Len Goff
Microeconometrics GR6414

Notes on conditional expectations and causal regression

Please let me know if you spot any typos, etc.!

Properties of CEFs

Suppose (X_i, Y_i) are jointly continuously distributed. Then the conditional density of Y on X is defined as $f_{y|x}(y|x) := \frac{f_{xy}(x,y)}{f_x(x)}$. For any fixed value of x for which $f_x(x) > 0$, this defines a new univariate density function $f_{y|x}(y|x)$ that is positive and integrates to one. The conditional expectation function (CEF) is simply the expected value of this conditional density, as a function of x: (note that I use the notation := for definitions)

$$m(x) := \mathbb{E}\left[Y_i|X_i = x\right] := \int y f_{y|x}(y|x)dy \tag{1}$$

When X and/or Y have discrete support, things are defined analogously with probability mass functions and sums instead of integrals. Why do we like CEF's? They have several nice properties, which I'll prove on the board in recitation (also see Chapter 3 of Angrist & Pischke 2008, henceforth MHE):

1. $m(x)$ is the function of X_i that serves as the best predictor of Y_i, in the mean-squared error sense (MHE 3.1.2):

   $$\mathbb{E}\left[Y_i|X_i = x\right] = \underset{h(X_i)}{\operatorname{argmin}}\ \mathbb{E}\left[(Y_i - h(X_i))^2\right]$$

   Note for the interested: $m(X_i)$ can be thought of as a "projection" of the random variable Y_i onto the space of random variables that are functions of X_i, where the inner-product between random variables is defined by the expectation of their product.

2. We can write $Y_i = \mathbb{E}\left[Y_i|X_i = x\right] + \epsilon_i$, where $\mathbb{E}\left[\epsilon_i|X_i = x\right] = 0$ (MHE 3.1.1). Hence $\mathbb{E}\left[h(X_i)\epsilon_i\right] = 0$ for any function $h(x)$.

This final property depends upon the law of iterated expectations (LIE), which is a fundamental tool in regression analysis and econometrics generally. In our context LIE states that

$$\mathbb{E}[Y_i] := \mathbb{E}\left[\mathbb{E}\left[Y_i|X_i\right]\right]$$

What does this expression mean? Note the notation $\mathbb{E}\left[Y_i|X_i\right] = m(X_i)$: this is the function $m(x)$ (which is not random, rather it is a fixed property of the joint distribution of X and Y), evaluated at an individual realization of the random variable X_i. Hence $\mathbb{E}\left[Y_i|X_i\right]$ is a random variable who's distribution depends on the distribution of X_i. LIE states that if we take the expected value of *this* random variable (i.e. integrate over the distribution of X_i), then we'll get back the unconditional expectation of Y_i. Woot!

For continuously distributed (X_i, Y_i), this can be seen as follows:

$$\mathbb{E}\left[\mathbb{E}\left[Y_i|X_i\right]\right] = \int \mathbb{E}\left[Y_i|X_i = x\right] f_x(x)dx = \int \left\{\int y f_{y|x}(y|x)dy\right\} f_x(x)dx = \int y f_y(y)dy = \mathbb{E}[Y_i]$$

where to move from the second to the third equality we reverse the order of integration, note that $f_{y|x}(y|x)f_x(x) = f_{xy}(x,y)$, and then evaluate the integral over x.

Linear regression

Suppose X is a k-dimensional vector (typically including a constant) and that the CEF of Y on X is *linear*:

$$m(x) = x'\beta$$

for some $\beta \in \mathbb{R}^k$. Then Property 2. from the previous section implies that $Y_i = X_i'\beta + \epsilon_i$, where $\mathbb{E}[X_i\epsilon_i] = 0$ (where 0 here is a k-dimensional vector of zeros). Look familiar? This is exactly what we expect of the residual in a linear regression. A linear CEF with coefficient vector β means that we can determine β by performing a linear regression of Y on X.

Even when $m(x)$ is not actually linear in x, we can still always define a parameter β by

$$\beta := \mathbb{E}[X_iX_i']^{-1}\mathbb{E}[X_iY_i] \tag{2}$$

so long as $\mathbb{E}[X_iX_i']$ is invertible.[1] This β is known as the *linear projection* coefficient of Y_i onto X_i (this "coefficient" is generally a vector). The linear projection coefficient is the solution to the OLS minimization problem (see Property 1. below).[2] In this sense, regression is always estimating the linear projection coefficient, and when the CEF is linear, this is also equal to the slope vector of the CEF. Note that if we define $\epsilon_i = Y_i - X_i'\beta$ with β defined by Equation (2), then $Y_i = X_i'\beta + \epsilon_i$ with $\mathbb{E}[\epsilon_i X_i] = 0$. When the CEF is linear, then we also have that $\mathbb{E}[\epsilon_i|X_i] = 0$, which is a stronger property.

The linear projection coefficient has two nice properties that show that it might be expected to do a good job of summarizing the relationship of Y and X, even when the CEF is not quite linear.

1. $X_i'\beta$ with β defined by Equation (2) provides the best linear predictor of Y_i, in the mean-squared error sense (MHE Thm. 3.1.5):

$$\beta = \underset{\gamma}{\operatorname{argmin}}\, \mathbb{E}\left[(Y_i - X_i'\gamma)^2\right]$$

[1] This is equivalent to there being no perfect (i.e. certain, occurring with probability one) linear dependence between the components of X_i, since: $\mathbb{E}[X_iX_i']$ invertible $\iff$ $\gamma'\mathbb{E}[X_iX_i']\gamma \neq 0$ for all $\gamma \in \mathbb{R}^k$ $\iff$ $\mathbb{E}[(X_i'\gamma)^2] > 0$ for all $\gamma \in \mathbb{R}^k$ $\iff$ for all $\gamma \in \mathbb{R}^k$, $P(X_i'\gamma \neq 0) > 0$.

[2] To see this, note that the first order condition to $\min\limits_{\gamma} \mathbb{E}\left[(Y_i - X_i'\gamma)^2\right]$ is $2\mathbb{E}[X_i\,(Y_i - X_i'\gamma)] = 0$ and rearrange.

and is the best linear approximation to the CEF, in the mean-squared error sense (MHE Th.m 3.1.6)

$$\beta = \underset{\gamma}{\operatorname{argmin}}\, \mathbb{E}\left[(\mathbb{E}[Y_i|X_i] - X_i'\gamma)^2\right]$$

so at least we know that linear regression is doing the "best job" possible among all linear functions of X_i

2. Regardless of the true functional form of the CEF, β is still related to it. By LIE, we have that

$$\beta = \mathbb{E}[X_iX_i']^{-1}\mathbb{E}[X_i\mathbb{E}[Y_i|X_i]] = \mathbb{E}[X_iX_i']^{-1}\mathbb{E}[X_im(X_i)]$$

which shows that the value of β depends on two things: the true CEF function $m(x)$, and the marginal distribution of X_i. But it's not obvious from this matrix expression exactly what β is telling us about $m(x)$. Yitzakhi (1989) shows that in the univariate case, β provides a weighted average of the derivative $m'(x)$ of the true CEF.[3] So, even if the true CEF $m(x)$ is not linear, linear regression still tells us a certain summary of how the CEF depends on x.

This all gets a bit more complicated when there are multiple variables, and more care is needed to relate the regression coefficient for one variable to the CEF when the CEF is nonlinear. Angrist & Krueger (1999, eq. 34) consider a condition under which an analogous property to Yitzhaki's holds with covariates in the regression.

Linear regression and causation

As much as we might worry about linearity, there's a deeper question about CEFs: under what conditions do they tell us about the *causal* effect of X on y? The simplest condition that could allow us to use regressions to get at causal effects is the conditional independence assumption (CIA), also known as *selection on observables* or *unconfoundedness.*

Consider the potential outcomes notation Y_{0i}, Y_{1i} for a binary treatment variable D_i. Then, the CIA states that

$$(Y_{0i}, Y_{1i}) \perp D_i|X_i$$

where X_i is a set of observed covariates (we'll exclude a constant from X_i in our notation). Now let's see how the CIA can fit into a setup in which linear regression reveals causal

[3] For the fascinated or incredulous: following Yitzakhi (1989), write $C(Y,X) = \mathbb{E}[Y(X-\mu_x)] = \mathbb{E}[m(X)(X-\mathbb{E}[X])] = \int f(x)m(x)(x-\mathbb{E}[X])dx$, using LIE. Then, integrate by parts where $u = f(x)(x-\mathbb{E}[X])$ and $dv = m(x)dx$ so that $C(Y,X) = v(x)g(x)|_{-\infty}^{\infty} - \int m'(x)v(x)dx$ where $v(x) = \int_{-\infty}^{x} f(t)(t-\mathbb{E}[X])$. The first term is zero because both $v(\infty)$ and $v(-\infty)$ are equal to zero (for $v(\infty)$ we assume X_i has a finite second moment). So, $\frac{C(Y,X)}{V(X)} = \int m'(x)w(x)dx$ where $w(x) = -v(x)/V(X)$. To see that $w(x)$ integrates to one, substitute $Y = X$ in which case $m'(x) = 1$. To see that the weights are positive, rewrite $v(x) = F(x)\mathbb{E}[X|X \leq x] - \mathbb{E}[X]F(x) = F(x)\left(\mathbb{E}[X|X \leq x] - \mathbb{E}[X]\right)$ and note that $\mathbb{E}[X|X \leq x] \leq \mathbb{E}[X]$ for all values of x.

effects. Let's assume that

$$\begin{aligned}
&(1)\quad Y_{1i} - Y_{0i} = \beta && \text{(homogenous treatment effects)} \\
&(2)\quad \mathbb{E}[Y_{0i}|X_i] = \alpha + X_i'\gamma && \text{(linearity of } Y_0 \text{ CEF)} \\
&(3)\quad \mathbb{E}[Y_{0i}|X_i, D_i] = \mathbb{E}[Y_{0i}|X_i] && \text{(CIA)}
\end{aligned}$$

Note that (3) is implied by the CIA, and that (2) can be written as $Y_{0i} = \alpha + \eta_i$ where $\mathbb{E}[\eta_i] = X_i'\gamma$. This just states that the CEF of Y_0 on X is linear, which justifies using linear regression to control for X_i. This assumption can be generalized, as can the assumption all units have the same treatment effect. But (1) and (2) simplify the math to help make the central point about causality and the CIA.

Note that (2) and (3) together imply that $\mathbb{E}[Y_{0i}|X_i, D_i] = \alpha + X_i'\gamma$ and then combining with (1) we have that:

$$\mathbb{E}[Y_i|X_i, D_i] = \alpha + \beta D_i + X_i'\gamma$$

where recall that $Y_i = Y_{D_i i} = (1 - D_i)Y_{0i} + D_i Y_{1i}$. Thus, since $\nu_i := Y_i - \mathbb{E}[Y_i|X_i, D_i]$ is mean independent of D_i and X_i, we can estimate β and γ from a linear regression (though we don't directly care about γ).

It's instructive to think about what we'd get if we didn't have the CIA. Suppose that instead of (2) and (3) we had $\mathbb{E}[Y_{0i}|X_i, D_i] = x_i'\gamma + \rho D_i$. Then we would have

$$\mathbb{E}[Y_i|X_i, D_i] = \alpha + (\beta + \rho)D_i + X_i'\gamma$$

The term ρ indicates the bias due to a failure of the CIA. If D_i is not "as good as randomly assigned" conditional on X_i, then "controlling" for X_i in a regression of Y_i on D_i is not sufficient to give the regression a causal interpretation.

To see how all of this generalizes to a treatment variable that may take on many values (e.g. schooling), see Section 3.2 of MHE.

References

Angrist, Joshua & Pischke, Jörn-Steffen. (2008). Mostly Harmless Econometrics: An Empiricist's Companion.

Yitzhaki, Shlomo. (1996). On Using Linear Regressions in Welfare Economics. *Journal of Business & Economic Statistics*, 14(4), 478-486. doi:10.2307/1392256

Len Goff
Microeconometrics GR6414

Notes on regression anatomy and omitted variables

Please let me know if you spot any typos, etc.!

Regression anatomy

Recall that in bivariate regression (also known as "simple linear regression"), where we have a single scalar X and an intercept:

$$\beta = \frac{C(Y_i, X_i)}{V(X_i)}$$

$$\alpha = E[Y_i] - \beta E[X_i]$$

where for random variables A and B: $C(A, B)$ denotes their covariance and $V(A)$ the variance of A.

