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 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

$$\text{Y}_{si} \equiv f_i\left(s\right),$$

and that

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

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:

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

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 $\eta\ _i$ and S_i are correlated, so that

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

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

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

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 $Cov(\eta_i,\text{Z}_i) = 0$, 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]

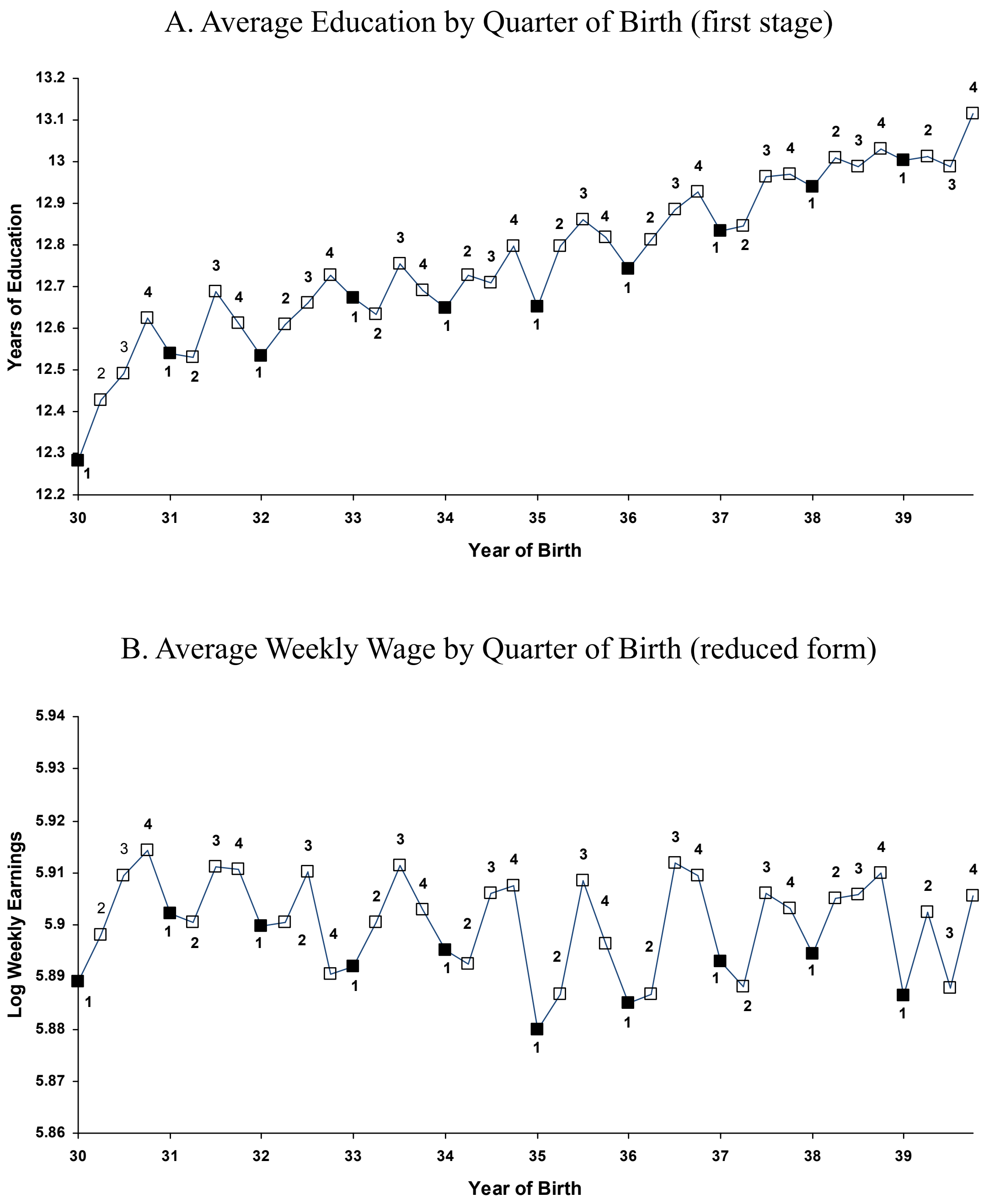
A mathematical representation of the story told by Figure 4.1.1 comes from the first-stage and reduced-form regression equations, spelled out below:

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

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

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.



A. Average Education by Quarter of Birth (first stage)

B. Average Weekly Wage by Quarter of Birth (reduced form)

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.

[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 $\frac{Cov(\mathrm{Y}_i,\tilde{z}_i)}{Cov(\mathrm{S}_i,\tilde{z}_i)}$.