \\ \mathrm{Y}_i &= \alpha'\mathrm{X}_i + \rho\hat{s}_i + [\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)] \end{aligned}$$

where X_i is a set of covariates, Z_i is a set of excluded instruments, and the first stage fitted values are $\hat{s}_i = \mathrm{X}_i'\hat{\pi}_{10} + \pi_{11}'\mathrm{Z}_i$.

Manual 2SLS takes some of the mystery out of canned 2SLS, and may be useful in a software crisis, but it opens the door to mistakes. For one thing, as we discussed earlier, the OLS standard errors from the manual second stage will not be correct (the OLS residual variance is the variance of $\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)$, while for proper 2SLS standard errors you want the variance of η_i only). There are more subtle risks as well.

Covariate Ambivalence

Suppose the covariate vector contains two sorts of variables, some (say, X_{0i}) that you are comfortable with, and others (say, X_{1i}) about which you are ambivalent. Griliches and Mason (1972) faced this scenario when

constructing 2SLS estimates of a wage equation that treats AFQT scores (an ability test used by the armed forces) as an endogenous control variable to be instrumented. The instruments for AFQT are early schooling (completed before military service), race, and family background variables. They estimated a system that can be described like this:

$$\begin{aligned} \text{S}_i &= \text{X}_{0i}'\pi_{10} + \pi_{11}'\text{Z}_i + \xi_{1i} \\ \text{Y}_i &= \alpha_0'\text{X}_{0i} + \alpha_0'\text{X}_{1i} + \rho\hat{s}_i + [\eta_i + \rho(\text{S}_i - \hat{s}_i)]. \end{aligned}$$

This looks a lot like manual 2SLS.

A closer look, however, reveals an important difference between the equations above and the usual 2SLS procedure: the covariates in the first and second stages are not the same. For example, Griliches and Mason included age in the second stage but not in the first, a fact noted by Cardell and Hopkins (1977) in a comment on their paper. This is a mistake. Griliches' and Mason's second stage estimates are not the same as 2SLS. What's worse, they are inconsistent where 2SLS might have been fine. To see why, note that the first-stage residual, $\text{S}_i - \hat{s}_i$, is uncorrelated with X_{0i} *by construction* since OLS residuals are always uncorrelated with included regressors. But because X_{1i} is not included in the first-stage it is likely to be correlated with the first-stage residuals (e.g., age is probably correlated with the AFQT residual from the Griliches and Mason (1972) first stage). The inconsistency from this correlation spills over to all coefficients in the second stage. The moral of the story: put the same exogenous covariates in your first and second stage. If a covariate is good enough for the second stage, it's good enough for the first.

Forbidden Regressions

Forbidden regressions were forbidden by MIT Professor Jerry Hausman in 1975, and while they occasionally resurface in an under-supervised thesis, they are still technically off-limits. A forbidden regression crops up when researchers apply 2SLS reasoning directly to nonlinear models. A common scenario is a dummy endogenous variable. Suppose, for example, the causal model of interest is

$$\text{Y}_i = \alpha'\text{X}_i + \rho\text{D}_i + \eta_i, \tag{4.6.1}$$

where D_i is a dummy variable for veteran status. The usual 2SLS first stage is

$$\text{D}_i = \pi_{10}'\text{X}_i + \pi_{11}'\text{Z}_i + \xi_{1i}, \tag{4.6.2}$$

a linear regression of D_i on covariates and regressors.

Because D_i is a dummy variable, the CEF associated with this first stage, $E[\text{D}_i|\text{X}_i,\text{Z}_i]$, is probably nonlinear. So the usual OLS first-stage is an approximation to the underlying nonlinear CEF. We might,

therefore, use a nonlinear first stage in an attempt to come closer to the CEF. Suppose that we use Probit to model $E[\mathrm{D}_i|\mathrm{X}_i,\mathrm{Z}_i]$. The Probit first stage is $\Phi[\mathrm{X}_i'\pi_{p0} + \pi_{p1}'\mathrm{Z}_i]$, where π_{p0} and π_{p1} are Probit coefficients, and the fitted values are $\hat{\mathrm{D}}_{pi} = \Phi[\mathrm{X}_i'\hat{\pi}_{p0} + \hat{\pi}_{p1}'\mathrm{Z}_i]$. The forbidden regression in this case is the second stage equation created by substituting $\hat{\mathrm{D}}_{pi}$ for D_i:

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\hat{\mathrm{D}}_{pi} + [\eta_i + \rho(\mathrm{D}_i - \hat{\mathrm{D}}_{pi})]. \tag{4.6.3}$$

The problem with (4.6.3) is that only OLS estimation of (4.6.2) is guaranteed to produce first-stage residuals that are uncorrelated with fitted values and covariates. If $E[\mathrm{D}_i|\mathrm{X}_i,\mathrm{Z}_i] = \Phi[\mathrm{X}_i'\pi_{p0} + \pi_{p1}'\mathrm{Z}_i]$, then residuals from the nonlinear model will be asymptotically uncorrelated with X_i and $\hat{\mathrm{D}}_{pi}$, but who is to say that the first stage CEF is really Probit? With garden-variety 2SLS, in contrast, we do not need to worry about whether the first-stage CEF is really linear.[35]

A simple alternative to the forbidden second step, (4.6.3), avoids problems due to an incorrect nonlinear first stage. Instead of plugging in nonlinear fitted values, we can use the nonlinear fitted values *as instruments*. In other words, use $\hat{\mathrm{D}}_{pi}$ as an instrument for (4.6.1) in a conventional 2SLS procedure (as always, the exogenous covariates, X_i, should also be in the instrument list). Use of fitted values as instruments is the same as plugging in fitted values when the first-stage is estimated by OLS, but not in general. Nonlinear-fits-as-instruments has the further advantage that, if the nonlinear model gives a better approximation of the first-stage CEF than the linear model, the resulting 2SLS estimates will be more efficient than those using a linear first stage (Newey, 1990).

But here, too, there is a drawback. The nonlinear-fits-as-instruments procedure implicitly uses nonlinearities in the first stage as a source of identifying information. To see this, suppose the causal model of interest includes the instruments, Z_i :

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \gamma'\mathrm{Z}_i + \rho\mathrm{D}_i + \eta_i. \tag{4.6.4}$$

Now, with the first stage given by (4.6.2), the model is unidentified and conventional 2SLS estimates of (4.6.4) don't exist. But 2SLS estimates using X_i, Z_i, $\hat{\mathrm{D}}_{pi}$ do exist, because $\hat{\mathrm{D}}_{pi}$ is a nonlinear function of X_i and Z_i that is excluded from the second stage. Should you use this nonlinearity as a source of identifying information? We usually prefer to avoid this sort of back-door identification since its not clear what the underlying experiment really is.

As a rule, naively plugging in first-stage fitted values in nonlinear models is a bad idea. This includes models with a nonlinear second stage as well as those where the CEF for the first stage is nonlinear. Suppose,

[35] The insight that consistency of 2SLS estimates in a traditional SEM does not depend on correct specification of the first-stage CEF goes back to Kelejian (1971). Use of a nonlinear plug-in first-stage may not do too much damage in practice - a probit first-stage can be pretty close to linear - but why take a chance when you don't have to?

for example, that you believe the causal relation between schooling and earnings is approximately quadratic (as in Card's [1995] structural model). In other words, the model of interest is

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho_1 \mathrm{S}_i + \rho_2 \mathrm{S}_i^2 + \eta_i. \tag{4.6.5}$$

Given two instruments, it's easy enough to estimate (4.6.5) treating both S_i and S_i^2 as endogenous. In this case, there are two first-stage equations, one for S_i and one for S_i^2. You need at least two instruments for this to work, of course. It's natural to use Z_i and its square (unless Z_i is a dummy, in which case you'll need a better idea).

