Regression Endogeneity

So far in this class we've considered a "selection-on-observables" approach to identifying causal effects—that the treatment (or other regressors of interest) are conditionally uncorrelated with potential outcomes (or potential outcome trends, in a panel data setting) given some observed controls. This is a flexible framework for using regression to overcoming "selection bias" or other threats to identification, as we've seen. But in many cases this sort of identification strategy may fail us: individuals may select into treatment on the basis of some unobservables (e.g. private information) that we may never have hope to measure and control for in our regression. More generally, the economic model of interest may suffer from "omitted variables bias" by involving terms that cannot be included in a regression.[1]

Broadly, the failure of regression-based identification is sometimes called *endogeneity*: a word economists appear to have made up for this situation.[2] In some cases this problem can be solved by *instrumental variables (IVs)*, a statistical technique which economists appear to have also made up but that is now widely used across many disciplines.[3] The basic idea of IV is as follows: when the causal or otherwise "structural" relationship between some Y_i and some X_i is "endogenous," we can use an "exogenous" Z_i that affects Y_i only through X_i to estimate the structural relationship. This definition, while compact, is very unclear without some more notation however...

Let's (as usual) consider the returns-to-schooling example: Y_i denotes an individual's adult earnings, X_i measures her completed schooling, and ε_i captures her (unobserved) ability or family characteristics. We posit a simple relationship of

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \tag{1}$$

with β being our usual returns-to-schooling parameter of interest. As written, equation (1) may look like a regression but of course we now know better: the model unobservable ε_i need not be uncorrelated with schooling X_i. When $Cov(X_i, \varepsilon_i) \neq 0$ we cannot recover β by the regression of Y_i on X_i; here we might say X_i is "endogenous," perhaps because more advantaged people (with higher ε_i) are more likely to select into high schooling levels X_i.

To see how IV can address this endogeneity challenge, let's suppose we have some Z_i that is randomly assigned across individuals but which affects schooling decisions. Concretely, let's suppose we randomly give some students a scholarship to attend college and not others: here, $Z_i = 1$ if individual i is a scholarship winner and $Z_i = 0$ otherwise. The randomization of Z_i ensures it is uncorrelated with student ability or background characteristics; if it only affects earnings Y_i through X_i we say it is "excludable" from the model of interest (1) and that $Cov(Z_i, \varepsilon_i) = 0$.[4] In this case, we can use (1) to write

$$\begin{aligned} Cov(Z_i, Y_i) &= \beta Cov(Z_i, X_i) + Cov(Z_i, \varepsilon_i) \\ &= \beta Cov(Z_i, X_i). \end{aligned} \tag{2}$$

[1] Here we are again using "bias" to mean "non-identification"—i.e., that the parameter of interest is not recovered by a particular regression estimand. This should not be confused with the statistical definition of bias—that an estimator is not, in expectation, equal to an estimand of interest. The fact that economists use the same word for these very different concepts is unfortunate, but hopefully not too confusing depending on the context.

[2] The first use of the term "endogenous" in an economics journal appears to be this 1953 poem(!) by Frederick Waugh, of FWL fame: `https://pbs.twimg.com/media/EVF1qLGUMAAj0te?format=png&name=small`. Waugh may have gotten this term from the natural sciences, as it is sometimes used in biology to refer to substances created by or put into an organism as early as the 1920s. These days "endogeneity" is used in the statistical sense across many fields, especially epidemiology.

[3] The inventor of IV appears to be Philip G. Wright who, in 1928, devised it as a solution to the classic simultaneity of supply and demand (discussed below). For more on this history, see `https://scholar.harvard.edu/files/stock/files/wr_5_w.pdf`.

[4] In general, the random assignment of such Z_i will not be enough for it to be excludable from such specifications. We might imagine, for example, that random scholarship offers make students more successful in school by allowing them to not work during college and earn better grades. In that case Z_i might affect earnings Y_i not only through college attendance but by an unmeasured channel of college GPA. Here we are abstracting from such concerns to introduce the IV concept simply.

Thus, provided $Cov(Z_i, X_i) \neq 0$ we can identify the returns-to-schooling parameter β by the estimand $Cov(Z_i, Y_i)/Cov(Z_i, X_i) = \beta$. This is the basic logic of IV; here Z_i is a *valid* instrument for X_i when $Cov(Z_i, \varepsilon_i) = 0$, and a *relevant* instrument when $Cov(Z_i, X_i) \neq 0$.

This simple example captures the core logic of IV, but there are many types of "endogeneity" that IV can solve. Each problem essentially boils down to finding some Z_i which is "valid" in the sense of being uncorrelated with a particular model unobservable and "relevant" in the sense of being correlated with a particular "endogenous variable" or treatment. Let's walk through a few more examples to see exactly how broad this set of problems can be.