Note: How do we know this? Recall that α and β minimize the quantity $E\left[(Y_i - \alpha - \beta X_i)^2\right]$. The first-order condition of this problem with respect to α is:

$$-2E\left[(Y_i - \alpha - \beta X_i)\right] = 0 \iff \alpha = E[Y_i] - \beta E[X_i]$$

The first-order condition of with respect to β is:

$$-2E\left[(Y_i - \alpha - \beta X_i)X_i\right] = 0 \iff \beta E[X_i^2] = E[X_iY_i] - \alpha E[X_i]$$

If we substitute our expression for α in, we get:

$$\beta\left(E[X_i^2] - E[X_i]^2\right) = E[X_iY_i] - E[X_i]E[Y_i]$$

which is the same as saying

$$\beta = \frac{C(Y_i, X_i)}{V(X_i)}$$

.

Now consider a regression of Y_i on K different variables $X_{1i}, X_{2i}, \dots X_{Ki}$. We want to know the coefficients $\beta_0, \beta_1, \dots \beta_K$ that minimize the quantity

$$E[(Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_K X_{Ki})^2] \tag{1}$$

(to simplify the notation we have renamed the constant term α to β_0). One approach to this problem is to write the set of coefficients as a vector and minimize Eq (1) with respect to this vector, which yields a matrix equation for β. Regression anatomy gives us another approach, expressing each element β_k as a bivariate regression coefficient.

In particular, the regression anatomy formula states that

$$\beta = \frac{C(Y_i, \tilde{X}_{ki})}{V(\tilde{X}_{ki})}$$

where $\tilde{X}_{ki}$ is defined as the residual from a regression of X_{ki} on a constant and all of the other $X's$ (excluding X_{ki}). To see this, write:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + e_i \tag{2}$$

where we know that for all $k = 1 \ldots K$, the first order condition of minimizing Eq (1) with respect to β_k is that $E[e_i X_{ki}] = 0$.

By the same logic, we can write the following regression equation of X_{ki} on a constant and all of the other $X's$:

$$X_{ki} = \tilde{\beta}_0^{[k]} + \tilde{\beta}_1^{[k]} X_{1i} + \cdots + \tilde{\beta}_{k-1}^{[k]} X_{k-1,i} + \cdots + \tilde{\beta}_{k+1}^{[k]} X_{k+1,i} + \cdots + \tilde{\beta}_K^{[k]} X_{K,i} + \tilde{X}_{ki} \tag{3}$$

We know that for all $j = 1 \ldots k-1, k+1, \ldots K$, the first order condition with respect to $\tilde{\beta}_j^{[k]}$ tells us that $E[\tilde{X}_{ki} X_{ji}] = 0$. Note that the first order condition with respect to $\tilde{\beta}_0^{[k]}$ is $E[\tilde{X}_{ki}] = 0$, which means that for any random variable V, $C(V, \tilde{X}_{ki}) = E[V\tilde{X}_{ki}]$. Together, we then have that $C(\tilde{X}_{ki}, X_{ji}) = 0$ for any $j = 1 \ldots k-1, k+1, \ldots K$.

Using Eq (2) and linearity of the covariance operator,

$$C(Y_i, \tilde{X}_{ki}) = 0 + \beta_1 C(X_{1i}, \tilde{X}_{ki}) + \cdots + \beta_K C(X_{Ki}, \tilde{X}_{ki}) + C(\epsilon_i, \tilde{X}_{ki}) \tag{4}$$

where the first term is zero since β_0 is not random. We've also established that $C(X_{ji}, \tilde{X}_{ki}) = 0$ for any $j = 1 \ldots k-1, k+1, \ldots K$. Finally, notice that we also have $C(\epsilon_i, \tilde{X}_{ki}) = 0$ because from Equation (3) we can write $\tilde{X}_{ki}$ as a linear function of the X_{ki} and $C(e_i, X_{ki}) = E[e_i X_{ki}] = 0$ for all $k = 1 \ldots K$. Thus, only the X_{ki} term in Eq (4) is nonzero and $C(Y_i, \tilde{X}_{ki}) = \beta_k C(X_{ki}, \tilde{X}_{ki})$.

Our last step to arrive at the regression anatomy formula is to show that $C(X_{ki}, \tilde{X}_{ki}) = V(\tilde{X}_{ki})$. To see this, substitute in Eq (3) for X_{ki}, and notice that only the last term is nonzero.

For those that are interested: if we collect as vectors all of the regression coefficients $\beta := (\beta_0, \beta_1 \dots \beta_K)'$ and regressors $X_i := (1, X_{1i} \dots X_{Ki})'$, recall that $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$. How can we show that the regression anatomy formula holds from this matrix expression? One (messy) way would be to reorganize the vectors like $X_i = (X_{ki}, X_{-k,i})'$ where $X_{-k,i} := (1, X_{1i} \dots X_{k-1,i}, X_{k+1,i} \dots X_{Ki})'$ and then use block matrix inversion identities.

A generalization of the regression anatomy formula is the so-called Frisch-Waugh-Lovell theorem, which gives us an expression for the vector of regression coefficients corresponding to any subset of the variables X_1 to X_K.

The OVB formula

Consider two regressions, "short" and "long":

$$\begin{aligned} Y_i &= \alpha + \beta X_i + e_i && \text{(short)} \\ Y_i &= \tau + \rho X_i + A_i'\gamma + \epsilon_i && \text{(long)} \end{aligned}$$

where X_i is a scalar and A_i can be a vector $(A_{1i}, \dots A_{Ki})'$. Note, neither the short or long regression is an assumption about the "true model". Rather, they are simply two regressions that we can run. The short regression can be thought of as defining an α and β such that the residual $e_i = Y_i - \alpha - \beta X_i$ is mean zero and uncorrelated with X_i. Similarly, the long regression can be thought of as defining τ, ρ, γ such that ϵ_i is mean zero and uncorrelated with X_i and A_i. In practice, we are often interested in cases where the long regression is likely to have a causal interpretation (see below), but that will not be important for deriving the omitted variables bias (OVB) formula.

The OVB formula relates ρ to β, i.e. the coefficients on X_i in the long and short regression, respectively. Start by recalling that $\beta = \frac{C(Y_i, X_i)}{V(X_i)}$. We can relate this to ρ by substituting the long regression equation for Y_i into the covariance operator:

$$\beta = \frac{\rho C(X_i, X_i) + C(A_i'\gamma, X_i) + C(\epsilon_i, X_i)}{V(X_i)} = \rho + \frac{C(\gamma' A_i, X_i)}{V(X_i)}$$

where we've used that $C(\epsilon_i, X_i) = 0$ and that $A_i'\gamma = \gamma' A_i$. Reversing the dot product is a convenient rearrangement because it sets us up to use the linearity of the covariance operator to notice that

$$\frac{C(\gamma' A_i, X_i)}{V(X_i)} = \sum_{j=1^K} \gamma_i \frac{C(A_{ji}, X_i)}{V(X_i)} = \gamma' \underbrace{\begin{pmatrix} C(A_{1i}, X_i)/V(X_i) \\ C(A_{2i}, X_i)/V(X_i) \\ \vdots \\ C(A_{Ki}, X_i)/V(X_i) \end{pmatrix}}_{:=\delta_{Ax}}$$

where we define the vector δ_{Ax} to be the vector of coefficients from K different simple bivariate regressions, in each of which we regress a component of A_i on X_i.

Thus, we have the OVB formula that $\beta = \rho + \gamma' \delta_{Ax}$.[1]

At risk of being redundant, we emphasize that the OVB formula is just algebra: it doesn't depend upon any model or assumptions about causality. However, the OVB comes up a lot when people are thinking about endogeneity. In a case where X_i is a treatment variable and we think that the CIA holds conditional on A_i, then under a few extra assumptions we can interpret the long regression as a causal regression in the sense described last week (see notes), where ρ is the treatment effect. With X_i binary, β is a simple difference in means between treatment and control groups, and the OVB formula tells us the difference between this "naive" estimate of the treatment effect and the true causal effect.

Saturated and "saturated in groups" regression

Suppose X_i is a random variable that takes on a finite set of possible values $x_1 \dots x_G$, such as $X_i \in \{$"super selective", "fairly selective", "not selective"$\}$ in the private school example (in this example $G = 3$). More generally, we can think of X_i as indicating which "group" an observation is in, where a group might be defined according to several variables, e.g.

$$X_i \in \{\underbrace{\text{“male and mother graduated high school”}}_{x_1}, \underbrace{\text{“male and mother didn't graduate high school”}}_{x_2},$$
$$\underbrace{\text{“female and mother graduated high school”}}_{x_3}, \underbrace{\text{“female and mother did not graduate high school”}}_{x_4}\}$$

In this example, $G = 4$, and the four values of X_i are map one-to-one with the four possible combinations of two binary variables: one for gender, and one for mother's high school completion (we could call these X_{1i} and X_{2i}).

Consider the CEF of a Y_i on X_i. Since X_i takes one on of G values, this CEF takes on at most G different values as well. Let $m_j := E[Y_i|X_i = x_j]$ for each $j = 1 \dots G$. Further define $d_{ij} := \mathbb{1}(X_i = x_j)$ to be an indicator variable that unit i belongs to "group j" (i.e., $X_i = x_j$). Then, for any unit i, we can write the CEF evaluated at that unit's X_i as:

$$E[Y_i|X_i] = \sum_{j=1}^{G} m_j d_{ij}$$

Note that conditioning on X_i is equivalent to conditioning on the values of all the variables d_{i1}, d_{i2}, and so on to d_{iG} (why is conditioning on only a subset of these not the

[1] What if we also had some set of variables W_i that were in both the short and the long regression? In this case we can use the regression anatomy formula show that the relation between β and ρ would be $\beta = \rho + \gamma' \delta^x_{Axw}$, where δ^x_{Axw} is the vector of coefficients on X_i in a set of regressions of A_{ji} on X_i and W_i, for $j = 1 \dots K$ (verifying this is a good exercise).

same?). Thus, the CEF of Y_i on the set of variables $d_{i1} \dots d_{iG}$ is linear in all of the d_{ij}. Cool!

The regression

$$Y_i = \beta_1 d_{i1} + \cdots + \beta_G d_{iG} + \epsilon_i$$

is called a *saturated model* (why is there no constant in this regression?). Since the CEF of Y_i on X_i is linear, the regression coefficients β_j are equal to the values of the CEF: $\beta_1 = m_1$, $\beta_2 = m_2$, $\dots \beta_G = m_G$. While regression always reveals the best linear approximation to the CEF, recall that when the CEF is linear it recovers it exactly.

Definition (from MHE 3.1.4): A *saturated model* is a regression in which the explanatory variables are discrete and the model includes a parameter for every possible combination of values of the explanatory variables.