You might be tempted, however, to work with a single first stage, say equation (4.6.2), and estimate the following second stage manually:

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho_1 \hat{s}_i + \rho_2 \hat{s}_i^2 + [\eta_i + \rho_1(\mathrm{S}_i - \hat{s}_i) + \rho_2(\mathrm{S}_i^2 - \hat{s}_i^2)].$$

This is a mistake since $\hat{s}_i$ can be correlated with $\mathrm{S}_i^2 - \hat{s}_i^2$ while $\hat{s}_i^2$ can be correlated with both $\mathrm{S}_i - \hat{s}_i$ and $\mathrm{S}_i^2 - \hat{s}_i^2$. On the other hand, as long as X_i and Z_i are uncorrelated with η_i in (4.6.5), and you have enough instruments in Z_i, 2SLS estimation of (4.6.5) is straightforward.

4.6.2 Peer Effects

A vast literature in social science is concerned with peer effects. Loosely speaking, this means the causal effect of group characteristics on individual outcomes. Sometimes regression is used in an attempt to uncover these effects. In practice, the use of regression models to estimate peer effects is fraught with peril. Although this is not really an IV issue *per se*, the language and algebra of 2SLS helps us understand why peer effects are hard to identify.

Broadly speaking, there are two types of peer effects. The first concerns the effect of group characteristics such as the average schooling in a state or city on individually-measured outcome variable. This peer effect links the average of one variable to individual outcomes as described by another variable. For example, Acemoglu and Angrist (2000) ask whether a given individual's earnings are affected by the average schooling in his or her state of residence. The theory of human capital externalities suggests that living in a state with a more educated workforce may make everyone in the state more productive, not just those who are more educated. This kind of spillover is said to be a *social return* to schooling: human capital that benefits everyone, whether or not they are more educated.

A causal model which allows for such externalities can be written

$$Y_{ijt} = \delta_j + \lambda_t + \gamma \overline{S}_{jt} + \rho \mathrm{S}_i + u_{jt} + \eta_{ijt}, \tag{4.6.6}$$

where Y_{ijt} is the log weekly wage of individual i in state j in year t, u_{jt} is a state-year error component, and η_i is an individual error term. The controls δ_j and λ_t are state-of-residence and year effects. The coefficient ρ is the returns to schooling for an individual, while the coefficient γ is meant to capture the effect of average schooling, $\overline{S}_{jt}$, in state j and year t.

In addition to the usual concerns about s_i, the most important identification problem raised by equation (4.6.6) is omitted variables bias from correlation between average schooling and other state-year effects embodied in the error component u_{jt}. For example, public university systems may expand during cyclical upturns, generating a common trend in state average schooling levels and state average earnings. Acemoglu and Angrist (2000) attempt to solve this problem using instrumental variables derived from historical compulsory attendance laws that are correlated with $\overline{S}_{jt}$ but uncorrelated with contemporary u_{jt} and η_i.

While omitted state-year effects are the primary concern motivating Acemoglu and Angrist's (2000) instrumental variables estimation, the fact that one regressor, $\overline{S}_{jt}$, is the average of another regressor, s_i, also complicates the interpretation of OLS estimates of equation (4.6.6). To see this, consider a simpler version of (4.6.6) with a cross-section dimension only. This can be written

$$Y_{ij} = \mu + \pi_0 \mathrm{s}_i + \pi_1 \overline{S}_j + \nu_i; \text{ where } E[\nu_i \mathrm{s}_i] = E[\nu_i \overline{S}_j] \equiv 0. \tag{4.6.7}$$

where Y_{ij} is he log weekly wage of individual i in state j and $\overline{S}_j$ is average schooling in the state. Now, let ρ_0 denote the coefficient from a bivariate regression of Y_{ij} on s_i only and let ρ_1 denote the coefficient from a bivariate regression of Y_{ij} on $\overline{S}_j$ only. From the discussion of grouping and 2SLS earlier in this chapter, it's clear that ρ_1 is the 2SLS estimate of the coefficient on s_i in a bivariate regression of Y_{ij} on s_i using a full set of state dummies as instruments. The Appendix uses this fact to show that the parameters in equation (4.6.7) can be written in terms of ρ_0 and ρ_1 as

$$\begin{aligned} \pi_0 &= \rho_1 + \phi(\rho_0 - \rho_1) \\ \pi_1 &= \phi(\rho_1 - \rho_0) \end{aligned} \tag{4.6.8}$$

where $\phi = \frac{1}{1-R^2} > 1$, and R^2 is the first-stage R-squared.

The upshot of (4.6.8) is that if, *for any reason,* OLS estimates of the bivariate regression of wages on individual schooling differ from 2SLS estimates using state-dummy instruments, the coefficient on average schooling in (4.6.7) will be nonzero. For example, if instrumenting with state dummies corrects for attenuation bias due to measurement error in s_i, we have $\rho_1 > \rho_0$ and the spurious appearance of positive external returns. In contrast, if instrumenting with state dummies eliminates the bias from positive correlation between s_i and unobserved earnings potential, we have $\rho_1 < \rho_0$, and the appearance of negative social returns.[36] In practice, therefore, it is very difficult to substantiate social effects by OLS estimation of an

[36] The coefficient on average schooling in an equation with individual schooling can be interpreted as the Hausman (1978)

equation like 4.6.6, though more sophisticated strategies where both the individual and group averages are treated as endogenous may work.

A second and even more difficult peer effect to uncover is the effect of the group average of a variable on the individual level of *this same variable.* This is not really an IV problem; it takes us back to basic regression issues. To see this point, suppose that $\overline{S}_j$ is the high-school graduation rate in school j, and we would like to know whether students are more likely to graduate from high school when everyone around them is more likely to graduate from high school. To uncover the peer effect in high school graduation rates, we might work with a regression model like:

$$\text{S}_{ij} = \mu + \pi_2 \overline{S}_j + \xi_{ij}, \tag{4.6.9}$$

where S_{ij} is individual i's high school graduation status and $\overline{S}_j$ is the average high school graduation rate in school j, which i attends.

At first blush, equation (4.6.9) seems like a sensible formulation of a well-defined causal question, but in fact it is nonsense. The regression of S_{ij} on $\overline{S}_j$ *always* has a coefficient of 1, a conclusion that can be drawn immediately once you recognize $\overline{S}_j$ as the first-stage fitted value from a regression of S_{ij} on a full set of school dummies.[37] Thus, an equation like (4.6.9) cannot possibly be informative about causal effects.

A modestly improved version of the bad peer regression changes (4.6.9) to

$$\text{S}_{ij} = \mu + \pi_4 \overline{S}_{(i)j} + \xi_{ij}, \tag{4.6.10}$$

where $\overline{S}_{(i)j}$ is the mean of S_{ij} in school j, excluding student i. This is a step in the right direction - by definition, i is not in the group used to construct $\overline{S}_{(i)j}$ - but still problematic because S_{ij} and $\overline{S}_{(i)j}$ are both affected by school-level random shocks. The presence of random effects in the error term raises important issues for statistical inference, issues discussed at length in Chapter 8. But in an equation like (4.6.10), group-level random shocks are more that a problem for standard errors: any shock common to the group (school) creates spurious peer effects. For example, particularly effective school principals may raise graduation rates for everyone in the schools at which they work. This looks like a peer effect since it induces correlation between S_{ij} and $\overline{S}_{(i)j}$ even if there is no causal link between peer means and individual student

test statistic for the equality of OLS estimates and 2SLS estimates of private returns to schooling using state dummies as instruments. Borjas (1992) discusses a similar problem affecting the estimation of ethnic-background effects.