First, consider *omitted variables bias* (OVB): a problem you previously saw in Chapter 6. The OVB problem is one in which we are interested in a parameter β from the "long" regression of

$$Y_i = \alpha + \beta X_i + \gamma W_i + \upsilon_i, \tag{3}$$

where $Cov(X_i, \upsilon_i) = Cov(W_i, \upsilon_i) = 0$. The issue is we do not observe W_i, and when it is omitted from our regression we may obtain a biased view of β. Indeed, the bivariate regression of Y_i on X_i gives

$$\begin{aligned} \frac{Cov(X_i, Y_i)}{Var(X_i)} &= \frac{Cov(X_i, \alpha + \beta X_i + \gamma W_i + \upsilon_i)}{Var(X_i)} \\ &= \beta + \gamma \underbrace{\frac{Cov(X_i, W_i)}{Var(X_i)}}_{\delta}. \end{aligned} \tag{4}$$

OVB here is the product of two terms: the "effect" of the omitted variable W_i on the outcome Y_i, γ, and the regression of the omitted variable on the included variable X_i, δ. If we knew the sign of γ and δ, we could sign OVB: if, for example, we knew X_i and W_i were positively correlated and that W_i conditionally positively correlates with Y_i given X_i then we would know both $\gamma > 0$ and $\delta > 0$; then we would know that the bivariate regression overstates the parameter of interest β. Moreover, if we can credibly argue that either γ or δ are zero then we know there is no OVB: the bivariate regression identifies β even without observing W_i. You will often see these sorts of arguments and discussions in papers and seminars when people think through the kinds of biases that may plague their regression estimates; they can be useful heuristics for determining whether one's estimates are likely inflated up or down.

In the OVB setting, a valid IV is one which is uncorrelated with the omitted variables: if $Cov(Z_i, W_i) = Cov(Z_i, \upsilon_i) = 0$ then $\beta = Cov(Z_i, Y_i)/Cov(Z_i, X_i)$, provided $Cov(Z_i, X_i) \neq 0$. The schooling/scholarship example above is basically a version of this more general setup.

A second example is *measurement error*. Here suppose the relationship of interest $Y_i = \alpha + \beta X_i^* + \upsilon_i$ is known to be unconfounded: i.e. we know $Cov(X_i^*, \upsilon_i) = 0$ and that this equation is a regression, for some reason. However we do not observe the regressor of interest X_i^* but instead observe a noisy measure $X_i = X_i^* + \eta_i$, where $Cov(X_i^*, \eta_i) = 0$ and $Cov(\eta_i, \upsilon_i) = 0$. In this case a regression of Y_i on X_i does not identify the parameter β:

$$\begin{aligned} \frac{Cov(X_i, Y_i)}{Var(X_i)} &= \frac{Cov(X_i^* + \eta_i, \alpha + \beta X_i^* + \upsilon_i)}{Var(X_i^* + \eta_i)} \\ &= \beta \frac{Var(X_i^*)}{Var(X_i^*) + Var(\eta_i)}. \end{aligned} \tag{5}$$

Here we have what is sometimes called *attenuation bias*: the regression estimand is a scaled version of the parameter of interest β, with a scaling factor of $\lambda \equiv Var(X_i^*)/Var(X_i^*) + Var(\eta_i)$ strictly between zero and one. In other words, when β is positive we will identify a smaller positive regression coefficient $\beta\lambda < \beta$.

In the measurement error setting we can undo attenuation bias by knowing the "signal-to-noise ratio" λ, which reduces to knowing the variance of X_i^* or the variance of η_i (since then we can solve out for λ from knowledge

of $Var(X_i)$). More directly, if we know $Var(X_i^*)$ we can directly estimate $Cov(X_i, Y_i)/Var(X_i^*) = \beta$. We can also "bound" the degree of attenuation bias if we know something about the possible range of $Var(X_i^*)$.

One example of a valid instrument in the measurement error case is a different mismeasured $Z_i = X_i^* + \xi_i$, satisfying $Cov(X_i^*, \xi_i) = Cov(\eta_i, \xi_i) = 0$. The trick here is that the measurement error in this instrument ξ_i is uncorrelated with the measurement error in the observed $X_i = X_i^* + \eta_i$, which might be true in some cases. Such an instrument is guaranteed to be relevant, since $Cov(Z_i, X_i) = Var(X_i^*) > 0$, and we again have $Cov(Z_i, Y_i)/Cov(Z_i, X_i) = \beta$.

