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*

Two things distinguish the discipline of econometrics from the 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 nonexperimental 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 from most other social scientists—is an arsenal of statistical tools that grew out of early

[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.

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, the IV method does more than allow us to consistently estimate the parameters in a system of simultaneous equations, though it allows us 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. On 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 toward 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 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

[2]Key historical references here are Wald (1940) and Durbin (1954), both discussed later in this chapter.

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 methods still comes from this framework. Today, we are more likely to find IV methods used to address measurement error problems than to estimate the parameters of an SEM. Undoubtedly, however, the most important contemporary use of IV methods is to solve the problem of omitted variables bias (OVB). IV methods solve the problem of missing or unknown control variables, much as a randomized trial obviates 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, or 2SLS). An initial focus on constant effects allows us to explain the mechanics of IV with a 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

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

and that

$$f_i(s) = \alpha + \rho s + \eta_i, \tag{4.1.1}$$

as in the discussion of regression and causality in section 3.2. Also, as in the earlier discussion, we imagine that there is a

[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.

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

$$E[s_i v_i] = 0.$$

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

$$y_i = \alpha + \rho s_i + A_i'\gamma + v_i. \tag{4.1.2}$$

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 $Cov(\eta_i, z_i) = 0$, 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.

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

$$\rho = \frac{Cov(y_i, z_i)}{Cov(s_i, z_i)} = \frac{Cov(y_i, z_i)/V(z_i)}{Cov(s_i, z_i)/V(z_i)}. \tag{4.1.3}$$

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 (called the reduced form) to the population regression of S_i on Z_i (called 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, like the CIA in chapter 3), and 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 a combination of institutional knowledge and ideas about the processes determining the variable of interest. For example, the economic model of education suggests that schooling decisions are based on the costs and benefits of alternative choices. Thus, one possible source of instruments for schooling is differences in costs due 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 is 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 OVB.

The starting point for the Angrist and Krueger (1991) quarter-of-birth strategy is the observation that most states require 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 31 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 a different degree, when they reach the legal dropout age. 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 U.S. census data. Panel A of figure 4.1.1 (adapted from Angrist and Krueger, 1991) 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 tended to have lower average schooling levels. Panel A of figure 4.1.1 is a graphical depiction of the first stage. The first stage in a general IV framework is the regression of the causal variable of interest on covariates and instruments. 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 dummies (covariates) and quarter-of-birth dummies (instruments).

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 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 instruments. 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, a covariate in the Angrist and Krueger (1991) setup. Importantly, this reduced-form relation parallels the