[37] Here is a direct proof that the regression of S_{ij} on $\overline{S}_j$ is always unity:

$$\begin{aligned} \frac{\sum_j \sum_i \text{S}_{ij}(\overline{S}_j - \overline{S})}{\sum_j n_j(\overline{S}_j - \overline{S})^2} &= \frac{\sum_j (\overline{S}_j - \overline{S}) \sum_i \text{S}_{ij}}{\sum_j n_j(\overline{S}_j - \overline{S})^2} \\ &= \frac{\sum_j (\overline{S}_j - \overline{S})(n_j \overline{S}_j)}{\sum_j n_j(\overline{S}_j - \overline{S})^2} = 1. \end{aligned}$$

achievement. We therefore prefer not see regressions like (4.6.10) either.

The best shot at a causal investigation of peer effects focuses on variation in *ex ante* peer characteristics, that is, some measure of peer quality which predates the outcome variable and is therefore unaffected by common shocks. A recent example is Ammermueller and Pischke (2006), who study the link between classmates' family background, as measured by the number of books in their homes, and student achievement in European primary schools. The Ammermueller and Pischke regressions are versions of

$$\text{S}_{ij} = \mu^* + \pi_4 \overline{B}_{(i)j} + \xi_{ij},$$

where $\overline{B}_{(i)j}$ is the average number of books in the home of student i's peers. This looks like (4.6.10), but with an important difference. The variable $\overline{B}_{(i)j}$ is a feature of the home environment that predates test scores and is therefore unaffected by school-level random shocks.

Angrist and Lang (2004) provide another example of an attempt to link student achievement with the *ex ante* characteristics of peers. The Angrist and Lang study looks at the impact of bused-in low-achieving newcomers on high-achieving residents' test scores. The regression of interest in this case is a version of

$$\text{S}_{ij} = \mu + \pi_3 \overline{m}_j + \xi_{ij}, \tag{4.6.11}$$

where $\overline{m}_j$ is the number of bused-in low-achievers in school j and S_{ij} is resident-student i's test score. Spurious correlation due to common shocks is not a concern in this context for two reasons. First, $\overline{m}_j$ is a feature of the school population determined by students outside the sample used to estimate (4.6.11). Second, the number of low-achievers is an *ex ante* variable biased on prior information about where the students come from and not the outcome variable, S_{ij}. School-level random effects remain an important issue for inference, however, since $\overline{m}_j$ is a group-level variable.

4.6.3 Limited Dependent Variables Reprise

In Section 3.4.2, we discussed the consequences of limited dependent variables for regression models. When the dependent variable is binary or non-negative, say, employment status or hours worked, the CEF is typically nonlinear. Most nonlinear LDV models are built around a non-linear transformation of a linear latent index. Examples include Probit, Logit, and Tobit. These models capture features of the associated CEFs (e.g., Probit fitted values are guaranteed to be between zero and one, while Tobit fitted values are non-negative). Yet we saw that the added complexity and extra work required to interpret the results from latent-index models may not be worth the trouble.

An important consideration in favor of OLS is a conceptual robustness that structural models often lack. OLS is always a MMSE linear approximation to the CEF. In fact, we can think of OLS as a scheme for computing marginal effects - a scheme that has the virtue of simplicity, automation, and comparability

across studies. Nonlinear latent-index models are more like GLS - they provide an efficiency gain when taken literally, but they require a commitment to functional form and distributional assumptions about which we do not usually feel strongly.[38] A second consideration is the difference between the latent-index parameters at heart of nonlinear models and the average causal effects that we believe should be the objects of primary interest in most research projects.

The arguments in favor of conventional OLS with LDVs apply with equal force to 2SLS and models with endogenous variables. IV methods capture local average treatment effects regardless of whether the dependent variable is binary, non-negative, or continuously distributed on the real line. With covariates, we can think of 2SLS as estimating LATE averaged across covariate cells. In models with variable or continuous treatment intensity, 2SLS gives us the average causal response or an average derivative. Although Abadie (2003) has shown that 2SLS does not, in general, provide the MMSE approximation to the complier causal response function, in practice, 2SLS estimates come out remarkably close to estimates using the more rigorously grounded Abadie procedure (and with a saturated model for covariates, 2SLS and Abadie are the same). And, of course, 2SLS estimates LATE directly; there is no intermediate step involving the calculation of marginal effects.

2SLS is not the only way to go. An alternative more elaborate approach tries to build up a causal story by describing the process generating LDVs in detail. A good example is bivariate Probit, which can be applied to the Angrist and Evans (1998) example like this. Suppose that a woman decides to have a third child by comparing costs and benefits using a net benefit function or latent index that is linear in covariates and excluded instruments, with a random component or error term, v_i. The bivariate Probit first stage can be written

$$\mathrm{D}_i = 1[\mathrm{X}_i'\gamma_0 + \gamma_1 \mathrm{Z}_i > v_i], \tag{4.6.12}$$

where Z_i is an instrumental variable that increases the benefit of a third child, conditional on covariates, X_i. For example, American parents appear to value a third child more when they have had either two boys or two girls, a sort-of portfolio-diversification phenomenon that can be understood as increasing the benefit

[38] The analogy between nonlinear LDV models and GLS is more than rhetorical. Consider a Probit model with nonlinear CEF $E[\mathrm{Y}_i|\mathrm{X}_i] = \Phi\left[\frac{\mathrm{X}_i'\beta}{\sigma}\right] \equiv p_i$. The first-order conditions for maximum likelihood estimation of this model are

$$\sum \frac{(\mathrm{Y}_i - p_i)\mathrm{X}_i}{p_i(1-p_i)} = 0.$$

Thus, maximum likelihood is the same as GLS estimation of the nonlinear model

$$\mathrm{Y}_i = \Phi\left[\frac{\mathrm{X}_i'\beta}{\sigma}\right] + \xi_i.$$

Consistency of the maximum likelihood estimator turns on the assumption that the conditional variance of Y_i is $p_i(1-p_i)$. It's worth noting that we can dispense with this assumption and simply fit Y_i to $\Phi\left[\frac{\mathrm{X}_i'\beta}{\sigma}\right]$ by nonlinear least squares (NLLS). This sort of agnostic NLLS shares the robustness properties of OLS; it gives the best MMSE fit in a class of approximating functions.

of a third child in families with same-sex sibships.

An outcome of primary interest in this context is employment status, a Bernoulli random variable with a conditional mean between zero and one. To complete the model, suppose that employment status, Y_i, is determined by the latent index

$$\mathrm{Y}_i = 1[\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{D}_i > \varepsilon_i], \tag{4.6.13}$$

where ε_i is a second random component or error term. This latent index can be seen as arising from a comparison of the costs and benefits of working.