Now suppose we have our X_i from before as well as a separate variable D_i, which we take to be a binary variable for simplicity. We might call the regression

$$Y_i = \rho D_i + \beta_1 d_{i1} + \cdots + \beta_G d_{iG} + \epsilon_i \tag{5}$$

"saturated in groups". We have a dummy variable for each of the possible groups, as well as D_i in the regression. But, we have not added interactions between the values of D_i and the groups. While a saturated linear regression will always recover the CEF exactly, the same is not true when we're only saturated in groups. That's because the conditional expectation of Y_i on both variables: $E[Y_i|D_i, X_i]$, is not guaranteed to be *additively separable* between D_i and X_i.[2]

Nevertheless, the saturated-in-groups regression will prove to be useful, because it ends up being equivalent to a certain *matching* estimator (well, the population version of it). To see this, we'll apply the regression anatomy formula to get the coefficient ρ from regression (5):

$$\rho = \frac{C(Y_i, \tilde{D}_i)}{V(\tilde{D}_i)}$$

where $\tilde{D}_i$ is the residual from a regression of D_i on $d_{1i} \dots d_{Gi}$. Since *this* regression is in-fact saturated, we know that it captures the CEF of D_i on X_i, and hence $\tilde{D}_i = D_i - E[D_i|X_i]$. This property is useful because then

$$\rho = \frac{C(Y_i, (D_i - E[D_i|X_i]))}{V(D_i - E[D_i|X_i])}$$

Warning: this is about to get messy. Let $\sigma_D^2(x)$ denote the conditional variance of D_i on X_i, i.e. $\sigma_D^2(x) := E\left[(D_i - E[D_i|X_i])^2 | \, X_i = x\right]$. Noting that $\tilde{D}_i$ is mean zero and

[2] If it is additively separable, then $E[Y_i|D_i, X_i] = \rho D_i + \beta_1 d_{i1} + \cdots + \beta_G d_{iG}$.

applying LIE to the above, we have:

$$\rho = \frac{E[Y_i(D_i - E[D_i|X_i])]}{E[(D_i - E[D_i|X_i])^2]} = \frac{E\left\{E[Y_i(D_i - E[D_i|X_i])|X_i]\right\}}{E[\sigma_D^2(X_i)]}$$

Consider the numerator. Applying the LIE over D_i:

$$\begin{aligned} E[Y_i(D_i - E[D_i|X_i])|X_i] &= \sum_{d\in\{0,1\}} P(D_i = d|X_i)E[Y_i(D_i - E[D_i|X_i])|D_i = d, X_i] \\ &= P(D_i = 1|X_i)E[Y_i|D_i = 1, X_i](1 - P(D_i = 1|X_i)) \\ &\qquad - P(D_i = 0|X_i)E[Y_i|D_i = 0, X_i]P(D_i = 1|X_i) \end{aligned}$$

To simplify notation, let's let $p(X_i) = P(D_i = 1|X_i) = E[D_i|X_i]$. Note that $\sigma_D^2(X_i) = p(X_i)(1 - p(X_i)) = P(D_i = 1|X_i)P(D_i = 0|X_i)$. Thus:

$$E[Y_i(D_i - E[D_i|X_i])|X_i] = \sigma_D^2(X_i)\left\{E[Y_i|D_i = 1, X_i] - E[Y_i|D_i = 0, X_i]\right\}$$

Thus, we've shown that

$$\rho = \frac{E[\delta(X_i)\sigma_D^2(X_i)]}{E[\sigma_D^2(X_i)]} \tag{6}$$

where $\delta(X_i) := E[Y_i|D_i = 1, X_i] - E[Y_i|D_i = 0, X_i]$. We can think of $\delta(x)$ as a function that "matches" treated ($D_i = 1$) and control ($D_i = 0$) units with the same value of $X_i = x$, and then conducts a simple comparison between the treated and control means for that x (note: since we haven't made the CIA, there's no reason yet to expect a causal interpretation to arise from the matching).

In fact, Expression (6) shows that ρ is a *weighted average* of $\delta(x_j)$ across the G possible values of X_i, i.e. $\rho = \sum_{j=1}^{G} w_j\delta(x_j)$ where the weights are $w_j = \frac{\sigma_D^2(x_j)P(X_i=x_j)}{E[\sigma_D^2(x_j)]}$. These weights are positive and sum to one. Thus, if we did make the CIA conditional on X_i, then ρ is a certain kind of average treatment effect, where the average is weighted according to the conditional variance of treatment (and the probability of X_i).

An analog to Equation (6) can be derived for cases where the "treatment variable" is not binary, but the regression is still saturated in groups (or the CEF of the treatment on X_i is otherwise linear). For the case of an ordered (but discrete) treatment variable, like years of schooling, see Eq (34) in Angrist & Krueger 1999. For a continuous treatment, see then end of Section 3.3.1 in MHE.

For the interested. So far we've contrasted a fully saturated model with a "saturated in groups" model, where we have a parameter for every value of X_i but make the regression equation additively separable between D_i and the d_{ij}. What if we did include interactions between D_i and all of the dummy variables for the possible values of X_i?

$$Y_i = \beta_1 d_{i1} + \cdots + \beta_G d_{iG} + \rho_1 D_i d_{i1} + \cdots + \rho_G D_i d_{iG} + \epsilon_i \tag{7}$$

In this case, with the $2G$ (non-redundant) parameters we would have a saturated model, and would exactly recover the CEF of Y_i on D_i and X_i (why did we drop the D_i term from the saturated-in-groups model?). The coefficients ρ_j from Eq (7) a matching interpretation. Except, rather than yielding a weighted average of the difference in means between matched treatment and control for a specific cell X_i, we will get each one separately. To see this, notice that regression (7) is equivalent to running G separate regressions on the subsample $X_i = x_j$, one for each value of X_i:

$$Y_i = \beta_j + \rho_j D_i + \epsilon_i^j \qquad \text{(on the } X_i = x_j \text{ subsample)}$$

for $j = 1 \ldots G$. We know that the coefficient on a dummy variable D_i a simple bivariate regression simply gives the difference in means. Since this is all also conditional on $X_i = x_j$, this means that ρ_j yields:

$$E[Y_i|D_i = 1, X_i = x_j] - E[Y_i|D_i = 0, X_i = x_j] = \delta(x_j)$$

Len Goff
Microeconometrics GR6414

Notes on two stage least squares

Please let me know if you spot any typos, etc.!

Review of 2SLS

Consider a setting where we are interested in estimating the coefficient ρ on S_i in the following equation:

$$Y_i = X_i'\alpha + \rho S_i + \eta_i \tag{1}$$

where X_i are some observed variables assumed to be uncorrelated with η_i. However, S_i is possibly correlated with η_i. Thus, Equation (1) is *not* a regression (if it were, the "residual" η_i would also be uncorrelated with S_i, by definition). The variable S_i is sometimes referred to as an *endogenous* variable.

Often, we think of (1) as describing a causal relationship that we're interested in, but where estimating the causal effect ρ is complicated by selection bias. Eq (1) doesn't necessarily need to be causal though: we might also think of ρ in Eq. (1) as simply the regression coefficient on S_i from a longer regression that we aren't able to run. For example, we might have $\eta_i = A_i'\theta + \epsilon_i$ where $C(S_i, \eta_i) = C(S_i, A_i'\theta) \neq 0$, and we don't observe A_i. Another example of endogeneity comes from there being measurement error in S_i.

Suppose that we have an vector of *instrumental variables* Z_i that we believe are each uncorrelated with η_i. We can think of Z_i as a vector $(Z_{i1}, Z_{i2}, \dots Z_{iL})'$ where each $C(Z_{ij}, \eta_i) = 0$. The instruments will help us identify ρ, even without random assignment or the CIA! This comes from two important properties: the so-called *exclusion restriction* (that $C(Z_{ij}, \eta_i) = 0$ for all j), and a *relevance condition.* To appreciate the relevance condition, let's consider a regression of S_i on X_i and Z_i:

$$S_i = X_i'\pi_{10} + Z_i'\pi_{11} + \xi_{1i} \tag{2}$$

The relevance condition states that at least one component of π_{11} is non-zero - essentially that at least one of the instruments are correlated with S_i after controlling for X_i (we'll see why this matters below). We'll refer to Equation (2) as the "first stage". The name can often be thought of capturing the idea of a process in which units "select" into a value of the treatment variable S_i, and that this choice is influenced by the instruments Z_i. But this interpretation is not necessary here. Unlike Equation (1), Equation (2) *is* a regression: $C(X_i, \xi_{1i}) = 0$ and $C(Z_i, \xi_{1i}) = 0$ by definition.

The *two stage least squares* (2SLS) approach begins by estimating Eq (2), which is easy since it's a regression. Running this regression allows the analyst to estimate fitted values of S_i from the first-stage. This amounts to subtracting off ξ_{1i} from S_i, and yields the "part" of the endogenous variable that can be explained by the instruments, and the covariates X_i:

$$\hat{S}_i := X_i'\pi_{10} + Z_i'\pi_{11} \tag{3}$$

Note: in keeping with our usual notation in this class, Eq. (3) is a "population" version of the fitted regression values. In practice, the actual estimates will have hats on π_{10} and $\pi_1 1$ (see the end of these notes).

Now, consider performing a regression of Y_i on X_i and $\hat{S}_i$. Let's call the coefficient on $\hat{S}_i$ in this regression ρ_{2sls}. Recall from the regression anatomy formula that if we regress Y_i on X_i and $\hat{S}_i$, the coefficient on $\hat{S}_i$ will be $\rho_{2sls} = C(Y_i, \hat{S}_i^*)/V(\hat{S}_i^*)$, where we define $\hat{S}_i^*$ to be the residual from a regression of $\hat{S}_i$ on the X_i. What happens if all entries of the vector π_{11} were equal to zero? In that case, we wouldn't be able to regression Y_i on both X_i and $\hat{S}_i$, since $\hat{S}_i$ would then be perfectly collinear with X_i. This is why we need the relevance condition that π_{11} has at least one non-zero component.

Now we show that 2SLS recovers ρ from Eq (1), i.e. $\rho_{2sls} = \rho$. If we substitute Eq (1) into our formula for ρ_{2sls}, we get:

$$\rho_{2sls} = \frac{C(\cancel{X_i'\alpha} + \rho S_i + \cancel{\eta_i}, \hat{S}_i^*)}{V(\hat{S}_i^*)} = \rho\frac{C(S_i, \hat{S}_i^*)}{V(\hat{S}_i^*)} = \rho$$

The steps we've made use of are the following:

- To eliminate the first term in the covariance, we note that since $\hat{S}_i^*$ is defined as a residual from a regression that includes X_i, it is uncorrelated with each component of X_i (and hence the sum $X_i'\alpha$).

- We can eliminate the η term in the covariance by the assumption that the components of Z_i and X_i are all are uncorrelated with η_i. Since the residual $\hat{S}_i^*$ can be written as $\hat{S}_i^* = \hat{S}_i - X_i'\gamma$ for some γ, and $\hat{S}_i = X_i'\pi_{10} + Z_i'\pi_{11}$ by Eq. (3), it follows that $\hat{S}_i^*$ is a linear combination of the X_i and the Z_i: $\hat{S}_i^* = X_i'(\pi_{10} - \gamma) + Z_i'\pi_{11}$. Thus, by the linearity of the covariance operator, it is uncorrelated with η_i.

- To achieve the final equality, we notice that the quantity $C(S_i, \hat{S}_i^*) = C(\underbrace{X_i'\gamma + \hat{S}_i^*}_{\hat{S}_i} + \xi_{1i}, \hat{S}_i^*)$ for some γ (the coefficient in a regression of $\hat{S}_i$ on X_i). Since $\hat{S}_i^*$ is the residual from a regression that includes X_i, it is uncorrelated with each component of X_i (and hence the sum $X_i'\gamma$). Also, the ξ_{1i} term is zero since $C(\xi_{1i}, \hat{S}_i^*) = C(\xi_{1i}, \hat{S}_i - X_i'\gamma) = 0$, and ξ_{1i} is uncorrelated both with X_i and $\hat{S}_i$. So, $C(S_i, \hat{S}_i^*) = V(\hat{S}_i^*)$.

So, we've shown mathematically that the 2SLS recovers ρ from Eq (1), despite the problem that $C(S_i, \eta_i) \neq 0$! But why does this work, intuitively? Consider the decomposition of S_i offered by the first stage regression (2). Notice that because of the assumption that X_i and Z_i are uncorrelated with η_i, we know that all of the correlation between S_i and η_i must come from the error term ξ_{1i}. What 2SLS does is use the first stage to "purge" ξ_{1i}

from S_i, getting rid of the part of it that is potentially correlated with η_i. The replacement of S_i by $\hat{S}_i$ allows us to treat Eq (1) as a regression, since $C(\hat{S}_i, \eta_i) = 0$.

Indirect least squares

There is another approach to using our instruments Z_i to identify Z_i, that goes by the name *indirect least squares* (ILS). The idea behind ILS is to run a third regression, which is like the first stage but with Y_i on the left hand side instead of S_i:

$$Y_i = X_i'\pi_{20} + Z_i'\pi_{21} + \xi_{2i} \tag{4}$$

where $C(X_i, \xi_{2i}) = 0$ and $C(Z_i, \xi_{2i}) = 0$ by definition.

Aside from being bored, why would we care to run regression (4)? It turns out that the coefficients on Z_i will be proportional to ρ, and that the proportionality can be determined from the first stage regression. Thus, we "indirectly" determine the value of ρ by running two regressions, and combining the estimates from each in a certain way. Equation (4) is referred to as the *reduced form*. The idea behind this name is that the instruments Z_i only effect Y_i through S_i. Thus, the coefficient π_{21} offers observable evidence of the "structural" (i.e. causal) effect ρ, while not directly measuring it (we'll need the to estimate the reduced form as well to do that).

When Z_i has L components $(Z_{i1} \dots Z_{iL})'$, we write $\pi_{21} = (\pi_{211} \dots \pi_{21L})'$, one coefficient for each component of the vector. The idea behind ILS is the observation that for any component $j \in 1 \dots L$: $\pi_{21j} = \rho\pi_{11j}$ where π_{11j} is the coefficient on Z_{ij} in the first stage regression. Thus, if we perform both the first stage and the reduced form regressions, we can solve for ρ. In the box below, we show why this works.