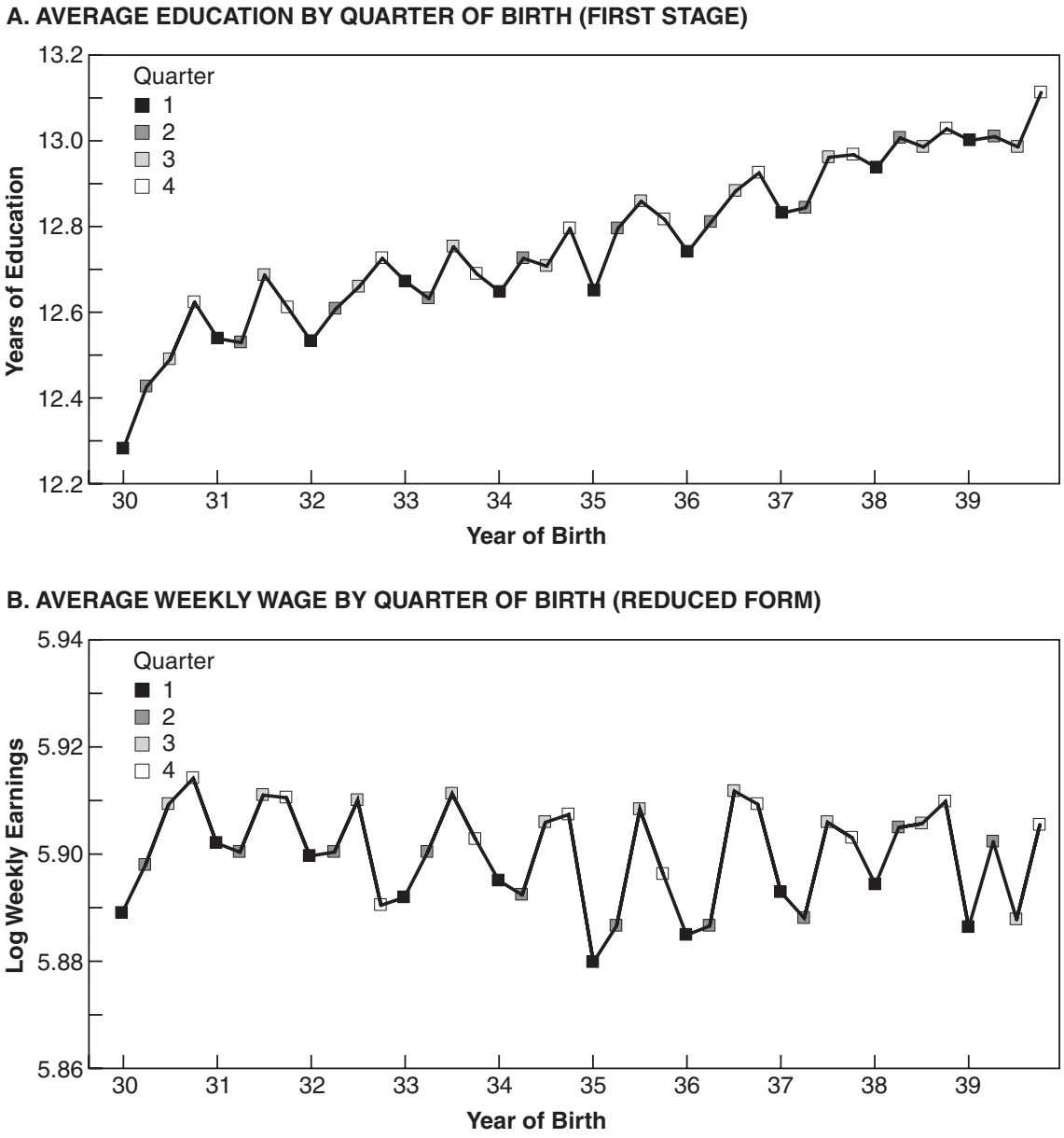

Figure 4.1.1 Graphical depiction of the first stage and reduced form for IV estimates of the economic return to schooling using quarter-of-birth instruments (from Angrist and Krueger, 1991).

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 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:

$$s_i = X_i'\pi_{10} + \pi_{11}z_i + \xi_{1i} \tag{4.1.4a}$$

$$y_i = X_i'\pi_{20} + \pi_{21}z_i + \xi_{2i}. \tag{4.1.4b}$$

The parameter π_{11} in equation (4.1.4a) captures the first-stage effect of z_i on s_i, adjusting for covariates, 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 Angrist and Krueger (1991), the instrument z_i is quarter of birth (or a dummy indicating quarter of birth) and the covariates are dummies for year of birth and state of birth. In the language of the SEM, the dependent variables in these two equations are said to be the *endogenous variables* (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 coefficients are the same. Hence, their ratio is

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{Cov(y_i, \tilde{z}_i)}{Cov(s_i, \tilde{z}_i)}, \tag{4.1.5}$$

[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 most pronounced at the schooling levels most affected by compulsory attendance laws.

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 IV formula, (4.1.3). Econometricians call the sample analog of equation (4.1.5) an indirect least squares (ILS) estimator of ρ in the causal model with covariates,

$$Y_i = \alpha' X_i + \rho S_i + \eta_i, \tag{4.1.6}$$

where η_i is the compound error term, $A_i'\gamma + v_i$. It's easy to use equation (4.1.6) to confirm directly that $Cov(Y_i, \tilde{z}_i) = \rho Cov(S_i, \tilde{z}_i)$, since $\tilde{z}_i$ is uncorrelated with X_i by construction and with η_i by assumption.