The source of omitted variables bias in the bivariate Probit setup is correlation between v_i and ε_i. In other words, unmeasured random determinants of childbearing are correlated with unmeasured random determinants of employment. The model is identified by assuming Z_i is independent of these components, and that the random components are normally distributed. Given normality, the parameters in (4.6.12) and (4.6.13) can be estimated by maximum likelihood. The log likelihood function is

$$\begin{aligned} &\sum \mathrm{Y}_i \ln \Phi_b \left(\frac{\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{D}_i}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0 + \gamma_1 \mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu} \right) \\ &+(1-\mathrm{Y}_i) \ln \left[1 - \Phi_b \left(\frac{\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{D}_i}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0 + \gamma_1 \mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu} \right) \right], \end{aligned} \tag{4.6.14}$$

where $\Phi_b(\cdot, \cdot; \rho_{\varepsilon\nu})$ is the bivariate normal distribution function with correlation coefficient $\rho_{\varepsilon\nu}$. Note, however, that we can multiply the latent index coefficients by a positive constant without changing the likelihood. The object of estimation is therefore the ratio of the index coefficients to the standard deviation of the error terms (e.g., $\beta_1/\sigma_\varepsilon$).

The potential outcomes defined by the bivariate Probit model are

$$\mathrm{Y}_{0i} = 1[\mathrm{X}_i'\beta_0 > \varepsilon_i] \text{ and } \mathrm{Y}_{1i} = 1[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i],$$

while potential treatment assignments are

$$\mathrm{D}_{0i} = 1[\mathrm{X}_i'\gamma_0 > v_i] \text{ and } \mathrm{D}_{1i} = 1[\mathrm{X}_i'\gamma_0 + \gamma_1 > v_i].$$

As usual, only one potential outcome and one potential assignment is observed for any one person. It's also clear from this representation that correlation between v_i and ε_i is the same thing as correlation between potential treatment assignments and potential outcomes.

The latent index coefficients do not themselves tell us anything about the size of the causal effect of childbearing on employment other than the sign. To see this, note that the average causal effect of childbearing is

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] = E\{1[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i] - 1[\mathrm{X}_i'\beta_0 > \varepsilon_i]\}$$

while the average effect on the treated is

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1] = E\{1[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i] - 1[\mathrm{X}_i'\beta_0 > \varepsilon_i]|\mathrm{X}_i'\gamma_0 + \gamma_1\mathrm{Z}_i > v_i\}.$$

Given alterative distributional assumptions for v_i and ε_i, these can be anything (If the error terms are heteroskedastic then even the sign is indeterminate).

Under normality, the average causal effects generated by the bivariate Probit model are easy to evaluate. The average causal effect is

$$\begin{aligned} & E\left\{1\left[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i\right] - 1\left[\mathrm{X}_i'\beta_0 > \varepsilon_i\right]\right\} \\ = \; & E\left\{\Phi\left[\frac{\mathrm{X}_i'\beta_0 + \beta_1}{\sigma}\right] - \Phi\left[\frac{\mathrm{X}_i'\beta_0}{\sigma}\right]\right\}, \end{aligned} \tag{4.6.15}$$

where $\Phi[\cdot]$ is the normal CDF. The effect on the treated is a little more complicated since it involves the bivariate normal CDF

$$\begin{aligned} & E\left[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1\right] \\ = \; & E\left\{\frac{\Phi_b\left(\frac{\mathrm{X}_i'\beta_0+\beta_1}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0+\gamma_1\mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu}\right) - \Phi_b\left(\frac{\mathrm{X}_i'\beta_0}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0+\gamma_1\mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu}\right)}{\Phi\left(\frac{\mathrm{X}_i'\gamma_0+\gamma_1\mathrm{Z}_i}{\sigma_\nu}\right)}.\right\} \end{aligned} \tag{4.6.16}$$

Since the bivariate normal CDF is a canned function in many software packages, this is easy enough to calculate in practice.

Bivariate Probit probably qualifies as harmless in the sense that it's not very complicated, and easy to get right using packaged software routines. Still, it shares the disadvantages of nonlinear latent-index modeling discussed in the previous chapter. First, some researchers become distracted by an effort to identify index coefficients instead of average causal effects. For example, a large literature in econometrics is concerned with the identification of index coefficients without the need for distributional assumptions. Applied researchers interested in causal effects can safely ignore this work.[39]

A second vice in this context is also a virtue. Bivariate Probit and other models of this sort can be used to identify population average causal effects and/or effects on the treated. 2SLS does not promise you average causal effects, only *local* average causal effects. But it should be clear from (4.6.15) that the assumed normality of the latent index error terms is essential for this. As always, the best you can do without a distributional assumption is LATE, the average causal effect for compliers. For bivariate Probit, we can

[39] Suppose the latent error term has an unknown distribution, with CDF $\Lambda[\cdot]$. The average causal effect in this case is

$$E\left\{\Lambda\left[\mathrm{X}_i'\beta_0 + \beta_1\right] - \Lambda\left[\mathrm{X}_i'\beta_0\right]\right\} = \Lambda'[\mathrm{X}_i'\beta_0 + \tilde{\beta}_1]\beta_1,$$

where $\tilde{\beta}_1$ is in $[0, \beta_1]$. This always depends on the shape of $\Lambda[\cdot]$.

write LATE as

$$\begin{aligned} & E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \; & E\{1[\mathrm{X}_i'\beta_0+\beta_1>\varepsilon_i]-1[\mathrm{X}_i'\beta_0>\varepsilon_i]|\mathrm{X}_i'\gamma_0+\gamma_1>v_i>\mathrm{X}_i'\gamma_0\}, \end{aligned}$$

which, like (4.6.16), can be evaluated using joint normality of v_i and ε_i. But you needn't bother using normality to evaluate $E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i}]$, since LATE can be estimated by IV for each X_i and averaged using the histogram of the covariates. Alternately, do 2SLS and settle for a variance-weighted average of covariate-specific LATEs.

You might be wondering whether LATE is enough. Perhaps you would like to estimate the average treatment effect or the effect of treatment on the treated and are willing to make a few extra assumptions to do so. That's all well and good, but in our experience, you can't get blood from a stone, even with heroic assumptions. Since local information is all that's in the data, in practice the average causal effects produced by bivariate Probit are likely to be similar to 2SLS estimates provided the model for covariates is sufficiently flexible. This is illustrated in Table 4.6.1**,** which reports 2SLS and bivariate Probit estimates of the effects of a third child on female labor supply using the Angrist-Evans (1998) same-sex instruments and the same 1980 census sample of married women with 2 or more children used in their paper**.** The dependent variable is a dummy for having worked the previous year; the endogenous variable is a dummy for having a third child. The first stage effect of a same-sex sibship on the probability of a third birth is about 7 percentage points.

Panel A of Table 4.6.1 reports estimates from a model with no covariates. The 2SLS estimate of -.138 in column 1 is numerically identical to the Abadie causal effect estimated using a linear model in column 2, as it should be in this case. Without covariates, the 2SLS slope coefficient provides the best linear approximation to the complier causal response function as does Abadie's kappa-weighting procedure. The marginal effect changes little if, instead of a linear approximation, we use nonlinear least squares with a Probit CEF. The marginal effect estimated by minimizing

$$E\left\{\kappa_i\left(\mathrm{Y}_i-\Phi\left[\frac{\beta_0+\beta_1\mathrm{D}_i}{\sigma_\varepsilon}\right]\right)^2\right\}$$

is -.137, reported in column 3. This is not surprising since the model without covariates imposes no functional form assumptions.