Rearranging the ILS expression, we have that $\rho = \frac{\pi_{21j}}{\pi_{11j}}$, the ratio of a coefficient from the reduced form regression to a coefficient from the first stage regression. However, notice that the RHS of this expression is indexed by j, while the LHS does not. This means that if $\pi_{11j} \neq 0$ for multiple values of j, then we have more than one expression for the same quantity, ρ!. This property is called *overidentification*, and it arises because we have L instrumental variables but only one endogenous variable S_i.

In practice, when we estimate ρ in a finite sample, we'd get a different numerical estimate for each j. This will always occur due to just statistical noise, but it can also be evidence that one or more of the instruments does not actually satisfy the exclusion restriction. This is the basis of so-called *over-identification testing*. Define $\rho_{ILS,j} = \frac{\pi_{21j}}{\pi_{11j}}$ for each j. Over-identification testing essentially asks the question of whether the different values of $\hat{\rho}_{ILS,j}$ across j can be accounted for by simple sampling uncertainty, or if they are evidence that their population counterparts $\rho_{ILS,j}$ are different, which would mean that at least one of our instrumental variables is invalid.

However, when we have just one instrumental variable ($L = 1$), the ILS procedure produces just one estimate of ρ. This estimate will be numerically equivalent to the one given by the 2SLS procedure. I'll show an example of this in Stata during recitation.

Why ILS works: Consider the coefficient on the j^{th} component of Z_i in Eq (4). By the regression anatomy formula, this is $\pi_{21j} = \frac{C(Y_{ij}, \tilde{Z}_{ij})}{V(\tilde{Z}_{ij})}$, where $\tilde{Z}_{ij}$ is the residual from a regression of Z_{ij} on X_i and the other components of Z_i. Now, let's substitute the causal "outcome equation" (1) for Y_i into this expression:

$$\pi_{21j} = \frac{C(\cancel{X_i'\alpha} + \rho S_i + \cancel{\eta_i}, \tilde{Z}_{ij})}{V(\tilde{Z}_{ij})} = \rho \frac{C(S_i, \tilde{Z}_{ij})}{V(\tilde{Z}_{ij})} = \rho \pi_{11j}$$

where π_{11j} is the j^{th} component of π_{11}.

- To eliminate the first term in the covariance, we use the fact that since $\tilde{Z}_{ij}$ is defined as a residual from a regression that includes X_i, it is uncorrelated with each component of X_i (and hence the sum $X_i'\alpha$).
- Similar to with 2SLS, we can eliminate the η term in the covariance by the assumption that all of the components of Z_i and X_i are uncorrelated with η_i. Since the residual $\tilde{Z}_{ij}$ can be written as $\tilde{Z}_{ij} = Z_{ij} - X_i'\lambda - \beta_1 Z_{i1} - \beta_2 Z_{i2} - \dots \beta_{j-1} Z_{i,j-1} - \beta_{j+1} Z_{i,j+1} - \dots \beta_L Z_{iL}$ for some λ and β, it is a linear combination of the Z_i's, which is also uncorrelated with η_i.
- To achieve the final equality, we notice that the quantity $C(S_i, \tilde{Z}_{ij})/V(\tilde{Z}_{ij})$ is exactly the regression anatomy formula for π_{22j} in the first stage regression (2). Notice that it's important that the first-stage and the reduced form regressions have exactly the same form (linear regressions with the same regressors on the right side). Otherwise this equality wouldn't be true.

2SLS as a simple IV

Recall that when there is just one instrument Z_i, and no covariates in the regression, the IV formula has a simple form (Eq. 2 in the slides):

$$\rho = \frac{C(Y_i, Z_i)}{C(S_i, Z_i)}$$

Let's call such a setting a "simple IV" setup. It turns out that we can think of the more complicated 2SLS as such a simple IV, but where the instrument is a certain function of the many Z_i and X_i.

Recall that by regression anatomy $\rho_{2sls} = C(Y_i, \hat{S}_i^*)/V(\hat{S}_i^*)$, where $\hat{S}_i^*$ is the residual from a regression of $\hat{S}_i$ on the X_i. We also showed in the proof for 2SLS that $C(S_i, \hat{S}_i^*) = V(\hat{S}_i^*)$. Thus, we could write

$$\rho_{2sls} = \frac{C(Y_i, \hat{S}_i^*)}{C(S_i, \hat{S}_i^*)}$$

which is the simple IV formula if our "instrument" is $\hat{S}_i^*$. We showed that this $\hat{S}_i^*$ will be uncorrelated with η_i, so it satisfies the exclusion restriction. And it will be correlated with S_i so long as at least one component of π_{11} is nonzero, satisfying the relevance condition.

This way of thinking about 2SLS is useful because it tells us that effectively what 2SLS does is aggregate all of our instruments $Z_{i1} \ldots Z_{iL}$ into one single instrument. It turns out that it does so in a nice way (it forms the statistically optimal linear combination of instruments under a homoskedasticity assumption).

2SLS as averaging simple IV estimates

When we have many instruments, we could have used each one separately to perform a just-identified IV regression, rather than combining them into one $2SLS$ estimation with all of the instruments. Recall that *just-identified* refers to a setting in which the number of instruments is equal to the number of endogenous variables. When this number is equal to one, as it would be with S_i and a single Z_{ij}, our estimator would be

$$\hat{\rho}_{IV,j} = \frac{\hat{C}(Y_i, Z_{ij})}{\hat{C}(S_i, Z_{ij})}$$

where $\hat{C}$ indicates the sample covariance. This is what we've called a "simple IV" setting in the last section.

It turns out that the 2SLS estimator $\hat{\rho}_{2sls}$ yields a weighted average of the $\hat{\rho}_{IV,j}$. For simplicity, we'll first consider the case where there are no covariates X_i (and then generalize for the brave). The 2SLS estimator is

$$\hat{\rho}_{2SLS} = \frac{\hat{C}(Y_i, \hat{S}_i)}{\hat{C}(S_i, \hat{S}_i)}$$

where $\hat{S}_i = Z_i'\hat{\pi}_{11} = \sum_{j=1}^{L} \hat{\pi}_{11j} Z_{ij}$. Note that we've been able to replace $\hat{S}_i^*$ in the general formula by $\hat{S}_i$ because there are no covariates X_i.[1]

By linearity of the covariance, the denominator in $\hat{\rho}_{2SLS}$ is $\hat{C}(S_i, \hat{S}_i) = \sum_{j=1}^{L} \hat{\pi}_{11j} \hat{C}(S_i, Z_{ij})$. Similarly, the numerator is

$$\hat{C}(Y_i, \hat{S}_i) = \sum_{j=1}^{L} \hat{\pi}_{11j} \hat{C}(Y_i, Z_{ij}) = \sum_{j=1}^{L} \hat{\pi}_{11j} \hat{C}(S_i, Z_{ij}) \frac{\hat{C}(Y_i, Z_{ij})}{\hat{C}(S_i, Z_{ij})}$$

where we've multiplied and divided by $\hat{C}(S_i, Z_{ij})$. Now notice that $\frac{\hat{C}(Y_i, Z_{ij})}{\hat{C}(S_i, Z_{ij})}$ is exactly $\hat{\rho}_{IV,j}$. If we define $w_j = \frac{\hat{\pi}_{11j}\hat{C}(S_i, Z_{ij})}{\sum_{j=1}^{L} \hat{\pi}_{11j}\hat{C}(S_i, Z_{ij})}$, then

$$\hat{\rho}_{2SLS} = \sum_{j=1}^{L} w_j \cdot \hat{\rho}_{IV,j}$$

It can be readily verified that these weights sum to one. To guarantee the weights to be positive, we need the additional assumption that $\hat{C}(S_i, Z_{ij})$ have the same sign as

[1] Aside from a constant, which has no effect on the covariance

$\hat{C}(S_i, \tilde{Z}_{ij})$ for each j, where $\tilde{Z}_{ij}$ is the residual from a regression of Z_j on all of the other instruments (this gives the sign of $\hat{\pi}_{11j}$). In a two-instrument case, this says that conditioning on one of the instruments doesn't change the sign of the relationship between the other instrument and S_i.

With covariates: We can generalize this result to the case with covariates X_i by making use of matrix notation. Let $Y = (Y_1 \dots Y_n)'$ denote a vector of observations of Y, and similarly for S. Let Z_j denote a vector of observations of Z_{ij}, and let $Z = [Z_1 \dots Z_j]'$ indicate a matrix with the Z_j as rows. Define a matrix X analogously for the covariates, including a column of ones. For any matrix A, let $P_A = A(A'A)^{-1}A'$ be the matrix that projects onto the subspace spanned by the columns of A, and $M_A = I - P_A$ be the matrix that projects onto its orthogonal complement.

In our matrix notation: $\hat{\rho}_{2sls} = \frac{Y'M_X\hat{S}}{\hat{S}'M_X\hat{S}}$ where $\hat{S} = P_{[XZ]}S$. Note that

$$M_X P_{[XZ]}S = M_X(X\hat{\pi}_{10} + Z\hat{\pi}_{11}) = M_X Z\hat{\pi}_{11} = \sum_{j=1}^{L} \hat{\pi}_{11j} M_X Z_j$$

Thus: $\hat{\rho}_{2sls} = \frac{\sum_{j=1}^{L} \hat{\pi}_{11j}\tilde{Y}'Z_j}{S'M_XP_ZS}$, where $\tilde{Y} = M_XY$.

The IV estimator with only Z_j as an instrument, but including the covariates X_i, is:

$$\hat{\rho}_{IV,j} = \frac{Y'M_XZ_j}{S'M_XZ_j} = \frac{\tilde{Y}'Z_j}{S'M_XZ_j}$$

Thus:

$$\hat{\rho}_{2sls} = \sum_{j=1}^{L} \underbrace{\hat{\pi}_{11j}\frac{S'M_XZ_j}{S'M_XP_ZS}}_{:=w_j} \hat{\rho}_{IV,j}$$

Noticing that $S'M_XP_ZS = S'M_XZ\hat{\pi}_{11} = \sum_{j=1}^{L} \hat{\pi}_{11j}S'M_XZ_j$, it follows that the w_j sum to one.

Len Goff
Microeconometrics GR6414

IV in groups and 2SLS

Please let me know if you spot any typos, etc.!

Setup: getting to the grouped regression

Consider a model in which

$$Y_i = \alpha + \rho S_i + \eta_i \tag{1}$$

where ρ is the parameter of interest and $C(S_i, \eta_i) \neq 0$. Suppose we have a binary instrument Z_i that is uncorrelated with η_i. We know from our IV analysis thus far that we can identify ρ, since it is equal to $C(Y_i, Z_i)/C(S_i, Z_i)$. Since Z_i is Bernoulli (i.e. binary), this ratio of covariances can also be written in "Wald" form:

$$\rho = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[S_i|Z_i = 1] - \mathbb{E}[S_i|Z_i = 0]} \tag{2}$$

In the military draft example (Angrist 1990), Y_i is earnings, $S_i \in \{0, 1\}$ is veteran status, and Z_i is eligibility for the draft. In practice, Z_i is itself a function of an individual's *draft lottery number* R_i, by the relation $Z_i = \mathbb{1}(R_i \leq C)$, where C was a draft eligibility "ceiling" that varied by cohort year (e.g. 195 for men born in 1950, 125 for men born in 1951, etc.). We'll focus on a single cohort, so we can treat C as constant across all individuals.

We start by observing that Z_i inherits its validity as an instrument from the fact that the underlying draft lottery number R_i is random, and hence

$$R_i \perp \eta_i \implies C(f(R_i), \eta_i) = 0$$

for any function $f(\cdot)$. We think of R_i as random because it corresponds to a random ordering of birthdays. Another property we can derive from R_i being independent of η_i is that $\mathbb{E}[\eta_i|R_i = r] = 0$ for any possible draft number r.[1] This suggests that instead of using draft-eligibility to estimate Eq. (2), we can also identify ρ from the relation

$$\rho = \frac{\mathbb{E}[Y_i|R_i = r] - \mathbb{E}[Y_i|R_i = r']}{\mathbb{E}[S_i|R_i = r] - \mathbb{E}[S_i|R_i = r']} \tag{3}$$

for any two values r, r'. To see that Equation (3) holds, substitute (1) into the expression for the numerator: $\mathbb{E}[Y_i|R_i = r] - \mathbb{E}[Y_i|R_i = r']$.