A final example of regression endogeneity is *simultaneity*, which has a long history with IV. The classic example of simultaneous data is supply and demand: suppose we are interested in a demand elasticity β from the system

$$\ln q = \alpha_D + \beta_D \ln p + \upsilon_i \tag{6}$$

$$\ln q = \alpha_S + \beta_S \ln p + \eta_i \tag{7}$$

where q denotes the quantity of some good and p denotes its price. Here equation (6) is a demand equation: β_D tells us how consumer demand increases with the offered price (as an elasticity). Equation (7) is the corresponding supply equation: β_S tells us how producer supply increases with the market price (again as an elasticity). We write υ_i and η_i as demand and supply "shocks" arising across different markets i, normalized to $E[\upsilon_i] = E[\eta_i] = 0$. We observe the equilibrium quantities and prices (Q_i, P_i) which solve the system given by (6) and (7) following the realization of these shocks.

It is easy to see how the simultaneous determination of quantities and prices from this system makes regressions of $\ln Q_i$ on $\ln P_i$ (or vice versa) difficult to interpret. Solving out for these variables, we have

$$\ln P_i = \frac{\alpha_S - \alpha_D + \eta_i - \upsilon_i}{\beta_D - \beta_S} \tag{8}$$

$$\ln Q_i = \frac{\beta_D \alpha_S - \beta_S \alpha_D + \beta_D \eta_i - \beta_S \upsilon_i}{\beta_D - \beta_S}, \tag{9}$$

so long as $\beta_D \neq \beta_S$. You can see from this that a regression of $\ln Q_i$ on $\ln P_i$, or vice versa, fails to identify either the demand or supply elasticity but instead gives some messy formula involving both β_D, β_S, and the relative variances of the shocks η_i and υ_i. This is intuitive, as the equilibrium relationship between prices and quantities is not driven by the variation along either the demand or supply curve, in general, but is instead given by the intersection of these curves as the different shocks move around the equilibrium.

A valid instrument in the simultaneous supply-and-demand case is one that isolates variation in shocks to one of the sides of the market: to identify the demand elasticity β_D we require a shock to the supply side (i.e. η_i) and to identify a supply elasticity β_S we require a shock to the demand side (i.e. υ_i). This is again intuitive, as when we have isolated variation that shifts around one of the two curves (e.g. supply) we are able to trace out the other curve (e.g. demand). Formally, if we have a Z_i with $Cov(Z_i, \eta_i) \neq 0$ (relevance) but $Cov(Z_i, \upsilon_i) = 0$ (validity), then we can see from equations (8) and (9) that

$$\begin{aligned} \frac{Cov(Z_i, \ln Q_i)}{Cov(Z_i, \ln P_i)} &= \frac{Cov(Z_i, \beta_D \alpha_S - \beta_S \alpha_D + \beta_D \eta_i - \beta_S \upsilon_i)}{Cov(Z_i, \alpha_S - \alpha_D + \eta_i - \upsilon_i)} \\ &= \beta_D \end{aligned} \tag{10}$$

and similarly for a "demand-side" instrument that identifies the supply elasticity β_S when $Cov(Z_i, \upsilon_i) \neq 0$ and $Cov(Z_i, \eta_i) = 0$.

In practice, of course, regression endogeneity can manifest in many ways (including combinations of the above stylized examples); the general formulation of the problem is the existence of some "structural" relationship that regression fails to identify due to the correlation between a regressor of interest and an unobserved model residual. We will next formalize the IV solution to such problems in both the simple bivariate case

discussed here, and the more general case with multiple endogenous regressors, multiple instruments, and controls. The basic logic of instrument validity and relevance will continue to hold in that case.

Instrument Validity and Relevance

Let's first define some terms in the simple (bivariate) case, where we have one outcome Y_i, one endogenous variable X_i, and one instrument Z_i. The IV estimand here is

$$\beta = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)} = \frac{Cov(Z_i, Y_i)/Var(Z_i)}{Cov(Z_i, X_i)/Var(Z_i)}. \tag{11}$$

In the second equality we've simply divided the numerator and denominator of the initial definition by $Var(Z_i)$. In doing so we can see that the IV estimand β can be written as the ratio of two regression estimands: $\beta = \rho/\pi$ where

$$Y_i = \kappa + \rho Z_i + \nu_i \tag{12}$$
$$X_i = \mu + \pi Z_i + \eta_i \tag{13}$$

denote bivariate regressions of Y_i and X_i, respectively, on the instrument. Here, by definition of regression, $Cov(Z_i, \nu_i) = Cov(Z_i, \eta_i) = 0$, with $\rho = Cov(Z_i, Y_i)/Var(Z_i)$ and $\pi = Cov(Z_i, X_i)/Var(Z_i)$. We sometimes call equation (12) the "reduced form" regression and equation (13) the "first stage" regression, for reasons that will become more clear shortly.