Perhaps more surprising is the fact that marginal effects and the average treatment effects calculated using (4.6.15) and (4.6.16) are also the same as the 2SLS and Abadie estimates. These results are reported in columns 4-6. The marginal effect calculated using a derivative to approximate to the finite difference in (4.6.15) is -.138 (in column 4, labelled MFX for marginal effects), while both average treatment effects are -.139 in columns 5 and 6. Adding a few covariates has little effect on the estimates, as can be seen in Panel

Table 4.6.1: 2SLS, Abadie, and bivariate probit estimates of the effects of a third child on female labor supply

	2SLS	Abadie Estimates		Bivariate probit		
		Linear	Probit	MFX	ATE	TOT
	(1)	(2)	(3)	(4)	(5)	(6)
			A. No Covariates			
Employment	-0.138	-0.138	-0.137	-0.138	-0.139	-0.139
	(0.029)	(0.030)	(0.030)	(0.029)	(0.029)	(0.029)
			B. Some covariates (no age controls)			
Employment	-0.132	-0.132	-0.131	-0.135	-0.135	-0.135
	(0.029)	(0.029)	(0.028)	(0.028)	(0.028)	(0.028)
			C. Some covariates plus age at first birth			
Employment	-0.129	-0.129	-0.129	-0.133	-0.133	-0.133
	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.026)
			D. Some covariates plus age at first birth and a dummy for age>30			
Employment	-0.124	-0.125	-0.125	-0.131	-0.131	-0.131
	(0.028)	(0.029)	(0.029)	(0.025)	(0.025)	(0.025)
			E. Some covariates plus age at first birth and age			
Employment	-0.120	-0.121	-0.121	-0.171	-0.171	-0.171
	(0.028)	(0.026)	(0.026)	(0.023)	(0.023)	(0.023)

Notes: Adapted from Angrist (2001). The table compares 2SLS estimates to alternative IV-type estimates of the effect of childbearing on labor supply using nonlinear models. Standard errors for the Abadie estimates were bootstrapped using 100 replications of subsamples of size 20,000. MFX denotes marginal effects; ATE is the average treatment effect; TOT is the average effect of treatment on the treated.

B. In this case, the covariates are all dummy variables, three for race (black, Hispanic, and other), and two indicating first and second-born boys (the excluded instrument is the interaction of these two). Panels C and D show that adding a linear term in age at first birth and a dummy for maternal age also leaves the estimates unchanged.

The invariance to covariates seems desirable: since the same-sex instrument is essentially independent of the covariates, control for covariates is unnecessary to eliminate bias and should primarily affect precision. Yet, as Panel E shows, the marginal effects generated by bivariate Probit are sensitive to the list of covariates. Swapping a dummy indicating mothers over 30 with a linear age term increases the bivariate Probit estimates markedly, to -.171, while leaving 2SLS and the Abadie estimators unchanged. This probably reflects the fact that the linear age change induces an extrapolation into cells where there is little data. Although there is no harm in reporting the results in Panel E, it's hard to see why the more robust 2SLS and Abadie estimators should not be featured as most likely more reliable.[40]

[40] Angrist (2001) makes the same point using twins instruments, and reports a similar pattern in a comparison of 2SLS,

4.6.4 The Bias of 2SLS★

It is a fortunate fact that the OLS estimator is not only consistent, it is also unbiased. 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.[41] The 2SLS estimator, in contrast, is consistent, but biased. This means that the 2SLS estimator only promises to be close the causal effect of interest in large samples. In small samples, the 2SLS estimator can differ systematically from the population estimand.

For many years, applied researchers have 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.[42]

The 2SLS estimator is most biased when the instruments are "weak," meaning the correlation with endogenous regressors is low, and when there are many over-identifying restrictions. When the instruments are both many and weak, the 2SLS estimator is biased towards the probability limit of the corresponding OLS estimate. In the worst-case scenario for many weak instruments, when the instruments are so weak that there really is no first-stage in the population, the 2SLS sampling distribution is centered on the 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 of the randomness in the first stage is due to the endogenous variable. This randomness turns into finite-sample correlation between first-stage fitted values and the 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 be variable across estimators). 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

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

Abadie, and nonlinear structural estimates of models for hours worked. Angrist (1991) compares 2SLS and bivariate Probit estimates in sampling experiments.

[41] A more precise statement is that OLS is unbiased when, either (a) the CEF is linear or, (b) the regressors are non-stochastic, i.e., fixed in repeated samples. In practice, these qualifications do not seem to matter much. As a rule, the sampling distribution of $\hat{\beta} = \left[\sum_i \mathrm{X}_i\mathrm{X}_i'\right]^{-1}\sum_i \mathrm{X}_i\mathrm{Y}_i$, 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.

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

The $N \times \mathrm{Q}$ 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

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

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

$$\widehat{\beta}_{2SLS} - \beta = (x'P_Z x)^{-1} \left(\pi'Z' + \xi'\right) P_Z \eta = (x'P_Z x)^{-1} \pi'Z'\eta + (x'P_Z x)^{-1} \xi'P_Z\eta \tag{4.6.19}$$

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 $(x'P_Z x)^{-1}$, 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,

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

This approximation is much better than the usual first-order 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.[43] Furthermore, because $E[\pi'Z'\xi] = 0$ and $E[\pi'Z'\eta] = 0$, we have

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

The approximate bias of 2SLS therefore comes from the fact that $E\left(\xi'P_Z\eta\right)$ 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:

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

[43] 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 instruments* go to infinity at the same time as the number of observations goes to infinity, thereby keeping the number of observations per instrument constant.

(see the appendix for a derivation). The term $(1/\sigma_\xi^2)E\left(\pi' Z' Z \pi\right)/\mathrm{Q}$ is the F-statistic for the joint significance of all regressors in the first stage regression.[44] Call this statistic F, so that we can write

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

From this we see that as the first stage F-statistic gets small, the bias of 2SLS approaches $\frac{\sigma_{\eta\xi}}{\sigma_\xi^2}$. The bias of the OLS estimator is $\frac{\sigma_{\eta\xi}}{\sigma_x^2}$, which also equals $\frac{\sigma_{\eta\xi}}{\sigma_\xi^2}$ 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 towards" 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 it should happen in large samples when $\pi \neq 0$.

When the instruments are weak, the F-statistic itself 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-squared. The model sum of squares, $E\left(\pi' Z' Z \pi\right)$, 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 $\widehat{x}_{pop} = Z\pi$ for the first-stage fitted values. These fitted values are uncorrelated with the second stage error. In practice, however, we use $\widehat{x} = P_Z x = Z\pi + P_Z\xi$, which differs from $\widehat{x}_{pop}$ 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 the correlation between errors in the first and second stages seeps in to our 2SLS estimates through the sampling variability in $\widehat{\pi}$. Asymptotically, this correlation is negligible, but real life does not play out in "asymptopia".

The bias formula, (4.6.21), shows that the bias in 2SLS is an increasing function of the number of instruments, so clearly bias is least in the just-identified case when the number of instruments is as low as it can get. It turns out, however, that 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 IV estimates tend to be highly unstable and imprecise.