Why bother with Eq (3), when we can just estimate (2)? The powerful thing about Eq. (3) is that it holds for every pair of values r and r'. Finding a clever way to take advantage of these multiple estimates for the same thing will improve the statistical efficiency of

[1] Really independence itself just implies that $\mathbb{E}[\eta_i|R_i = r] = \mathbb{E}[\eta_i]$ for all r, but we can take the unconditional expectation to be zero by absorbing it into α in Eq. (1).

our estimator. In practice, that allows us to have more confidence that our finite sample estimates are close to the population quantity of interest ρ. However, we note that while Eq. (3) will lead to many estimators for the same quantity ρ under the assumption of the model in Equation (1), this will generally not be the case for IV with heterogeneous treatment effects (essentially, a setting where ρ_i can vary by individual). We'll talk about this context in the next lecture.

Suppose that we estimated Equation (3) for every r, r' pair, leading to $J(J-1)$ estimators

$$\hat{\rho}_{r,r'} = \frac{\hat{\mathbb{E}}[Y_i|R_i = r] - \hat{\mathbb{E}}[Y_i|R_i = r']}{\hat{\mathbb{E}}[S_i|R_i = r] - \hat{\mathbb{E}}[S_i|R_i = r']} \tag{4}$$

where J is the number of values in the support of R (which we take to be discrete as in the draft example), and $\hat{\mathbb{E}}$ is the empirical expectation function $\frac{1}{N}\sum_{j=1}^{N}$. It's not immediately obvious how one would combine these various estimators to get a really good single one, e.g. if one was to add them up what weights should they use?

An alternative is inspired by observing that Equation (1) implies that

$$\mathbb{E}[Y_i|R_i = r] = \alpha + \rho\mathbb{E}[S_i|R_i = r] \tag{5}$$

Thus, if we form empirical estimates of the conditional expectations: $\bar{y}_r = \frac{1}{N_r}\sum_{i:R_i=r} Y_i$, and $\bar{s}_r = \frac{1}{N_r}\sum_{i:R_i=r} S_i$ where $N_r := |\{i : R_i = r\}|$, then we could estimate ρ from a regression on groups:

$$\bar{y}_r = \alpha + \rho\bar{s}_r + \bar{\eta}_r \tag{6}$$

where we have one "observation" for each of the J groups, and $\bar{\eta}_r = \frac{1}{N_r}\sum_{i:R_i=r} \eta_i$.[2] Each group corresponds to a value r of the instrumental variable R_i. Note that in Angrist 1990, a value of R_i in the data really corresponds to five different draft lottery numbers, which have been aggregated by bins of five consecutive numbers (for privacy reasons). Thus, the number J of groups in our analysis is only 70, rather than 365. However, this doesn't change the structure of our discussion at all: the following holds with this pre-aggregation just as it would with a unique group for every birthdate.

The grouped regression is 2SLS

In Equation (6), it's apparent that each of our "observations" (a value of r) will correspond with a different number N_r of actual observations from our data: the N_r folks who had birthdays putting them in a single group r. Thus it's natural that from a statistical perspective, we should let a value of r for which N_r is high count more in the regression, then a value of r for corresponding to only a few individuals.

This leads us to so-called generalized least squares (GLS), which is a generalization of the OLS estimator: $\hat{\rho}_{OLS} = \frac{\frac{1}{J}\sum_{r=1}^{J}\bar{y}_r(\bar{s}_r - \bar{s})}{\frac{1}{J}\sum_{r=1}^{J}\bar{s}_r(\bar{s}_r - \bar{s})}$. When observations (in our regression, we

[2] Regression (6) does have measurement error in the dependent variable because $\bar{s}_r \neq \mathbb{E}[S_i|R_i = r]$ in a finite sample, but the resulting bias will go away asymptotically, so the estimator will still be consistent.

have one "observation" for each of the J groups) are independent of one another (as they are in our case), GLS simply adds a weight w_r to each observation:

$$\hat{\rho}_{GLS} = \frac{\sum_{r=1}^{J} w_r \bar{y}_r (\bar{s}_r - \bar{s}_w)}{\sum_{r=1}^{J} w_r \bar{s}_r (\bar{s}_r - \bar{s}_w)}$$

where $\bar{s}_w = \sum_{r=1}^{J} w_r \bar{s}_r$.

It turns out that GLS is better than OLS (in a statistical sense) whenever there is heteroskedasticity in the data. In particular, one can show that when the errors $\bar{\eta}_r$ are uncorrelated across r, GLS with weights w_r (often referred to as *weighted least squares*) is the best linear unbiased estimator of ρ (in the sense of minimizing the asymptotic variance of the estimator), when $w_r \propto \mathbb{V}(\bar{\eta}_r|\bar{s}_r)$ and the weights are normalized to sum to one. This expression means that we want to make the weights proportional to the inverse of the variance of the error term in Regression (6) for a given group r. Since $\bar{\eta}_r$ is simply an average of N_r η_i's, it's easy to work out that if the η_i's all have the same variance σ_η^2, then the weights should be

$$w_r = \frac{N_r/\sigma_\eta}{\sum_{r=1}^{J} N_r/\sigma_\eta} = N_r/N$$

One of the great things about assuming homoskedasticity is that we don't even need to know the value of σ_η to implement GLS, since it simply cancels out in the weights. We will maintain this assumption that there exists some σ_η such that $\mathbb{V}(\bar{\eta}_r|\bar{s}_r) = \sigma_\eta^2$ for all r.

Note also that with $w_r = N_r/N$:

$$\sum_{r=1}^{J} w_r \bar{y}_r = \frac{1}{N}\sum_{r=1}^{J} \cancel{N_r} \frac{1}{\cancel{N_r}} \sum_{i:R_i=r} Y_i = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{y}$$

Thus:

$$\hat{\rho}_{GLS} = \frac{\sum_{r=1}^{J} w_r \bar{y}_r \bar{s}_r - \bar{y}\bar{s}_w}{\sum_{r=1}^{J} w_r \bar{s}_r (\bar{s}_r - \bar{s}_w)} = \frac{\sum_{r=1}^{J} w_r \bar{s}_r (\bar{y}_r - \bar{y})}{\sum_{r=1}^{J} w_r \bar{s}_r (\bar{s}_r - \bar{s}_w)} = \frac{\sum_{r=1}^{J} N_r \bar{s}_r (\bar{y}_r - \bar{y})}{\sum_{r=1}^{J} N_r \bar{s}_r (\bar{s}_r - \bar{s})} \tag{7}$$

where in the last step we've used that with $w_r = N_r/N$, $\bar{s}_w = \bar{s}$.

The final form of Eq (7) will be convenient for what follows, in which we will see that $\hat{\rho}_{GLS}$ is equivalent to a 2SLS estimator where we use a group indicator $\mathbb{1}(R_i = r)$ for each value r of R_i as a separate instrument to estimate ρ. Nifty! To see this, let $\hat{S}_i$ indicate the population fitted value from a regression of S_i on a full set of group indicators, the *first stage*:

$$\hat{S}_i = \pi_{10} + \pi_{11}\mathbb{1}(R_i = 1) + \cdots + \pi_{1J}\mathbb{1}(R_i = J)$$

To indicate the sample estimate, we'll put a hat on the π's and a second hat on $\hat{S}_r$ (why not wear two hats?). This regression is saturated, we know that it will recover the CEF: $\pi_{1r} = \mathbb{E}[S_i|R_i = r]$. The estimated version $\hat{\hat{S}}_i$ will simply be the sample mean $\bar{s}_r$ of S_i

within the group (i.e. value of R_i) to which observation i belongs. We can write this as: $\hat{\hat{S}}_i = \sum_{r=1}^{J} \mathbb{1}(R_i = r)\bar{s}_r$. Using this:

$$\begin{aligned}
\hat{\rho}_{2SLS} &= \frac{\hat{C}(Y_i, \hat{\hat{S}}_i)}{\hat{V}(\hat{\hat{S}}_i)} = \sum_{r=1}^{J} \bar{s}_r \frac{\hat{C}(Y_i, \mathbb{1}(R_i = r))}{\hat{V}(\hat{\hat{S}}_i)} \\
&= \frac{1}{\hat{V}(\hat{\hat{S}}_i)} \sum_{r=1}^{J} \bar{s}_r \left\{ \hat{\mathbb{E}}[Y_r \mathbb{1}(R_i = r))] - \hat{\mathbb{E}}[Y_i]\hat{\mathbb{E}}[\mathbb{1}(R_i = r)] \right\} \\
&= \frac{1}{\hat{V}(\hat{\hat{S}}_i)} \sum_{r=1}^{J} \hat{P}(R_i = r)\bar{s}_r \left(\hat{\mathbb{E}}[Y_i | R_i = r)] - \hat{\mathbb{E}}[Y_i] \right) \\
&= \frac{1}{N\hat{V}(\hat{\hat{S}}_i)} \sum_{r=1}^{J} N_r \bar{s}_r \left(\bar{y}_r - \bar{y} \right)
\end{aligned}$$

where $\bar{y}$ is the unconditional mean of Y_i in the sample. Comparing with Eq (7), we see that to finish demonstrating that $\rho_{2SLS} = \rho_{GLS}$, we only need to show that

$$N\hat{V}(\hat{\hat{S}}_i) = \sum_{r=1}^{J} N_r \bar{s}_r (\bar{s}_r - \bar{s})$$

To see this, substitute $\hat{\hat{S}}_i = \sum_{r=1}^{J} \mathbb{1}(R_i = r)\bar{s}_r$ into the empirical variance:

$$\begin{aligned}
\hat{V}(\hat{\hat{S}}_i) &= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{r=1}^{J} \mathbb{1}(R_i = r)\bar{s}_r \right)^2 - \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{r=1}^{J} \mathbb{1}(R_i = r)\bar{s}_r \right)^2 \\
&= \frac{1}{N} \sum_{r=1}^{J} \sum_{i=1}^{N} \mathbb{1}(R_i = r)\bar{s}_r^2 - \left(\frac{1}{N} \sum_{r=1}^{J} \sum_{i=1}^{N} \mathbb{1}(R_i = r)\bar{s}_r \right)^2 \\
&= \frac{1}{N} \sum_{r=1}^{J} N_r \bar{s}_r^2 - \left(\frac{1}{N} \sum_{r=1}^{J} N_r \bar{s}_r \right)^2 \\
&= \frac{1}{N} \sum_{r=1}^{J} N_r \bar{s}_r^2 - \bar{s}^2 = \frac{1}{N} \sum_{r=1}^{J} N_r \bar{s}_r (\bar{s}_r - \bar{s})
\end{aligned}$$

Len Goff
Microeconometrics GR6414

Notes on IV with heterogeneous treatment effects

Please let me know if you spot any typos, etc.!

The LATE theorem

We'll focus on the case of a single binary treatment and a binary instrument. Recall that the assumptions of the LATE model are that for all units i (i.e. with probability one):

1. **Independence:** $(Y_i(d,z), D_i(z)) \perp Z_i$ for all d, z
2. **Exclusion:** $Y_i(d,z) = Y_i(d)$ for all d, z
3. **Monotonicity:** $D_i(1) \geq D_i(0)$

where $Y_i(d,z)$ indicates the potential outcome where treatment status is $d \in \{0,1\}$ *and* the instrument is equal to $z \in \{0,1\}$, while $D_i(z)$ indicates a potential treatment: what the value of treatment for unit i would be when the instrument takes value $z \in \{0,1\}$.

Independence is our assumption that the instrument is as good as randomly assigned, and is thus statistically independent of heterogeneity across individuals' potential outcomes and potential treatments. Exclusion states that $Y_i(d,z)$ doesn't depend on z and is our assumption that Z only effects Y *through* D. Thus, the potential outcome $Y_i(d,z)$ doesn't change with z if d is held fixed. We could write Assumptions 1 and 2 together as

$$(Y_i(0), Y_i(1), D_i(0), D_i(1)) \perp Z_i$$

Monotonicity states that the causal effect of the instrument on treatment status is to move all units weakly in the *same direction.* That is, we can't have some unit i for whom $D_i(0) = 0$ and $D_i(1) = 1$, and some other unit j for whom $D_j(0) = 1$ and $D_j(1) = 0$. The direction $\geq$ of the weak inequality in Assumption 3. is arbitrary, if the instrument taking a value of one were to move all units *out* of treatment or not at all (i.e. a $\leq$ instead of $\geq$), we could simply redefine the instrument by swapping the labels of $z = 0$ and $z = 1$.