An alternative but equivalent way to define this IV starts with the "second stage"

$$Y_i = \alpha + \beta X_i + U_i \tag{14}$$

where (α, β) (and thus U_i) are such that $Cov(Z_i, U_i) = 0$. This parallels our definition of population regression as the parameters (and residual) satisfying $Cov(X_i, U_i) = 0$; it is a proper definition so long as $Cov(Z_i, X_i) \neq 0$, since then it can be shown there are unique (α, β) satisfying $Cov(Z_i, U_i) = 0$.[5] This parallels the "no perfect multicollinearity" condition with regression which uniquely defines the regression coefficients. As with regression, we can always define the IV estimand β when this $Cov(Z_i, X_i) \neq 0$; there is always a residual U_i satisfying $Cov(Z_i, U_i) = 0$, just as before how there was always a regression residual satisfying $Cov(X_i, U_i) = 0$. The aim of identification is to make sufficient assumptions on the model such that Z_i is uncorrelated with the model's residual, in which case it coincides with this U_i and β identifies an interesting parameter.

The more general definition of IV starts with a $J \times 1$ vector of endogenous variables $\boldsymbol{X}_i$, a $L \times 1$ vector of instruments $\boldsymbol{Z}_i$, and a $K \times 1$ vector of controls $\boldsymbol{W}_i$ (which includes a constant). Suppose, from an economic model, we arrive at a linear relationship of

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + e_i, \tag{15}$$

where the model parameter of interest is the coefficient vector $\boldsymbol{\beta}$. Here e_i is a model residual (e.g. something related to "potential outcomes"), and need not be orthogonal to $\boldsymbol{X}_i$. We however think the vector of instruments is orthogonal to ε_i after controlling for $\boldsymbol{W_i}$: that is, we think the coefficient on $\boldsymbol{Z}_i$ from the population regression of ε_i on $\underline{\boldsymbol{Z}}_i = [\boldsymbol{Z}_i', \boldsymbol{W}_i']'$ is the $L \times 1$ vector of zeros. The controls here may thus account for some observed confounding between $\boldsymbol{Z}_i$ and ε_i, as with the kind of "selection-on-observables" stories told before (but now with $\boldsymbol{Z}_i$ instead of the actual "treatment" $\boldsymbol{X}_i$; more on this below). To accommodate them, let's imagine projecting e_i on the control vector to obtain

$$e_i = \boldsymbol{W}_i'\boldsymbol{\gamma} + \varepsilon_i, \tag{16}$$

[5] Here $\alpha = \kappa - \beta\mu$ and $U_i = \nu_i - \beta\eta_i$, by substituting the reduced form and first stage equations into the second stage equation. The reduced form and first stage coefficients are unique provided $Var(Z_i) \neq 0$, which is implied by $Cov(Z_i, X_i) \neq 0$. Further, as shown above, $\beta = \rho/\pi$.

where ε_i is by construction orthogonal to $\boldsymbol{W}_i$. Combining (16) and (15), we have our second stage equation:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \varepsilon_i. \tag{17}$$

How can we use instrument exogeneity in this case? Motivated by the simple case above, let's consider the first stage regression of the endogenous variable on the instrument and controls. Here we have one such regression for each row of $\boldsymbol{X}_i$; stacking these, we have

$$\boldsymbol{X}_i = \boldsymbol{\Pi}\boldsymbol{Z}_i + \boldsymbol{\mu}\boldsymbol{W}_i + \boldsymbol{\eta}_i, \tag{18}$$

where $\boldsymbol{\Pi}$ is a $J \times L$ matrix of coefficients from regressing each $\boldsymbol{X}_{ij}$ on $\boldsymbol{Z}_i$ while controlling for $\boldsymbol{W}_i$. By construction, $\boldsymbol{\eta}_i$ is orthogonal to both $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$. Substituting this series of first stage regressions into the second stage (17), we obtain

$$\begin{aligned} Y_i &= (\boldsymbol{\Pi}\boldsymbol{Z}_i + \boldsymbol{\mu}\boldsymbol{W}_i + \boldsymbol{\eta}_i)'\boldsymbol{\beta} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \varepsilon_i \\ &= (\boldsymbol{\Pi}\boldsymbol{Z}_i)'\boldsymbol{\beta} + \boldsymbol{W}_i'(\boldsymbol{\mu}'\boldsymbol{\beta} + \boldsymbol{\gamma}) + (\boldsymbol{\eta}_i'\boldsymbol{\beta} + \varepsilon_i). \end{aligned} \tag{19}$$

From this we can see that $\boldsymbol{\beta}$ is identified, by the population regression of Y_i on $\boldsymbol{\Pi}\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$, under two conditions.