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 have:

$$\begin{aligned} Y_i &= \alpha' X_i + \rho[X_i'\pi_{10} + \pi_{11} Z_i + \xi_{1i}] + \eta_i \\ &= X_i'[\alpha + \rho\pi_{10}] + \rho\pi_{11} Z_i + [\rho\xi_{1i} + \eta_i] \\ &= X_i'\pi_{20} + \pi_{21} Z_i + \xi_{2i}, \end{aligned} \tag{4.1.7}$$

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 rearrangement of (4.1.7) gives

$$Y_i = \alpha' X_i + \rho[X_i'\pi_{10} + \pi_{11} Z_i] + \xi_{2i}, \tag{4.1.8}$$

where $[X_i'\pi_{10} + \pi_{11} Z_i]$ 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 $[X_i'\pi_{10} + \pi_{11} Z_i]$ in the population regression of Y_i on X_i and $[X_i'\pi_{10} + \pi_{11} Z_i]$ equals ρ.

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

$$\hat{s}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11} Z_i,$$

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,"

$$Y_i = \alpha' X_i + \rho \hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)], \tag{4.1.9}$$

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 the covariates and first-stage fitted values are uncorrelated with both η_i and $(s_i - \hat{s}_i)$.

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). 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—that is, generated by the instrument z_i.

2SLS is a many-splendored thing. For one, it is an IV estimator: the 2SLS estimate of ρ in (4.1.9) is the sample analog of $\frac{Cov(Y_i,\hat{s}_i^*)}{Cov(s_i,\hat{s}_i^*)}$, 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 $Cov(s_i, \hat{s}_i^*) = V(\hat{s}_i^*)$. 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.[5]

[5]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 the sample analog of $\left[\frac{Cov(Y_i,\tilde{z}_i)}{V(\tilde{z}_i)}\right](\hat{\pi}_{11})^{-1}$. But the sample analog of the numerator, $\frac{Cov(Y_i,\tilde{z}_i)}{V(\tilde{z}_i)}$, 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, that is, the ratio of the reduced-form effect of the instrument to the corresponding first-stage effect where both the first-stage and reduced-form equations include covariates.

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,"

$$Y_i = \alpha' X_i + \rho \hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)], \tag{4.1.9}$$

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 the covariates and first-stage fitted values are uncorrelated with both η_i and $(s_i - \hat{s}_i)$.

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). 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—that is, generated by the instrument z_i.

2SLS is a many-splendored thing. For one, it is an IV estimator: the 2SLS estimate of ρ in (4.1.9) is the sample analog of $\frac{Cov(Y_i,\hat{s}_i^*)}{Cov(s_i,\hat{s}_i^*)}$, 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 $Cov(s_i, \hat{s}_i^*) = V(\hat{s}_i^*)$. 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.[5]

[5]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 the sample analog of $\left[\frac{Cov(Y_i,\tilde{z}_i)}{V(\tilde{z}_i)}\right](\hat{\pi}_{11})^{-1}$. But the sample analog of the numerator, $\frac{Cov(Y_i,\tilde{z}_i)}{V(\tilde{z}_i)}$, 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, that is, the ratio of the reduced-form effect of the instrument to the corresponding first-stage effect where both the first-stage and reduced-form equations include covariates.

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 accomplishes this 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

$$s_i = X_i'\pi_{10} + \pi_{11}z_{1i} + \pi_{12}z_{2i} + \pi_{13}z_{3i} + \xi_{1i}, \qquad (4.1.10a)$$

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 exogenous covariates. The exclusion restriction in this case is the claim that the quarter-of-birth dummies in (4.1.10a) are uncorrelated with η_i in equation (4.1.6).

The results of 2SLS estimation of the economic returns to schooling using quarter-of-birth dummies as instruments 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.