Given the normalization that $D_i(1) \geq D_i(0)$ for all i (rather than $D_i(1) \leq D_i(0)$), the monotonicity assumption implies that we can separate all units i into three mutually exclusive categories:

Name	**Meaning**
"never-takers"	$D_i(0) = 0$ & $D_i(1) = 0$
"always-takers"	$D_i(0) = 1$ & $D_i(1) = 1$
"compliers"	$D_i(0) = 0$ & $D_i(1) = 1$
~~"defiers"~~	~~$D_i(0) = 1$ & $D_i(1) = 0$~~

where we've crossed out defiers, because by monotonicity we assume they do not exist. The names given in this table give intuitive meaning to the three groups: never-takers would not receive treatment regardless of the value of the instrument, always-takers would receive treatment regardless of the instrument, and compliers take the treatment if and only if the instrument takes a value of one.

Now consider the IV estimand ρ_{IV}:

$$\rho_{IV} = \frac{C(Y,Z)}{C(S,Z)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

(the equality can be shown with some simple algebra).

Consider the term $E[Y_i|Z_i = 1]$. Note that $D_i(1)$ and $D_i(0)$ are each random variables, so by the law of iterated expectations:

$$\begin{aligned} E[Y_i|Z_i = 1] &= p_{nevertaker} \cdot E[Y_i|Z_i = 1, D_i(0) = 0, D_i(1) = 0] \\ &+ p_{alwaystaker} \cdot E[Y_i|Z_i = 1, D_i(0) = 1, D_i(1) = 1] \\ &+ p_{complier} \cdot E[Y_i|Z_i = 1, D_i(0) = 0, D_i(1) = 1] \end{aligned}$$

where by independence assumption $P(D_i(0) = 0, D_i(1) = 0|Z_i = z)$ doesn't depend on z and let $p_{nevertaker}$ be the unconditional probability $P(D_i(0) = 0, D_i(1) = 0)$ (and similarly for always-takers and compliers). Now, having conditioned on Z_i as well as a unit's potential treatments, we know their realized treatment $D_i = D_i(Z_i)$, and hence which potential outcome we are observing in Y_i:

$$\begin{aligned} E[Y_i|Z_i = 1] &= p_{nevertaker} \cdot E[Y_i(0)|Z_i = 1, D_i(0) = 0, D_i(1) = 0] \\ &+ p_{alwaystaker} \cdot E[Y_i(1)|Z_i = 1, D_i(0) = 1, D_i(1) = 1] \\ &+ p_{complier} \cdot E[Y_i(1)|Z_i = 1, D_i(0) = 0, D_i(1) = 1] \end{aligned}$$

Following the same logic for $E[Y_i|Z_i = 0]$

$$\begin{aligned} E[Y_i|Z_i = 0] &= p_{nevertaker} \cdot E[Y_i(0)|Z_i = 0, D_i(0) = 0, D_i(1) = 0] \\ &+ p_{alwaystaker} \cdot E[Y_i(1)|Z_i = 0, D_i(0) = 1, D_i(1) = 1] \\ &+ p_{complier} \cdot E[Y_i(0)|Z_i = 0, D_i(0) = 0, D_i(1) = 1] \end{aligned}$$

Now, the great thing about having replaced the Y_i's by the corresponding potential outcomes is that the potential outcomes themselves (as well as the potential treatments) are independent of the instrument Z_i, so we can drop the conditioning on Z_i.[1] Abbreviating

[1] Consider a term $E[Y_i(d)|Z_i = z, D_i(0) = d, D_i(1) = d^*]$ with specific values of d, z and z^*. Really what we're using is the joint independence condition $(Y_i(d), D_i(0), D_i(1)) \perp Z_i$. Assume for simplicity that $Y_i(d)$ has discrete support. Then

$$E[Y_i(d)|Z_i = z, D_i(0) = d, D_i(1) = d^*] = \sum_y yP(Y_i(d) = y|Z_i = z, D_i(0) = d, D_i(1) = d^*) = \sum_y y\frac{P(Y_i(d) = y, Z_i = z, D_i(0) = d, D_i(1) = d^*)}{P(Z_i = z, D_i(0) = d, D_i(1) = d^*)}$$

By the independence condition this is equal to

$$\sum_y y\frac{\cancel{P(Z_i = z)}P(Y_i(d) = y, D_i(0) = d, D_i(1) = d^*)}{\cancel{P(Z_i = z)}P(D_i(0) = d, D_i(1) = d^*)} = \sum_y yP(Y_i(d) = y|D_i(0) = d, D_i(1) = d^*) = E[Y_i(d)|D_i(0) = d, D_i(1) = d^*]$$

$p_{nevertaker}$, $p_{alwaystaker}$, and $p_{complier}$ as p_n, p_a and p_c respectively, we have that:

$$\begin{aligned} E[Y_i|Z_i = 1] &= p_n \cdot E[Y_i(0)|D_i(0) = 0, D_i(1) = 0] \\ &+ p_a \cdot E[Y_i(1)|D_i(0) = 1, D_i(1) = 1] \\ &+ p_c \cdot E[Y_i(1)|D_i(0) = 0, D_i(1) = 1] \end{aligned} \quad \text{and} \quad \begin{aligned} E[Y_i|Z_i = 0] &= p_n \cdot E[Y_i(0)|D_i(0) = 0, D_i(1) = 0] \\ &+ p_a \cdot E[Y_i(1)|D_i(0) = 1, D_i(1) = 1] \\ &+ p_c \cdot E[Y_i(0)|D_i(0) = 0, D_i(1) = 1] \end{aligned}$$

Thus, in the difference, the always-taker and never-taker terms cancel out, leaving:

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= p_{complier} \left(E[Y_i(1)|D_i(0) = 0, D_i(1) = 1] - E[Y_i(0)|D_i(0) = 0, D_i(1) = 1] \right) \\ &= p_{complier} E[Y_i(1) - Y_i(0)|D_i(0) = 0, D_i(1) = 1] \end{aligned}$$

Often the event $D_i(0) = 0, D_i(1) = 1$ is written as $D_i(1) > D_i(0)$, which is equivalent.

Now consider in the terms in the denominator of ρ_{IV}. By similar steps, $E[D_i|Z_i = 1]$ ends up just being the probability that a unit is an always-taker or a complier:

$$\begin{aligned} E[D_i|Z_i = 1] &= p_{nevertaker} E[D_i|Z_i = 1, D_i(0) = D_i(1) = 0] \\ &+ p_{alwaystaker} E[D_i|Z_i = 1, D_i(0) = D_i(1) = 1] \\ &+ p_{complier} E[D_i|Z_i = 1, D_i(1) > D_i(0)] \\ &= p_{nevertaker} * 0 + p_{alwaystaker} * 1 + p_{complier} * 1 \\ &= p_{alwaystaker} + p_{complier} \end{aligned}$$

And by similar steps:

$$E[D_i|Z_i = 0] = p_{alwaystaker}$$

Thus $E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = p_{complier}$ and we have finally the result that

$$\rho_{IV} = E[Y_i(1) - Y_i(0)|D_i(1) > D_i(0)]$$

Latent-index models

To get the LATE theorem, we have made assumptions about potential outcomes/treatments and their distributions, but we haven't committed to an explicit model of how these outcomes come about. An alternative/complimentary approach might characterize the selection process by constructing a "structural" model of who chooses treatment.

For instance, we might think that each unit i is a utility-maximizing agent who's utility is

$$u_i = \begin{cases} \gamma_0 + \gamma_1 Z_i & \text{if they receive treatment } (i.e.\ D_i = 1) \\ \nu_i & \text{if they don't } (i.e.\ D_i = 0) \end{cases}$$

Agents will choose treatment when it gives them higher utility, and so they will choose:

$$D_i = \mathbb{1}(\gamma_0 + \gamma_1 Z_i > \nu_i)$$

where we've assumed that ties go to non-treatment.

In this model, heterogeneity among D_i comes from agent's having different values of the instrument, as well as a different "random utility" ν_i in the non-treatment state. If γ_1 is positive, the instrument incents individuals towards treatment, since:

$$D_i(0) = \mathbb{1}(\gamma_0 > \nu_i) \text{ and } D_i(1) = \mathbb{1}(\gamma_0 + \gamma_1 > \nu_i)$$

so the monotonicity condition is immediately satisfied: $D_i(1) \geq D_i(0)$. If γ_1 were negative, we'd have monotonicity in the other direction.

Len Goff
Microeconometrics GR6414

Notes on distribution treatment effects

Please let me know if you spot any typos, etc.!

Moving beyond the mean

Average treatment effects are a convenient and intuitive summary of heterogeneous treatment effects. In the potential outcomes notation, we typically define individual i's treatment effect as $\Delta_i = Y_{1i} - Y_{0i}$. Recall for example that the local average treatment effect

$$LATE = E[\Delta_i | D_{1i} > D_{0i}]$$

is identified (i.e. can be estimated consistently from the data) when we have an instrumental variable satisfying the LATE model assumptions.

Great! Nevertheless, when treatment effects Δ_i are highly heterogeneous within the population of compliers, the average could be misleading. In the extreme case, imagine that treatment has a huge effect just for some small subgroup of the compliers. Then we might see a substantially positive LATE, even if treatment has a very small or even negative effect for most of the compliers. Is there any way to empirically distinguish this case from one in which all the compliers had the same treatment effect (which would then be equal to the LATE)?

Estimating the marginal distributions of potential outcomes

It turns out that we have a great tool at our disposal to "move beyond the mean" – under the standard LATE assumptions of independence, exclusion, and monotonicity, we can actually determine the effect of treatment on the whole distribution of Y among compliers. Even greater!

The result is based on two simple tricks: the first is that CDF of a random variable Y can be written as an expectation: $F_Y(y) = P(Y_i \leq y) = E[\mathbb{1}(Y_i \leq y)]$. The second is that if we want to learn about the distribution of one particular potential outcome, say Y_1, we can slip in a D_i into this expression:

$$\begin{aligned} E[D_i \mathbb{1}(Y_i \leq y)] &= P(D_i = 1) E[\mathbb{1}(Y_i \leq y) | D_i = 1] \\ &= P(D_i = 1) P(Y_{1i} \leq y | D_i = 1) = P(D_i = 1) F_{Y_1|D=1}(y) \end{aligned}$$

Thus the LHS, which can be estimated from the data, tells us something about the conditional distribution of the Y_1 potential outcome, up to a proportionality that we can also estimate.

This intuition underlies a more general result by Abadie (2003), which states that:

Lemma 2.1 from Abadie (2003). *Let $g(y)$ be any function at assume that we have the standard LATE assumptions of independence, exclusion and monotonicity (as defined in the previous set of recitation notes), and that we have a "non-zero first stage"($P(Z_i = 1) \in (0,1)$ and $P(D_{1i} > D_{0i}) > 0$). Then:*

$$E[g(Y_{1i})|D_{1i} > D_{0i}] = \frac{E[D_i g(Y_i)|Z_i = 1] - E[D_i g(Y_i)|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

and

$$E[g(Y_{0i})|D_{1i} > D_{0i}] = \frac{E[(1 - D_i)g(Y_i)|Z_i = 1] - E[(1 - D_i)g(Y_i)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]}$$

The result implies that if we pick some possible value y for Y_i, and let $g(Y_i) = \mathbb{1}[Y_i \leq y]$, then by the Lemma it follows that the CDF of Y_0 and Y_1 conditional on being a complier are each identified:

$$F_{Y_1|D_1>D_0}(y) = \frac{E[D_i \mathbb{1}(Y_i \leq y)|Z_i = 1] - E[D_i \mathbb{1}(Y_i \leq y)|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

and

$$F_{Y_0|D_1>D_0}(y) = \frac{E[(1 - D_i)\mathbb{1}(Y_i \leq y)|Z_i = 1] - E[(1 - D_i)\mathbb{1}(Y_i \leq y)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]}$$

where for clarity, by $F_{Y_d|D_1>D_0}(y)$ we mean $P(Y_{di} \leq y|D_1 > D_0)$ for each $d \in \{0, 1\}$.

Note that this type of result isn't specific to the IV research design: you might be interested to know that something analogous can also be done in an RDD setup (see Frandsen et al. 2012), and under more complicated assumptions in a diff-in-diff design too (Callaway 2015).