The first condition is the generalized IV validity assumption, that $E[\boldsymbol{\Pi}\boldsymbol{Z}_i\varepsilon_i] = 0$. This is enough to ensure equation (19) is a regression, since we know $\boldsymbol{\eta}_i$ is by construction orthogonal to both $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$ (so the linear combination $\boldsymbol{\eta}_i'\boldsymbol{\beta}$ is orthogonal to both the linear combination $\boldsymbol{\Pi}Z_i$ and to $\boldsymbol{W}_i$) and we know that ε_i is orthogonal to $\boldsymbol{W}_i$ by definition. A sufficient condition for IV validity is the conditional orthogonality of $\boldsymbol{Z}_i$ and ε_i given $\boldsymbol{W}_i$, by the Frisch-Waugh-Lovell theorem. For example, if ε_i denoted fixed student ability, $\boldsymbol{Z}_i$ were a vector of randomized scholarship offers, and $\boldsymbol{W}_i$ contained information on which scholarship lotteries a student was entered into, we may expect $E[\boldsymbol{\Pi}\boldsymbol{Z}_i\varepsilon_i] = 0$.

The second condition in the generalized IV relevance condition, which here resolves to a no perfect collinearity assumption on $[(\boldsymbol{\Pi}\boldsymbol{Z}_i)', \boldsymbol{W}_i']'$. This in turn resolves to no perfect multicollinearity in $[\boldsymbol{Z}_i', \boldsymbol{W}_i']'$ (which allows us to define the first stage regressions) and an assumption that $\boldsymbol{\Pi}$ is of full row rank. This rank condition could fail when, for example, there are fewer instruments than endogenous variables ($L < K$) or more generally when the instruments do not generate independent variation in all of the endogenous variables.

As above we've derived the IV validity and relevance condition by starting with a model satisfying them, but when the relevance condition holds it can be shown there is always a second stage residual satisfying the validity condition. Namely, so long as $\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$ are not perfectly collinear we can always define the first stage regression (18). Let $\underline{\boldsymbol{X}}_i = [\boldsymbol{X}_i', \boldsymbol{W}_i']'$ collect the endogenous variables and controls, let $\underline{\boldsymbol{Z}}_i = [\boldsymbol{Z}_i', \boldsymbol{W}_i']'$ collect the instruments and controls, and let

$$\underline{\boldsymbol{\Pi}} = \begin{bmatrix} \boldsymbol{\Pi} & \boldsymbol{\mu} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \tag{20}$$

collect the first-stage coefficients. Then it can be shown the IV "moment condition"

$$\boldsymbol{E}[\underline{\boldsymbol{\Pi}}\underline{\boldsymbol{Z}}_i U_i] = E[\underline{\boldsymbol{\Pi}}\underline{\boldsymbol{Z}}_i(Y_i - \underline{\boldsymbol{X}}_i'\underline{\boldsymbol{\beta}})] = 0,$$

which imposes orthogonality of the second-stage residual $U_i = Y_i - (\boldsymbol{\Pi}\boldsymbol{Z}_i)'\boldsymbol{\beta} - \boldsymbol{W}_i'\boldsymbol{\gamma}$ with the regressors $\boldsymbol{\Pi}\boldsymbol{Z}_i$ and $\boldsymbol{W}_i$, has a unique solution

$$\underline{\boldsymbol{\beta}} = [\boldsymbol{\beta}', \boldsymbol{\gamma}']' = \boldsymbol{E}[\underline{\boldsymbol{\Pi}}\underline{\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']^{-1}\boldsymbol{E}[\underline{\boldsymbol{\Pi}}\underline{\boldsymbol{Z}}_i Y_i]. \tag{21}$$

Thus, we can always define the general IV estimand $\underline{\boldsymbol{\beta}}$ when relevance holds, just as we could define the "simple" IV estimand when $Cov(Z_i, X_i) \neq 0$ or define the population regression of Y_i on X_i when $Var(X_i)$. As in these cases, the identification question is whether the statistical residual U_i, which imposes instrument validity, coincides with the residual of a particular model for how the data are generated. If it does, then we know the coefficients of the IV regression coincide with the coefficients from that "structural" second stage

equation. Furthermore, we can see that the IV estimand (21) is relatively straightforward to estimate, as it is a relatively simple function of second moments; more on that soon.