The LIML estimator is approximately median-unbiased for over-identified 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

[44] Sort of; the actual F-statistic is $(1/\hat{\sigma}_\xi^2)\hat{\pi}' Z' Z \hat{\pi}/\mathrm{Q}$, where hats denote estimates. $(1/\sigma_\xi^2)E\left(\pi' Z' Z \pi\right)/\mathrm{Q}$ 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.

large-sample distribution as 2SLS (under constant effects) while providing finite-sample bias reduction. A number of estimators 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.[45]

We use a small Monte Carlo experiment to illustrate some of the theoretical results from the discussion above. The simulated data are drawn from the following model,

$$\begin{aligned} y_i &= \beta x_i + \eta_i \\ x_i &= \sum_{j=1}^{\mathrm{Q}} \pi_j z_{ij} + \xi_i \end{aligned}$$

with $\beta = 1$, $\pi_1 = 0.1$, $\pi_j = 0 \; \forall j > 1$,

$$\left(\begin{array}{c} \eta_i \\ \xi_i \end{array} \right) \Bigg| Z \sim N \left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} 1 & 0.8 \\ 0.8 & 1 \end{array} \right) \right),$$

where the z_{ij} are independent, normally distributed random variables with mean zero and unit variance. The sample size is 1000.

Figure 4.6.1 shows the Monte Carlo distributions of four estimators: OLS, just identified IV (i.e. 2SLS with Q= 1, labeled IV), 2SLS with two instruments (for Q= 2, labeled 2SLS), and LIML with Q= 2. The OLS estimator is biased and centered around a value of about 1.79. IV is centered around 1, the value of β. 2SLS with one weak and one uninformative instrument is moderately biased towards OLS (the median is 1.07). The distribution function for LIML with Q= 2 is basically indistinguishable from that for just-identified IV, even though the LIML estimator uses a completely uninformative instrument.

Figure 4.6.2 reports simulation results where we set Q= 20. Thus, in addition to the one informative but weak instrument, we added 19 worthless instruments. The figure again shows OLS, 2SLS, and LIML distributions. The bias in 2SLS is now much worse (the median is 1.53, close to the OLS median). The sampling distribution of the 2SLS estimator is also much tighter than in the Q= 2 case. LIML continues to

[45] 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 unbiased. LIML turns out to be just such a "combination estimator". Like the bias of 2SLS, the approximate unbiasedness of LIML can be shown using a Bekker-style group-asymptotic sequence that fixes the ratio of instruments to sample size. Its worth mentioning, however, that LIML is biased in models with a certain type of heteroskedasticity; See Hausman, Newey, and Wouterson (2006) for details.

perform well and is centered around $\beta = 1$, with a bit more dispersion than in the Q= 2 case.

Finally, Figure 4.6.3 reports simulation results from a model that is truly unidentified. In this case, we set $\pi_j = 0$; $j = 1, ..., 20$. Not surprisingly, all the sampling distributions are centered around the same value as OLS. On the other hand, the 2SLS sampling distribution is much tighter than the LIML distribution. We would say advantage-LIML in this case because the widely dispersed LIML sampling distribution correctly reflects the fact that the sample is uninformative about the parameter of interest.

What does this mean in practice? Besides retaining a vague sense of worry about your first stage, we recommend the following:

1. Report the first stage and think about whether it makes sense. Are the magnitude and sign as you would expect, or are the estimates too big or large but wrong-signed? If so, perhaps your hypothesized first-stage mechanism isn't really there, rather, you simply got lucky.

2. Report the F-statistic on the excluded instruments. The bigger this is, the better. Stock, Wright, and Yogo (2002) suggest that F-statistics above about 10 put you in the safe zone though obviously this cannot be a theorem.

3. Pick your best single instrument and report just-identified estimates using this one only. Just-identified IV is median-unbiased and therefore unlikely to be subject to a weak-instruments critique.

4. Check over-identified 2SLS estimates with LIML. LIML is less precise than 2SLS but also less biased. If the results come out similar, be happy. If not, worry, and try to find stronger instruments.

5. Look at the coefficients, t-statistics, and F-statistics for excluded instruments in the reduced-form regression of dependent variables on instruments. Remember that the reduced form is proportional to the causal effect of interest. Most importantly, the reduced-form estimates, since they are OLS, are unbiased. As Angrist and Krueger (2001) note, if you can't see the causal relation of interest in the reduced form, it's probably not there.[46]

We illustrate some of this reasoning in a re-analysis of the Angrist and Krueger (1991) quarter-of-birth study. Bound, Jaeger, and Baker (1995) argued that bias is a major concern when using quarter birth as an instrument for schooling, in spite of the fact that sample size exceeds 300,000. "Small sample" is clearly relative. Earlier in the chapter, we saw that the QOB pattern in schooling is clearly reflected in the reduced form, so there would seem to be little cause for concern. On the other hand, Bound, Jaeger, and Baker (1995) argue that the most relevant models have additional controls not included in these reduced forms. Table 4.6.2 reproduces some of the specifications from Angrist and Krueger (1991) as well as other specifications in the spirit of Bound, Jaeger, and Baker (1995).

[46] A recent paper by Chernozhukov and Hansen (2007) formalizes this maxim.

Table 4.6.2: Alternative IV estimates of the economic returns to schooling

	(1)	(2)	(3)	(4)	(5)	(6)
2SLS	0.105	0.435	0.089	0.076	0.093	0.091
	(0.020)	(0.450)	(0.016)	(0.029)	(0.009)	(0.011)
LIML	0.106	0.539	0.093	0.081	0.106	0.110
	(0.020)	(0.627)	(0.018)	(0.041)	(0.012)	(0.015)
F-statistic (excluded instruments)	32.27	0.42	4.91	1.61	2.58	1.97
	Controls					
Year of birth	✓	✓	✓	✓	✓	✓
State of birth					✓	✓
Age, Age squared		✓		✓		✓
	Excluded Instruments					
Quarter of birth	✓	✓				
Quarter of birth*year of birth			✓	✓	✓	✓
Quarter of birth*state of birth					✓	✓
Number of excluded instruments	3	2	30	28	180	178

Notes: The table compares 2SLS and LIML estimates using alternative sets of instruments and controls. The OLS estimate corresponding to the models reported in columns 1-4 is .071; the OLS estimate corresponding to the models reported in columns 5-6 is .067. Data are from the Angrist and Krueger (1991) 1980 Census sample. The sample size is 329,509. Standard errors are reported in parentheses.

The first column in the table reports 2SLS and LIML estimates of a model using three quarter of birth dummies as instruments with year of birth dummies as covariates. The OLS estimate for this specification is 0.071, while the 2SLS estimate is a bit higher at 0.105. The first-stage F-statistic is over 32, well above the danger zone. Not surprisingly, the LIML estimate is almost identical to 2SLS in this case.

Angrist and Krueger (1991) experimented with models that include age and age squared measured in quarters as additional controls. These controls are meant to pick up omitted age effects that might confound the quarter-of-birth instruments. The addition of age and age squared reduces the number of instruments to two, since age in quarters, year of birth, and quarter of birth are linearly dependent. As shown in column 2, the first stage F-statistic drops to 0.4 when age and age squared are included as controls, a sure sign of trouble. But the 2SLS standard error is high enough that we would not draw any substantive conclusions from this estimate. The LIML estimate is even less precise. This model is effectively unidentified.

Columns 3 and 4 report the results of adding interactions between quarter of birth dummies and year of birth dummies to the instrument list, so that there are 30 instruments, or 28 when the age and age squared variables are included. The first stage F-statistics are 4.9 and 1.6 in these two specifications. The 2SLS estimates are a bit lower than in column 1 and hence closer to OLS. But LIML is not too far away from 2SLS. Although the LIML standard error is pretty big in column 4, it is not so large that the estimate is uninformative. On balance, there seems to be little cause for worry about weak instruments, even with the age quadratic included.