The first pair of IV estimates, reported in columns 3 and 4, are from models without exogenous covariates. The instrument used to construct the estimate in column 3 is a single dummy for first-quarter births, while the instruments used to construct the estimate in column 4 are three dummies indicating first-, second-, and third-quarter births. These estimates range from .10 to .11. The results from models including year-of-birth and state-of-birth dummies as exogenous 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	.071	.067	.102	.13	.104	.108	.087	.057
	(.0004)	(.0004)	(.024)	(.020)	(.026)	(.020)	(.016)	(.029)
Exogenous 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 = 2				✓		✓	✓	✓
dummy for QOB = 3				✓		✓	✓	✓
QOB dummies interacted with year-of-birth dummies (30 instruments total)							✓	✓

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the Angrist and Krueger (1991) 1980 census sample. This sample includes native-born men, born 1930–39, with positive earnings and nonallocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QOB denotes quarter of birth.

(reported in columns 5 and 6) 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 association between schooling and earnings is not driven by omitted variables such as ability and family background.

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

$$\begin{aligned} s_i = X_i'\pi_{10} + \pi_{11}z_{1i} + \pi_{12}z_{2i} + \pi_{13}z_{3i} \qquad (4.1.10b) \\ + \sum_j (b_{ij}z_{1i})\kappa_{1j} + \sum_j (b_{ij}z_{2i})\kappa_{2j} + \sum_j (b_{ij}z_{3i})\kappa_{3j} + \xi_{1i}, \end{aligned}$$

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 quarter and year interaction terms. 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 .019 to .016 as we move from column 6 to column 7.[6] (The first-stage and reduced-form effects plotted in figure 4.1.1 are from this fully interacted specification.)

The last 2SLS model reported in table 4.1.1 adds controls for linear and quadratic terms in age in quarters to the list of exogenous covariates. 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

[6]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 section 4.6.4.

finely coded age variable provides a partial control for the fact that small differences in age may be an omitted variable that confounds the quarter-of-birth identification strategy. As long as the effects of age are reasonably smooth, the quadratic age-in-quarters model will pick them up.

Columns 7 and 8 in table 4.1.1 illustrate the interplay between identification and estimation. (In traditional SEM theory, a parameter is said to be *identified* if we can figure it out from the reduced form.) For the 2SLS procedure to work, there must be some variation in the first-stage fitted values conditional on whatever exogenous 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 would be manifest by perfect multicollinearity (i.e., linear dependence between X_i and $\hat{s}_i$). 2SLS estimates with quadratic age controls exist, but the variability "left over" in the first-stage fitted values is reduced when the covariates include variables such as 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, in other words, 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 IV, 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 dummy 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 one 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 estimator* for a bivariate regression with mismeasured

regressors.[7] In our context, the Wald formula provides an appealingly transparent implementation of the IV strategy for the elimination of OVB. 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 dummy 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 fourth quarters of the year is −.0135, while the corresponding difference in schooling is −.151. 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 .089. 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 both are 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 American men were at risk of being drafted for military service. Concerns about the fairness of the U.S. conscription policy led to the institution of a draft lottery in 1970 that was used to determine priority for conscription. A promising instrument for Vietnam veteran status is therefore draft eligibility, since this was determined by a lottery over birthdays. Specifically,

[7] As noted in the introduction to this chapter, measurement error in regressors tends to shrink regression coefficients toward 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 1st Quarter of Year	(2) Born in 4th Quarter of Year	(3) Difference (Std. Error) (1) − (2)
ln (weekly wage)	5.892	5.905	−.0135 (.0034)
Years of education	12.688	12.839	−.151 (.016)
Wald estimate of return to education			.089 (.021)
OLS estimate of return to education			.070 (.0005)

Notes: From 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 162,515.

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 a cutoff were eligible for the draft, while men with numbers above the cutoff 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, but draft eligibility provides a dummy instrument highly correlated with Vietnam-era veteran status.