You shouldn't be too surprised if all of this is sounding familiar; the link between IV and linear regression is tight because the latter is a special case of the former. Formally, when $\boldsymbol{X}_i = \boldsymbol{Z}_i$, the first stage regression of $\boldsymbol{X}_i$ on $\boldsymbol{Z}_i$ fits perfectly; the first stage matrix is then $\underline{\boldsymbol{\Pi}} = \boldsymbol{I}$ and the IV estimand

$$\begin{aligned}\underline{\boldsymbol{\beta}} &= \boldsymbol{E}[\underline{\boldsymbol{\Pi}\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']^{-1}\boldsymbol{E}[\underline{\boldsymbol{\Pi}\boldsymbol{Z}}_i Y_i] \\ &= \boldsymbol{E}[\underline{\boldsymbol{X}}\,\underline{\boldsymbol{X}}_i']^{-1}\boldsymbol{E}[\underline{\boldsymbol{X}} Y_i],\end{aligned}$$

is just population regression. The IV relevance condition here is satisfied just by the lack of perfect multicollinearity in $\underline{\boldsymbol{Z}}_i = \underline{\boldsymbol{X}}_i$ and IV validity is simply $\boldsymbol{E}[\underline{\boldsymbol{X}}_i\varepsilon_i] = 0$. Working through this special case is useful for showing how IV allows some endogenous "slippage" between $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$; instead of regressing Y_i on $\boldsymbol{X}_i$, we regress on the component of $\boldsymbol{X}_i$ which is predicted by the exogenous instrument (the fitted values).

Of course, since IV arises just from regression the Frisch-Waugh-Lovell theorem applies. We can, for example, think of regressing Y_i on the residuals from projecting $\boldsymbol{\Pi}\boldsymbol{Z}_i$ on $\boldsymbol{W}_i$ by the first part of the FWL. This will come in handy for analyzing some IV coefficients, as before.

It may also be handy to work with the generalized IV's reduced form and first stage expressions:

$$Y_i = \boldsymbol{Z}_i'\boldsymbol{\rho} + \boldsymbol{W}_i'\boldsymbol{\kappa} + \nu_i \tag{22}$$

$$\boldsymbol{X}_i = \boldsymbol{\Pi}\boldsymbol{Z}_i + \boldsymbol{\mu}\boldsymbol{W}_i + \boldsymbol{\eta}_i, \tag{23}$$

where, per (19), we have $\boldsymbol{\rho} = \boldsymbol{\Pi}'\boldsymbol{\beta}$, $\boldsymbol{\kappa} = \boldsymbol{\mu}'\boldsymbol{\beta} + \boldsymbol{\gamma}$, and $\nu_i = \boldsymbol{\eta}'\boldsymbol{\beta} + U_i$. Consider the case where $L = dim(\boldsymbol{Z}_i) = dim(\boldsymbol{X}_i) = K$; we call this case "just-identified," in that there are just as many instruments as endogenous variables. Here $\boldsymbol{\Pi}$ is a square matrix, which is invertible (i.e. full rank) when the relevance condition holds. Thus we can define $\boldsymbol{\beta} = \Pi'^{-1}\boldsymbol{\rho}$ in this case, generalizing how we defined IV as the reduced form $\rho = Cov(Z_i, Y_i)/Var(Z_i)$ divided by the first stage $\pi = Cov(Z_i, X_i)/Var(Z_i)$ in the simple (bivariate) case above, which was just-identified. The FWL also of course applies here, allowing us to study Π and ρ by first residualizing out the controls.

In the just-identified case, where $\boldsymbol{\Pi}$ is invertible, the IV validity condition becomes equivalent to the orthogonality of the instrument vector with the residual: $E[\boldsymbol{\Pi}\boldsymbol{Z}_i\varepsilon_i] = 0$ if and only if $\boldsymbol{\Pi}^{-1}E[\boldsymbol{\Pi}\boldsymbol{Z}_i\varepsilon_i] = \boldsymbol{\Pi}^{-1}0$ or $E[\boldsymbol{Z}_i\varepsilon_i] = 0$. In the just-identified case any full-rank linear combination of $\boldsymbol{Z}_i$ is valid and gives the same IV estimand. In the "overidentified" case of $L = dim(\boldsymbol{Z}_i) > dim(\boldsymbol{X}_i) = K$, where we have more instruments than endogenous variables, this is no longer true: different linear combinations of the $\boldsymbol{Z}_i$ may or may not be valid and will yield different IV estimands. If we assume the stronger validity condition holds, that $E[\boldsymbol{Z}_i\varepsilon_i] = 0$, then we can use any $\boldsymbol{M}\boldsymbol{Z}_i$ as a set of L instruments, for any full-rank $J \times L$ matrix $\boldsymbol{M}$. More specifically, we can consider the class of IV estimands