The most worrisome specifications are those reported in columns 5 and 6. These estimates were produced by adding 150 interactions between quarter of birth and state of birth to the 30 interactions between quarter of birth and year of birth. The rationale for the inclusion of state-of-birth interactions in the instrument list is to exploit differences in compulsory schooling laws across states. But this leads to highly over-identified models with 180 (or 178) instruments, many of which are weak. The first stage F-statistics for these models are 2.6 and 2.0, well into the discomfort zone. On the plus side, the LIML estimates again look fairly similar to 2SLS. Moreover, the LIML standard errors differ little from the 2SLS standard errors in this case. This suggests that you can't always determine instrument relevance using a mechanical rule such as "F>10". In some cases, a low F may not be fatal.[47]

Finally, it's worth noting that in applications with multiple endogenous variables, the conventional first-stage F is no longer appropriate. To see why, suppose there are two instruments for two endogenous variables and that the first instrument is strong and predicts both endogenous variables well while the second instrument is weak. The first-stage F-statistics in each of the two first stage equations are likely to be high but the model is weakly identified because one instrument is not enough to capture two causal effects. A simple modification of the first-stage F for this case is given in the appendix.

[47] Cruz and Moreira (2005) similarly conclude that, low F-statistics notwithstanding, there is little bias in the Angrist and Krueger (1991) 180-instrument specifications.

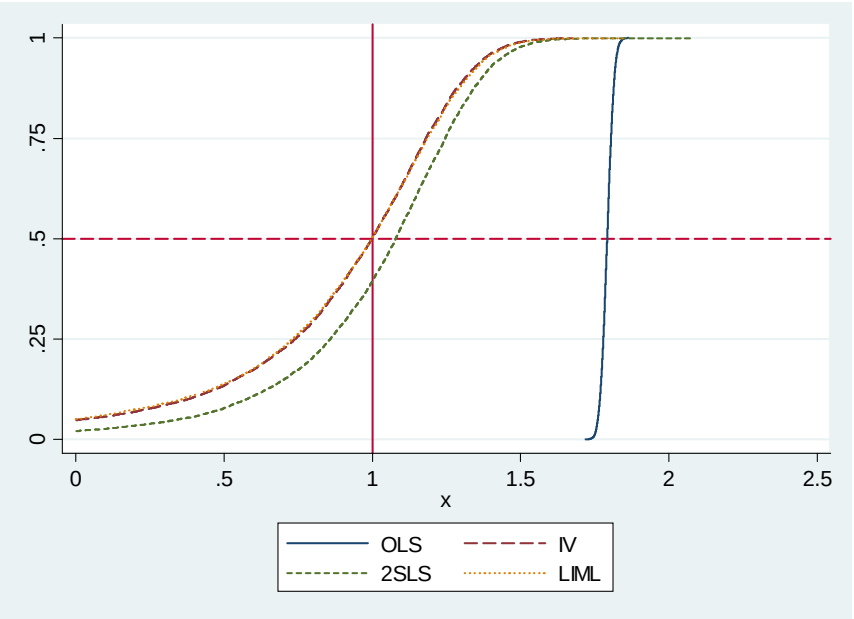

Figure 4.6.1: Distribution of the OLS, IV, 2SLS, and LIML estimators. IV uses one instrument, while 2SLS and LIML use two instruments.

4.7 Appendix

Derivation of Equation (4.6.8)

Rewrite equation (4.6.7) as follows

$$Y_{ij} = \mu^* + \pi_0 \tau_i + (\pi_0 + \pi_1)\overline{S}_j + \nu_i;$$

where $\tau_i \equiv \mathrm{s}_i - \overline{S}_j$. Since τ_i and $\overline{S}_j$ are uncorrelated by construction, we have:

$$\begin{aligned} \rho_1 &= \pi_0 + \pi_1. \\ \pi_0 &= \frac{C(\tau_i, Y_{ij})}{V(\tau_i)}. \end{aligned}$$

Simplifying the second line,

$$\begin{aligned} \pi_0 &= \frac{C[(s_i - \overline{S}_j), Y_{ij}]}{[V(\mathrm{s}_i) - V(\overline{S}_j)]} \\ &= \left[\frac{C(\mathrm{s}_i, Y_{ij})}{V(\mathrm{s}_i)}\right]\left[\frac{V(\mathrm{s}_i)}{V(\mathrm{s}_i) - V(\overline{S}_j)}\right] - \left[\frac{C(\overline{S}_j, Y_{ij})}{V(\overline{S}_j)}\right]\left[\frac{V(\overline{S}_j)}{V(\mathrm{s}_i) - V(\overline{S}_j)}\right] \\ &= \rho_0\phi + \rho_1(1-\phi) = \rho_1 + \phi(\rho_0 - \rho_1) \end{aligned}$$

where $\phi \equiv \frac{V(s_i)}{V(\mathrm{s}_i) - V(\overline{S}_j)}$. Solving for π_1, we have

$$\pi_1 = \rho_1 - \pi_0 = \phi(\rho_1 - \rho_0).$$

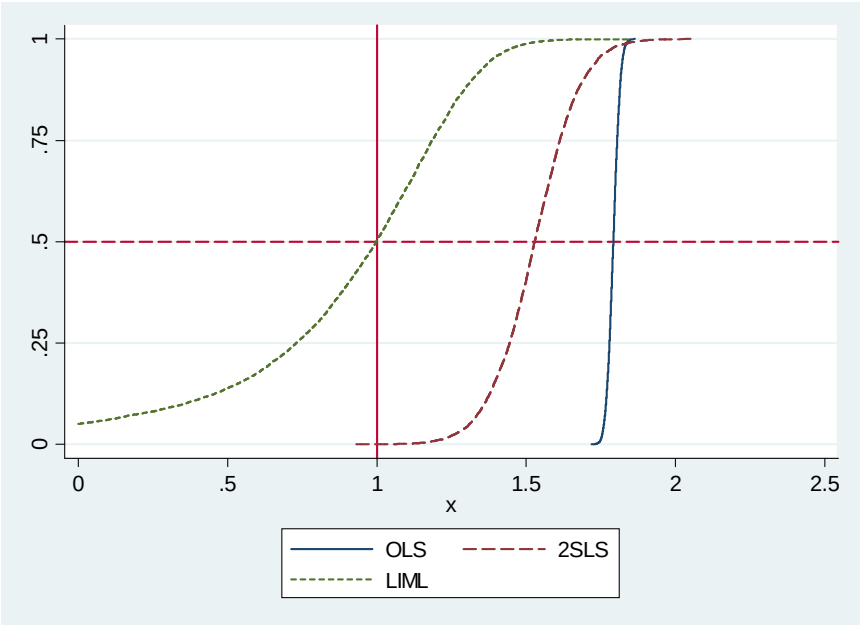

Figure 4.6.2: Distribution of the OLS, 2SLS, and LIML estimators with 20 instruments