The RHS of the above two equations can be estimated from the data for each value of y. If we repeat this computation for all values of y, then we know the whole distribution function of each potential outcome, conditional on being a complier.

In the slides, we saw how this can be used to plot the complier distributions of Y_1 and Y_0 as densities in the charter school example. We saw in these figures that the charter school treatment appears to move the whole distribution of test scores to the right, consistent with the idea that the treatment effect is spread somewhat broadly among the compliers and not concentrated on just a few of them. If that were the case, we'd expect the Y_1 density to look like the Y_0 density in most places, but with one "piece" moved to the right.

Interlude: quantile treatment effects:

One thing that having $F_{Y_1|D_1>D_0}(y)$ and $F_{Y_0|D_1>D_0}(y)$ lets us compute is so-called *quantile treatment effects* (QTEs) among the compliers. For notational simplicity, let's drop the conditioning on being a complier: $D_{i1} > D_{i0}$. The (unconditional)

QTE is defined as

$$QTE(u) = F_1^{-1}(u) - F_0^{-1}(u)$$

where F_d^{-1} is the quantile function associated with potential outcome Y_d: $F_d^{-1}(u) = \inf\{y : P(Y_{id} \leq y) \geq u\}$ is the u^{th} quantile of Y_d, and u is a specified quantile level $u \in (0, 1)$. For example, if we measured test scores in points in the charter school example, and we found that $QTE(0.5) = 10$ (I just made up this number), then this would mean that the median outcome (test score) among compliers when they go to a KIPP school is 10 points higher than the median outcome among compliers when they do not go to a KIPP school. Knowing the QTE for all levels u is a way to summarize the difference between the two density curves plotted together in the slides.

Note that the QTEs *are* causal: they do tell us about the difference between the distribution of Y_1 and Y_0 (as opposed to the distributions $Y_1|D_i = 1$ and $Y_1|D_i = 0$, which are always identified in an observational study, but might be confounded by selection/endogeneity). Nevertheless, the QTEs do not tell us directly about the individual treatment effects Δ_i or their distribution, without further assumptions. The reason is that unlike the expectation function, the quantile function is not linear–thus: $QTE_i(u) \neq F_\Delta^{-1}(u)$. There is a notable exception: if we assume that each students' *rank* were the same in both the treated and untreated distributions: $F_0(Y_{0i}) = F_1(Y_{1i})$ for all i, then the u-quantile treatment effect is equal to the treatment effect for a student with rank u. However, this is a strong assumption that's hard to justify in general.

Without additional assumptions such as this *rank invariance* assumption, the marginal distributions $F_1(y)$ and $F_0(y)$ do generally place bounds on the distribution of treatment effects, which are sometimes informative. See for example Fan and Park (2009).

Proving the lemma from Abadie (2003)

Consider the first equality:

$$E[g(Y_{1i})|D_{1i} > D_{0i}] = \frac{E[D_i g(Y_i)|Z_i = 1] - E[D_i g(Y_i)|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \tag{1}$$

Since each term in the numerator is conditioned on a value of Z_i, we can rewrite it as:

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

where the notation $Y_{D_{1i}i}$ indicates either Y_{1i} or Y_{0i} depending on the value of D_{1i}, and so on (one way to express it is $Y_{D_{1i}i} = D_{1i}Y_{1i} + (1 - D_{1i})Y_{0i}$). Now, using independence between $(Y_{0i}, Y_{1i}, D_{0i}, D_{1i})$ and Z_i, we can drop the conditioning and our expression is equal to:

$$E[D_{1i}g(Y_{D_{1i}i})] - E[D_{0i}g(Y_{D_{0i}i})] = E[D_{1i}g(Y_{D_{1i}i}) - D_{0i}g(Y_{D_{0i}i})]$$

Now note that when $D_{1i} = D_{0i}$, the quantity $D_{1i}g(Y_{D_{1i}i}) - D_{0i}g(Y_{D_{0i}i})$ is equal to zero. Thus, if we apply LIE over the random variable $\mathbb{1}(D_{1i} = D_{0i})$, we have that:

$$\begin{aligned} E[D_{1i}g(Y_{D_{1i}i}) - D_{0i}g(Y_{D_{0i}i})] &= P(D_{1i} \neq D_{0i})E[D_{1i}g(Y_{D_{1i}i}) - D_{0i}g(Y_{D_{0i}i})|D_{1i} \neq D_{0i}] + 0 \\ &= P(D_{1i} > D_{0i})E[D_{1i}g(Y_{D_{1i}i}) - D_{0i}g(Y_{D_{0i}i})|D_{1i} > D_{0i}] \\ &= P(D_{1i} > D_{0i})E[g(Y_{1i}) - 0|D_{1i} > D_{0i}] \end{aligned}$$

where to move from the first line to the second line we use that by the monotonicity there are no defiers, so the event $D_{1i} \neq D_{0i}$ is the same as the event $D_{1i} > D_{0i}$, and to move to the third line we replace D_{1i} and D_{0i} by their values for compliers.

Now note that as with our original proof of the LATE theorem (see last set of notes), $E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = P(D_{1i} > D_{0i})$, and thus we've shown Eq. (1).

Recall that the second result of the Lemma is that:

$$E[g(Y_{0i})|D_{1i} > D_{0i}] = \frac{E[(1 - D_i)g(Y_i)|Z_i = 1] - E[(1 - D_i)g(Y_i)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]} \tag{2}$$

here the proof proceeds in exactly the same way. By the same steps, we can show that the numerator equals

$$\begin{aligned} &P(D_{1i} > D_{0i})E[(1 - D_{1i})g(Y_{D_{1i}i}) - (1 - D_{0i})g(Y_{D_{0i}i})|D_{1i} > D_{0i}] \\ &\quad = -P(D_{1i} > D_{0i})E[g(Y_{0i})|D_{1i} > D_{0i}] \end{aligned}$$

and the denominator equals $-P(D_{1i} > D_{0i})$, so the second result is proved.

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003. ISSN 01621459. doi: 10.1198/016214502753479419.

Brantly Callaway. Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with only Two Time Periods. 2015.

Yanqin Fan and Sangsoo Park. Sharp Bounds on the Distribution of the Treatment Effects and Their Statistical Inference. *Econometric Theory*, 26(3), 2009.

Brigham R Frandsen, Markus Frölich, and Blaise Melly. Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, 168(2):382–395, 2012. ISSN 03044076. doi: 10.1016/j.jeconom.2012.02.004.

Len Goff
Microeconometrics GR6414

Heterogeneous treatment effects in the fuzzy RDD

Please let me know if you spot any typos, etc.!

Motivation: fuzzy regression discontinuity with constant effects

In class we introduced the fuzzy regression discontinuity design (RDD), in which compliance with the mechanism $D_i = \mathbb{1}(X_i \geq c)$ is not perfect: there are some units that are untreated but above the cutoff $(X_i \geq c, D_i = 0)$, and some units that are treated but are below the cutoff $(X_i < c, D_i = 1)$. However, the probability of treatment is assumed to be discontinuous at the cutoff: $\lim_{x \downarrow c} E[D_i|X_i = x] > \lim_{x \uparrow c} E[D_i|X_i = x]$.

Let's start by assuming that all units have the same treatment effect, that is: $Y_i(1) = Y_i(0) + \rho$ for all units i. This is a restrictive assumption, but it will help us motivate what it is that we want to estimate in a fuzzy regression discontinuity. It turns out that this same quantity–given in Eq. (1)–yields an average treatment effect when effects are not constant (see next section). Let

$$E[D_i|X_i = x] = \begin{cases} g_0(x) & x < c \\ g_1(x) & x \geq c \end{cases}$$

where $g_1(c) - g_0(c) = \pi > 0$. Then, since $Y_i = Y_i(0) + \rho D_i$:

$$E[Y_i|X_i = x] = \begin{cases} E[Y_i(0)|X_i = x] + \rho g_0(x) & x < c \\ E[Y_i(0)|X_i = x] + \rho g_1(x) & x \geq c \end{cases}$$

Thus, assuming that $E[Y_i(0)|X_i = x]$ is continuous at $x = c$, the discontinuity $\lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]$ in the CEF of X at the cutoff is equal to $\pi\rho$, and thus the treatment effect ρ is:

$$\rho = \frac{\lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]}{\lim_{x \downarrow c} E[D_i|X_i = x] - \lim_{x \uparrow c} E[D_i|X_i = x]} \tag{1}$$

We can estimate ρ from this equality, since Y_i, X_i, and D_i are all observed.

A LATE model for the regression discontinuity

What is identified in the fuzzy research design when treatment effects can vary by individual? We'll see here that just as with instrumental variables, we can write down a LATE model in which we have a notion of *compliers* and fuzzy RDD estimates a treatment effect among them (in particular, we'll learn about the treatment effect for compliers close to the threshold c). In these notes I'll follow a formalization of the fuzzy RDD that can be found in, for example, the Frandsen, Frölich and Melly paper cited in the last set of slides.

The key piece of notation that we'll need is to model the assignment mechanism. For each individual i, we'll define a potential treatment for each value of the running variable: $d_i(x)$. A unit's realized treatment assignment will be this function evaluated at their actual value of the running variable: $D_i = d_i(X_i)$. Just as we discussed for the sharp design in class, we think of this assignment rule as deterministic: once you know X_i, you know D_i. The only difference is now we are letting the rule vary by individual.[1]

However, just as with IV, we'll make a monotonicity assumption about the way $d_i(x)$ behaves at the cutoff. Define $D_i^+ = \lim_{x \downarrow c} d_i(x)$ and $D_i^- = \lim_{x \uparrow c} d_i(x)$. Our monotonicity assumption is that:

$$\textbf{Monotonicity: } P(D_i^+ \geq D_i^-) = 1$$

This assumption says that for no individual i is the limit of their treatment assignment function from the left equal to one and the limit of their treatment assignment function from the right equal to zero–these would be "defiers"(recall that $d_i(x) \in \{0, 1\}$). Just as with IV, this implies that we can separate the population into three groups: always-takers, never-takers, and compliers.

1. Always-takers: $D_i^- = D_i^+ = 1$
2. Never-takers: $D_i^- = D_i^+ = 0$
3. Compliers: $D_i^- = 0, D_i^+ = 1$

The second main assumption we'll make is that all distributions of potential outcomes and potential treatments are continuous at the threshold, conditional on compliance group:

$$\textbf{Continuity: } E[Y_i(1)|X_i = x, D_i^+ = d, D_i^- = d'], E[Y_i(0)|X_i = x, D_i^+ = d, D_i^- = d'],$$
$$\text{and } P(D_i^+ = d, D_i^- = d'|X_i = x) \text{ are continuous at } x = c \text{ for all } d, d' \in \{0, 1\}$$

The continuity assumption captures the idea that everything that is discontinuous at the cutoff is due to treatment, and in particular reflects the response of compliers to crossing the threshold. This results in the following result:

Theorem. *If **continuity** and **monotonicity** hold, and $P(D_i^+ > D_i^-) > 0$ (i.e. there are some compliers) along with some regularity conditions, then:*

$$\frac{\lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]}{\lim_{x \downarrow c} E[D_i|X_i = x] - \lim_{x \uparrow c} E[D_i|X_i = x]} = E[Y_i(1) - Y_i(0)|X_i = c, D_i^+ > D_i^-] \quad (2)$$

This theorem tells us that from the joint distribution of (Y_i, X_i, D_i), which lets us compute the LHS of this expression, we can identify the average treatment effect among compliers close to the cutoff.[2] This result can be extended to look at things other than

[1] The function $d_i(x)$ can be correlated with all sorts of things about the individual–there's no restriction about that.

[2] Since X_i is taken to be continuously distributed, being *at* the cutoff is a measure zero event. So in practice we should interpret this as compliers that are very close to the cutoff, or more formally, we can identify the limit of the average treatment effect among compliers within some bandwidth of the cutoff, as that bandwidth goes to zero

the average treatment effect, e.g. quantile treatment effects (see the Frandsen, Frölich and Melly paper).

Note also that this theorem nests the sharp RDD as a special case. In the sharp RDD there are no always-takers or never-takers, and the denominator of Eq (2) is simply one. In the sharp case, since all units are compliers, we are learning about the average treatment effect among all units close the cutoff. However, there still may be difficulties generalizing this result to units that are far from the cutoff (since treatment effects could be correlated with X_i).