$$\underline{\boldsymbol{\beta}} = \boldsymbol{E}[\underline{\boldsymbol{M}\boldsymbol{Z}}_i\underline{\boldsymbol{X}}_i']^{-1}\boldsymbol{E}[\underline{\boldsymbol{M}\boldsymbol{Z}}_i Y_i] \tag{24}$$

for any $\underline{\boldsymbol{M}}$, not just $\underline{\boldsymbol{\Pi}}$. This class is quite large, containing some well-studied IV estimands such as the Nagar (1962), k-class, or limited information maximum likelihood (LIML) procedures that you may encounter in future econometrics classes. In this class, however, we will focus on the $\underline{\boldsymbol{M}} = \underline{\boldsymbol{\Pi}}$ case even when overidentified.

Where do Instruments Come From?

So far we've talked about instruments in the abstract, and from that perspective they sound pretty magical: what are these "exogenous" Z_i and how are they actually used in practice? Here we will walk through three examples, from Abdulkadiroglu et al. (2016) (on charter school effectiveness) and Angrist and Krueger (1991) (on—what else?—the returns to schooling).

Some of the best candidates for instruments come from true experiments, in which Z_i is as-good-as-randomly assigned across observations i. Abdulkadiroglu et al. (2016), for example, use the random assignment of

offers to attend charter middle schools as an instrument for charter school enrollment. The idea is that when students apply to an "oversubscribed" charter, with more applicants than available seats, the school runs a simple lottery to determine who is eligible to attend. Those who receive offers can decline, and go somewhere else, while other students may find their way into the charter through later admission rounds. Thus, while the randomized offers are a strong predictor of charter enrollment there is still a considerable amount of "slippage" in terms of enrollment (both conditional on application and unconditionally, as most students do not apply to enroll in a charter school). Formally, Abdulkadiroglu et al. (2016) study the effects of charter enrollment $X_i \in \{0, 1\}$ on subsequent test score achievement Y_i, instrumenting by a $Z_i \in \{0, 1\}$ that indicates student i got an offer to attend a charter on the night of the lottery.[6] We estimate this regression in the sample of charter applicants, controlling for strata $\boldsymbol{W}_i$ indicating different lottery years and schools. The upshot is we estimate very large test score effects from charter enrollment, consistent with other papers in a recent literature on charter effectiveness.

A virtue of IVs like charter enrollment offers is that they are (conditionally) randomly assigned; we thus know for sure that we can estimate their "reduced form" effects on test scores Y_i as well as the "first stage" effect on charter enrollment X_i. That is, we can estimate the ATT $E[Y_{i1} - Y_{i0}]$ of getting an charter *offer* on test scores—sometimes in the IV context this is referred to as an "intent-to-treat" effect (the idea being Z_i captures the randomized "intent" to receive the endogenous enrollment treatment X_i)—as well as $E[X_{i1} - X_{i0}]$. A key point to recognize, however, is that random assignment is *not* sufficient for such Z_i to be a valid instrument for X_i. For this we also need an *exclusion restriction*, that Z_i only affects Y_i through X_i. In the charter school case of Abdulkadiroglu et al. (2016) this seems fairly defensible: admission offers likely only affect later achievement via enrollment decisions, having no real effects other than giving a student access to attend a charter school. This sort of logic is often found in randomized control trials (RCTs) with "imperfect compliance," where an offer to participate in a program is randomized but people can opt out or in. These days most researchers understand that instrumenting by offers can identify causal program effects despite such imperfect incompliance, though economists were a driving force behind making this clear to different fields.[7]

Absent literal randomization of Z_i, researchers may still credibly argue that it is "as-good-as-randomly" assigned (perhaps conditional on some $\boldsymbol{W}_i$). The idea here is to appeal to a kind of "natural experiment" which generates Z_i in a way that is plausibly unrelated to the second-stage model of interest. Angrist and Krueger (1991) give a now-famous example of such a setting when estimating the returns to schooling in the early 20th century. They leverage two institutional features of this time: compulsory schooling laws, which typically required a student to stay in school until their 16th birthday, and the fact that most schools require students to enter school in the calendar year they turn six. Consequently, students born in different quarters who plan to drop out as soon as they are able will tend to have different completed years of schooling. A student born in January, for example, will start school at six and eight months and at her 16th birthday will have nine years of completed schooling. A student born in December, in contrast, will start school at five and eight months and at her 16th birthday will have completed ten years of schooling. Angrist and Krueger (1991) thus use the "natural experiment" of quarter-of-birth as an instrument for completed years of schooling, controlling for the year- and state-of-birth (to help with the instrument's first-stage power).