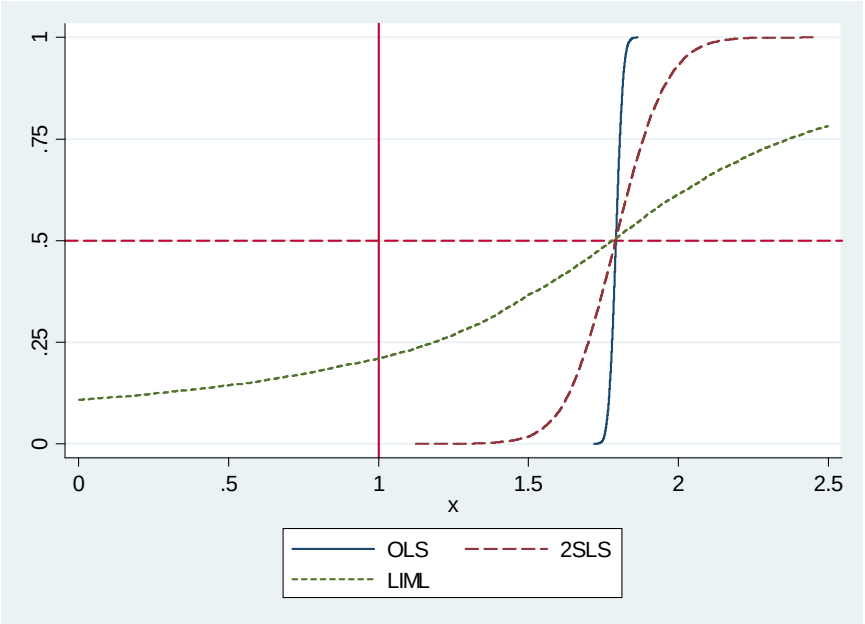

Figure 4.6.3: Distribution of the OLS, 2SLS, and LIML estimators with 20 worthless instruments

Derivation of the approximate bias of 2SLS

Start from the last equality in (4.6.20):

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

The magic of linear algebra helps us simplify this expression: The term $\xi' P_Z \xi$ is a scalar and therefore equal to its trace; the trace is a linear operator which passes through expectations and is invariant to cyclic permutations; finally, the trace of P_Z, an idempotent matrix, is equal to it's rank, Q. Using these facts, we have

$$\begin{aligned} E\left(\xi' P_Z \xi\right) &= E\left[\mathrm{tr}\left(\xi' P_Z \xi\right)\right] \\ &= E\left[\mathrm{tr}\left(P_Z \xi \xi'\right)\right] \\ &= \mathrm{tr}\left(P_Z E\left[\xi \xi'\right]\right) \\ &= \mathrm{tr}\left(P_Z \sigma_\xi^2 I\right) \\ &= \sigma_\xi^2 \mathrm{tr}\left(P_Z\right) \\ &= \sigma_\xi^2 \mathrm{Q}, \end{aligned}$$

where we have assumed that ξ_i is homoskedastic. Similarly, applying the trace trick to $\xi' P_Z \eta$ shows that this term is equal to $\sigma_{\eta\xi}\mathrm{Q}$. Therefore,

$$\begin{aligned} E[\widehat{\beta}_{2SLS} - \beta] &\approx \left[E\left(\pi' Z' Z \pi\right) + \sigma_\xi^2 \mathrm{Q}\right]^{-1} E\left[\mathrm{tr}\left(\xi' P_Z \eta\right)\right] \\ &= \sigma_{\eta\xi}\mathrm{Q}\left[E\left(\pi' Z' Z \pi\right) + \sigma_\xi^2 \mathrm{Q}\right]^{-1} \\ &= \frac{\sigma_{\eta\xi}}{\sigma_\xi^2}\left[\frac{E\left(\pi' Z' Z \pi\right)/\mathrm{Q}}{\sigma_\xi^2} + 1\right]^{-1}. \end{aligned}$$

Multivariate first-stage F-statistics

Assume any exogenous covariates have been partialled out of the instrument list and that there are two endogenous variables, x_1 and x_2 with coefficients δ_1 and δ_2. We are interested in the bias of the 2SLS estimator of δ_2 when x_1 is also treated as endogenous. The second stage equation is

$$y = P_Z x_1 \delta_1 + P_Z x_2 \delta_2 + [\eta + (x_1 - P_Z x_1)\delta_1 + (x_2 - P_Z x_2)\delta_2]. \tag{4.7.1}$$

where $P_Z x_1$ and $P_Z x_2$ are the first-stage fitted values from regressions of x_1 and x_2 on Z. By the usual anatomy formula for multivariate regression, δ_2 in (4.7.1) is the bivariate regression of y on the residual from

a regression of $P_Z x_2$ on $P_Z x_1$. This residual is

$$[I - P_Z x_1 (x_1' P_Z x_1)^{-1} x_1' P_Z] P_Z x_2 = M_{1z} P_Z x_2,$$

where $M_{1z} = [I - P_Z x_1 (x_1' P_Z x_1)^{-1} x_1' P_Z]$ is the relevant residual-maker matrix. In addition, note that $M_{1z} P_Z x_2 = P_Z [M_{1z} x_2]$.

From here we conclude that the 2SLS estimator of δ_2 is the OLS regression on $P_Z[M_{1z} x_2]$, in other words, OLS on the fitted values from a regression of $M_{1z} x_2$ on Z. This is the same as 2SLS using P_Z to instrument $M_{1z} x_2$. So the 2SLS estimator of δ_2 can be written

$$[x_2' M_{1z} P_Z M_{1z} x_2]^{-1} x_2' M_{1z} P_Z y = \delta_2 + [x_2' M_{1z} P_Z M_{1z} x_2]^{-1} x_2' M_{1z} P_Z \eta.$$

The explained sum of squares (numerator of the F-statistic) that determines the bias of the 2SLS estimator of δ_2 is therefore the expectation of $[x_2' M_{1z} P_Z M_{1z} x_2]$, while the bias comes from the fact that the expectation $E[\xi' M_{1z} P_Z \eta]$ is non-zero when η and ξ are correlated.

Here's how to compute this F-statistic in practice: (a) Regress the first stage fitted values for the regressor of interest, $P_Z x_2$, on the other first-stage fitted values and any exogenous covariates. Save the residuals from this step; (b) Construct the F-statistic for excluded instruments in a first-stage regression of the residuals from (a) on the excluded instruments. Note that you should get the 2SLS coefficient of interest in a 2SLS procedure where the residuals from (a) are instrumented using Z, with no other covariates or endogenous variables. Use this fact to check your calculation.

Chapter 5

Parallel Worlds: Fixed Effects, Differences-in-differences, and Panel Data

> The first thing to realize about parallel universes . . . is that they are not parallel.
>
> Douglas Adams, *Mostly Harmless* (1995)

The key to causal inference in chapter 3 is control for observed confounding factors. If important confounders are unobserved, we might try to get at causal effects using IV as discussed in Chapter 4. Good instruments are hard to find, however, so we'd like to have other tools to deal with unobserved confounders. This chapter considers a variation on the control theme: strategies that use data with a time or cohort dimension to control for unobserved-but-fixed omitted variables. These strategies punt on comparisons in levels, while requiring the counterfactual *trend* behavior of treatment and control groups to be the same. We also discuss the idea of controlling for lagged dependent variables, another strategy that exploits timing.

5.1 Individual Fixed Effects

One of the oldest questions in Labor Economics is the connection between union membership and wages. Do workers whose wages are set by collective bargaining earn more because of this, or would they earn more anyway? (Perhaps because they are more experienced or skilled). To set this question up, let Y_{it} equal the (log) earnings of worker i at time t and let D_{it} denote his union status. The observed Y_{it} is either Y_{0it} or Y_{1it}, depending on union status. Suppose further that

$$E(\text{Y}_{0it}|A_i, \text{X}_{it}, t, \text{D}_{it}) = E(\text{Y}_{0it}|A_i, \text{X}_{it}, t),$$