Among white men who were at risk of being drafted in the 1970 draft lottery, draft eligibility is clearly associated with lower earnings in the years after the lottery. This is documented in table 4.1.3, which reports the effect of randomized draft eligibility status on Social Security–taxable earnings in column 2. Column 1 shows average annual earnings for purposes of comparison. For men born in 1950, there are

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

	Earnings		Veteran Status		
Earnings Year	Mean (1)	Eligibility Effect (2)	Mean (3)	Eligibility Effect (4)	Wald Estimate of Veteran Effect (5)
1981	16,461	−435.8	.267	.159	−2,741
		(210.5)		(.040)	(1,324)
1971	3,338	−325.9			−2,050
		(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 Income and Program Participation. There are about 13,500 individuals in the sample.

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 casual 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 4 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 interpret. Let D_i denote Vietnam-era veteran status and Z_i indicate 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 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 effects 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 Angrist's (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 that assigned RSNs to the 1953 birth cohort in February 1972, no one born in 1953 was actually drafted (the draft officially ended in July 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. 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 probably would 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 second 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 family background.

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 of the 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 estimated 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

TABLE 4.1.4
Wald estimates of the effects of family size on labor supply

			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	.528	−.167 (.002)	.625 (.011)	−.083 (.017)	.067 (.002)	−.135 (.029)
Weeks worked	19.0	−8.05 (.09)		−3.83 (.76)		−6.23 (1.29)
Hours/week	16.7	−6.02 (.08)		−3.39 (.64)		−5.54 (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. The first stage is the same for all dependent variables.

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 (showing an employment reduction of .135, for example). 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.

4.1.3 Grouped Data and 2SLS

The Wald estimator is the mother of all IV 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 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 to 365) can usually be grouped without much loss of information (e.g., a single dummy for draft eligibility status, or dummies for groups of 25 lottery numbers).[8]

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

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

discussion suggests the possibility of many more comparisons, for example men with $R_i \leq 25$ versus men with $R_i \in [26-50]$, men with $R_i \in [51-75]$ versus men with $R_i \in [76-100]$, and so on, until these 25-number intervals are exhausted. We might also make the intervals finer, comparing, say, men in five-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).

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

$$Y_i = \alpha + \rho D_i + \eta_i, \tag{4.1.14}$$

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

$$E[Y_i|R_i] = \alpha + \rho P[D_i = 1|R_i], \tag{4.1.15}$$

since $P[D_i = 1|R_i] = E[D_i|R_i]$. 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 we estimate e by 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 $j = 1, \ldots, \text{J}$. 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 interval 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, generalized least squares (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 WLS, 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_η^2, this variance is $\frac{\sigma_\eta^2}{n_j}$, where n_j is the group size. Therefore, we should weight by the group size, as discussed in a different context in section 3.4.1.

The GLS (or WLS) estimator of ρ in equation (4.1.16) is especially important for two reasons. First, the GLS slope estimate constructed from J grouped observations is an asymptotically efficient linear combination of any full set of $\text{J} - 1$ linearly independent Wald estimators (Angrist, 1991). This can be seen without any mathematics: GLS and any linear combination of 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 estimator of equation (4.1.16) is 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 $Z_i \equiv \{r_{ji} = 1[\text{R}_i = j]; j = 1, \ldots J-1\}$, where $1[\cdot]$ denotes the indicator function used to construct dummy variables. 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 the same as the slope from WLS 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 using 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 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) recode continuous weather-based instruments into three dummy variables, *stormy, mixed*, and *clear*, which they then use to estimate the demand for fish. This dummy variable parameterization

seems to capture the main features of the relationship between weather conditions and the price of fish.[9]

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 military 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.

Sometimes labor economists refer to grouped data plots for discrete instruments as visual instrumental variables (VIV).[10] An example appears in Angrist (1990), reproduced here as figure 4.1.2. This figure shows the relationship between average earnings in five-number RSN cells and the probability of service in these cells, for the 1981–84 earnings of white men born in 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).

4.2 Asymptotic 2SLS Inference

4.2.1 The Asymptotic 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 $V_i \equiv [\mathrm{X}_i' \; \hat{s}_i]'$ denote the vector of regressors in the 2SLS second stage, equation (4.1.9). The 2SLS

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

[10] See, for example, the preface to Borjas (2005).