One's quarter-of-birth may appear as-good-as-randomly assigned with respect to the labor market conditions (and other factors) one faces in adulthood. Even though people do not time the conception of their children by a lottery, this "natural experiment" seems fairly plausible. Again, however, we must consider not only the independent assignment of this Z_i across individuals but also ponder the key exclusion restriction: does

[6] As you'll see in the paper, and course slides, we actually use two instruments in the main specifications of the paper: an "immediate offer" to enroll on lottery night, and a "waitlist offer" to enroll later. The latter comes from the fact that we know each student's position on the school waiting list, which is randomized on lottery night. So we can define arbitrary cutoffs on this randomized wait list and use it as an instrument too. In practice we get very similar estimates with both instruments or just the "immediate offer" IV.

[7] A famous result by Imbens and Angrist (1994), formalized this approach by showing when Z_i and X_i are binary such IV regressions identify "local average treatment effects" (LATEs), defined as the average effect of X_i on Y_i among marginal individuals ("compliers") who are induced to the treatment by the randomized offer. Such an interpretation requires an additional "monotonicity" condition which says the randomized offer can only shift people into taking the treatment. We may have the opportunity to say more about LATEs and related parameters in the final lectures or TA sessions.

quarter-of-birth only affect adult earnings through completed schooling? More recent studies have, for example, found that being older in your class in grade school can have direct effects on both mental and physical development which may conceivably violate the exclusion restriction. It turns out that the Angrist and Krueger (1991) student may also have suffered from a different problem, related to estimation instead of identification, as we will cover in the next Chapter. Still, it is a compelling story at first glance as well as an influential early example of our modern view of such "natural experiments."

Both of these examples highlight a useful way of thinking about instruments, by separating a statistical assumption of (as-good-as) random assignment from the more model-based exclusion restriction. We can think of other ways to ensure random assignment by leveraging what we've seen in "reduced form" treatment effect estimation. For example we might tell a "selection-on-observables" story that makes a given instrument Z_i as-good-as-randomly assigned conditional on some $\boldsymbol{W}_i$, even though the ultimate treatment of D_i is not unconfounded. With panel data, we might use a difference-in-differences approach to argue $Z_i Post_t$ satisfies a parallel trends assumption controlling for Z_i and $Post_t$ main effects. Given such arguments we still need to be able to credibly argue an exclusion restriction holds, in order to relate the (say) reduced-form difference-in-difference estimates of the effect of Z_i on Y_i to first-stage difference-in-difference estimates of the effect of Z_i on X_i. Again, such arguments tend to be "model-based," requiring us to rule out stories of other direct effects of the instruments on outcomes. Abdulkadiroglu et al. (2016) pursue such an approach in their study of non-lotteried "takeover" charters, which sits alongside their lottery analysis discussed above. See the course slides for an illustration of these two approaches.

This discussion of where IVs come from, and the example of Abdulkadiroglu et al. (2016) in particular, highlight a general tradeoff between *internal* and *external* validity of observational (possibly IV-based) analyses. In many ways the "gold-standard" for estimating causal effects is a randomized treatment; since analyzing such an RCT requires minimal assumptions (besides the existence of potential outcomes) we sometimes say that it has high internal validity. In contrast, an observational study of the treatment's effects which makes hard-to-swallow selection-on-observables assumptions may have low internal validity (in that it requires assumptions that are not guaranteed by virtue of randomization). But (far from) all treatments of interest can be or are randomized, and those that are often can only be deployed on selected individuals. In the Abdulkadiroglu et al. (2016) example, we have high internal validity for estimating the effects of charter school enrollment among the students who enter the admission lottery. We may worry about the external validity (i.e. generalizability) of such studies, especially when individuals self-select into the study population (by, e.g., applying to a charter school). To probe external validity, it is often helpful to turn to more observational studies (i.e. those that make a selection-on-observables argument or rely on difference-in-difference-type identification) on a more representative population. The point is that IV can help on both fronts, but not all IVs are created equal in this regard. The exclusion restriction can fail even when the instrument is randomized in a lottery, and parallel trends can be very credible even in a non-experimental setting.

As with our discussion of population regression, we'll next turn to the question of how we estimate IV estimands from data. The key insight, as before, is that these $\underline{\boldsymbol{\beta}}$ are also relatively simple functions of second moments; we can thus consider their sample analogues to construct an estimator of $\underline{\boldsymbol{\beta}}$, and follow similar steps as with OLS to characterize its asymptotic behavior. There are, however, a few new practical considerations with IV that we didn't have before. These are related to the fact that we now have some "slippage" between the exogenous $\boldsymbol{Z}_i$ and endogenous $\boldsymbol{X}_i$, and can be tricky to deal with in practice.