Proof of the result

To clean up notation, for any function $f(x)$ let $f(c^+) := \lim_{x\downarrow c} f(x)$ and $f(c^-) := \lim_{x\uparrow c} f(x)$, so $D_i^+ = d_i(c^+)$, $D_i^- = d_i(c^-)$, etc. Let's also let $p_{nevertaker|x} = P(D_i^+ = D_i^- = 0|X_i = x)$, $p_{complier|x} = P(D_i^+ > D_i^-|X_i = x)$ etc.

Consider the term $E[Y_i|c^+] = \lim_{x\downarrow c} E[Y_i|X_i = x]$ in the numerator of Eq (2). By the law of iterated expectations:

$$\begin{aligned} E[Y_i|c^+] &= p_{nevertaker|c} \cdot E[Y_i|X_i = c^+, D_i^+ = D_i^- = 0] \\ &+ p_{alwaystaker|c} \cdot E[Y_i|X_i = c^+, D_i^+ = D_i^- = 1] \\ &+ p_{complier|c} \cdot E[Y_i|X_i = c^+, D_i^+ > D_i^-] \end{aligned}$$

where we've used that the probabilities of being in any of the compliance groups $p_{nevertaker|x}$ etc. are continuous at $x = c$. However, since we haven't yet turned the Y_i into potential outcomes, we've kept the CEFs expressed as limits (with c^+), since Y_i itself is not continuous at the cutoff. But, since we've conditioned on D_i^+ and D_i^-, we know the realized treatment in each term, so:

$$\begin{aligned} E[Y_i|c^+] &= p_{nevertaker|c} \cdot E[Y_i(0)|X_i = c, D_i^+ = D_i^- = 0] \\ &+ p_{alwaystaker|c} \cdot E[Y_i(1)|X_i = c, D_i^+ = D_i^- = 1] \\ &+ p_{complier|c} \cdot E[Y_i(1)|X_i = c, D_i^+ > D_i^-] \end{aligned}$$

where since the potential outcome CEF's are continuous at c, we can replace c^+ with c after we've changed each Y_i to the corresponding potential outcome.

Following the same logic for $E[Y_i|Z_i = 0]$

$$\begin{aligned} E[Y_i|c^-] &= p_{nevertaker|c} \cdot E[Y_i(0)|X_i = c, D_i^+ = D_i^- = 0] \\ &+ p_{alwaystaker|c} \cdot E[Y_i(1)|X_i = c, D_i^+ = D_i^- = 1] \\ &+ p_{complier|c} \cdot E[Y_i(0)|X_i = c, D_i^+ > D_i^-] \end{aligned}$$

Thus, in the numerator of Eq (2): the always-taker and never-taker terms cancel out and:

$$E[Y_i|c^+] - E[Y_i|c^-] = p_{complier|c} \cdot E[Y_i(1) - Y_i(0)|X_i = c, D_i^+ > D_i^-]$$

The denominator of Eq (2) is exactly $p_{complier|c}$ since by continuity

$$E[D_i|c^+] - E[D_i|c^-] = \big(\cancel{p_{alwaystaker|c^+}} + p_{complier|c^+}\big) - \cancel{p_{alwaystaker|c^-}} = p_{complier|c}$$

and the theorem is proved.

Note how similar this proof was for the proof we presented for IV LATE, almost all the steps were in perfect analogy. Thus we see that in a fuzzy RDD, continuity plays the role of the independence assumption, and the two values of our binary instrument are replaced with limits from the right or the left of the cutoff. This formal analogy is so strong that we can literally use $Z_i = \mathbb{1}(x_i \geq c)$ as an instrument for treatment and estimate Eq (2) by 2SLS, as described in the slides.

Len Goff
Microeconometrics GR6414

Notes on double-robustness

Please let me know if you spot any typos, etc.!

Causal effects under unconfoundedness

Suppose that we're interested causal effects of a binary treatment D_i and we beleive that an *unconfoundedness* condition holds: $(Y_{0i}, Y_{1i}) \perp D_i | X_i$ where Y_{0i} and Y_{1i} are potential outcomes and X_i are a set of variables we observe. Under this assumption, the average treatment effect will be

$$\begin{aligned}
\mathbb{E}\left[Y_{1i} - Y_{0i}\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_{1i} - Y_{0i}|X_i, D_i\right]\right] && (L.I.E) \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_{1i}|X_i, D_i\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_{0i}|X_i, D_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_{1i}|X_i, D_i = 1\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_{0i}|X_i, D_i = 0\right]\right] && (unconfound.) \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i|X_i, D_i = 1\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i|X_i, D_i = 0\right]\right] \\
&= \mathbb{E}\left[g(1, X_i) - g(0, X_i)\right]
\end{aligned}$$

where the functions $g(0, X_i)$ and $g(1, X_i)$ are defined as $g(d, X_i) = \mathbb{E}\left[Y_i|D_i = d, X_i\right]$. These expectations can in principle be evaluated from our data, which let us learn the joint distribution of (Y_i, D_i, X_i). Yay!

Problem is: if X_i is high dimensional, as we might imagine in cases where the unconfoundedness assumption is plausible, we may have a hard time actually estimating these conditional expectations, because of the curse of dimensionality. We may be forced to make semiparametric or parametric restrictions, which could be wrong.

Note that there is another approach to the average treatment effect we might also consider using, that is valid under unconfoundedness:

$$\begin{aligned}
\mathbb{E}\left[\frac{D_iY_i}{m(X_i)} - \frac{(1-D_i)Y_i}{1-m(X_i)}\right] &= \mathbb{E}\left[\frac{\mathbb{E}\left[D_iY_i|X_i\right]}{m(X_i)} - \frac{\mathbb{E}\left[(1-D_i)Y_i|X_i\right]}{1-m(X_i)}\right] && (L.I.E) \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[D_iY_i|X_i\right]}{m(X_i)} - \frac{\mathbb{E}\left[(1-D_i)Y_i|X_i\right]}{1-m(X_i)}\right] \\
&= \mathbb{E}\Bigg[\frac{P(D_i = 1|X_i)\mathbb{E}\left[1 \cdot Y_i|X_i, D_i = 0\right] + 0}{m(X_i)} \\
&\qquad - \frac{0 + P(D_i = 0|X_i)\mathbb{E}\left[1 \cdot Y_i|X_i, D_i = 0\right]}{1-m(X_i)}\Bigg] \\
&= \mathbb{E}\left[\cancel{\frac{P(D_i = 1|X_i)}{m(X_i)}}\mathbb{E}\left[Y_{1i}|X_i, D_i = 1\right] - \cancel{\frac{P(D_i = 0|X_i)}{1-m(X_i)}}\mathbb{E}\left[Y_{0i}|X_i, D_i = 0\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_{1i}|X_i\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_{0i}|X_i\right]\right] && (unconfound.) \\
&= \mathbb{E}\left[Y_{1i} - Y_{0i}\right]
\end{aligned}$$

where $m(x) = \mathbb{E}\left[D_i|X_i = x\right]$ is a propensity score function, which can in principle be computed from the data. The above expression, which is like the population version of "re-weighting" each observation by the inverse of its propensity score, is merely a different way of expressing the same quantity as in the first regression-based approach, and suggests a different route towards estimating it.

However, we have a similar problem here, which is that the CEF $m(x)$ may be hard to estimate if x is very high dimensional, and we may introduce misspecification bias if we impose restrictions intended to improve estimation.

The doubly-robust approach

Now consider an alternative estimation approach, based on the following equality:

$$\mathbb{E}\left[Y_{1i} - Y_{0i}\right] = \mathbb{E}\left[g(1, X_i) - g(0, X_i) + \frac{D_i(Y_i - g(1, X_i))}{m(X_i)} - \frac{(1 - D_i)(Y_i - g(0, X_i))}{1 - m(X_i)}\right]$$

(this equation can be verified with L.I.E and the unconfoundedness assumptions, similar to the steps above for the propensity score re-weighting calculation).[1]

To implement this approach, we'd need estimators of the $g(d, x)$ functions as well as the propensity score function $m(x)$, but with them we could estimate $\mathbb{E}\left[Y_{1i} - Y_{0i}\right]$ by

$$\begin{aligned}\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{g}_n(1, X_i) - \hat{g}_n(0, X_i) + \frac{D_i(Y_i - \hat{g}_n(1, X_i))}{\hat{m}_n(X_i)} - \frac{(1 - D_i)(Y_i - \hat{g}_n(0, X_i))}{1 - \hat{m}_n(X_i)}\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{g}_n(1, X_i) + \frac{D_i(Y_i - \hat{g}_n(1, X_i))}{\hat{m}_n(X_i)}\right\} - \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{g}_n(0, X_i) + \frac{(1 - D_i)(Y_i - \hat{g}_n(0, X_i))}{1 - \hat{m}_n(X_i)}\right\}\end{aligned}$$

Let $\delta_n^d(x)$ be the error in our estimate of the function $g(d, x)$, i.e. $g(d, x) = \hat{g}_n(d, x) + \delta_n^d(x)$, and similarly let $m(x) = \hat{m}_n(x) + \eta^n(x)$. Furthermore let ϵ_i be the difference between Y_i and it's conditional expectation on D_i and X_i: $\epsilon_i = Y_i - g(D_i, X_i)$, and similar for D_i: $\nu_i := D_i - m(X_i)$. Call the first term in the above expression $\hat{\theta}_1$, which is

$$\begin{aligned}\hat{\theta}_1 =& \frac{1}{n}\sum_{i=1}^{n}\left\{g(1, X_i) - \delta_n^1(X_i) + \frac{(\hat{m}_n(X_i) + \eta_n(X_i) + \nu_i)(\cancel{\hat{g}_n(1, X_i)} + \delta_n^1(X_i) + \epsilon_i - \cancel{\hat{g}_n(1, X_i)})}{\hat{m}_n(X_i)}\right\}\\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{g(1, X_i) - \cancel{\delta_n^1(X_i)} + \cancel{\delta_n^1(X_i)} + \epsilon_i + \frac{(\eta_n(X_i) + \nu_i)(\delta_n^1(X_i) + \epsilon_i)}{\hat{m}_n(X_i)}\right\}\end{aligned}$$

Since the purely stochastic errors ϵ_i and ν_i are mean zero conditional on X_i, the terms proportional to them will not contribute to this expression as the sample gets very large. On the other hand, the functions $\hat{g}_n(d, x)$ and $\hat{m}_n(x)$ may be subject to specification error, in which case $\delta_n^d(x)$ and $\eta_n(x)$ may not converge to zero asymptotically. Let $\delta^d(x) = \text{plim}\delta^d(x)$ and $\eta(x) = \text{plim}\eta_n(x)$. Then, under regularity conditions the probability limit

[1] Note, there is another version of this that is also valid, which starts from the propensity score approach and adds terms like $\frac{D_i - m(X_i)}{m(X_i)}g(1, X_i)$ to achieve double-robustness.

of $\hat{\theta}_1$ is

$$\begin{aligned}
\hat{\theta}_1 &\xrightarrow{p} \mathbb{E}\left[g(1, X_i) + \epsilon_i + \frac{(\eta(X_i) + \nu_i)(\delta^1(X_i) + \epsilon_i)}{m(X_i) + \eta(X_i)}\right] \\
&= \mathbb{E}\left[g(1, X_i) + \mathbb{E}\left[\epsilon_i | X_i\right] + \frac{(\eta(X_i) + \mathbb{E}\left[\nu_i | X_i\right])(\delta^1(X_i) + \mathbb{E}\left[\epsilon_i | X_i\right])}{m(X_i) + \eta(X_i)}\right] \\
&= \mathbb{E}\left[g(1, X_i) + \frac{\eta(X_i)\delta^1(X_i)}{m(X_i) + \eta(X_i)}\right]
\end{aligned}$$

Notice that if *either* $\hat{g}_n(1, x)$ or $\hat{m}_n(x)$ are consistent estimators for all z (i.e. $\delta^1(x) = 0$ or $\eta(x) = 0$ for all x), then the second term will be xero, making $\hat{\theta}_1$ consistent for $\mathbb{E}\left[g(1, X_i)\right]$. That is, we can misspecify the model for the propensity score, or the CEF of the outcome variable, just not both. Furthermore, if both models are misspecified, but the asymptotic bias is small, then our doubly-robust estimator is only off by a factor that's proportional to the product of the two errors.

Naturally, all of the above carries through for the second term (the $D = 0$ term), yielding the double robustness property for the full treatment effect estimator $\hat{\theta}$.