14.381 Notes

Rebekah Dix and Shinnosuke Kikuchi *

December 13, 2019

Contents

*Notes written during 14.381 Fall 2019 at MIT. Taught by Josh Angrist.

Todo list

1 RCT and Regression Recap

1.1 Conditional Expectation Function

Definition 1.1 (Conditional Expectation Function (CEF))**.** The **Conditional Expectation Function (CEF)** for a dependent variable Y_i given a $K \times 1$ vector of covariates X_i is the expectation (or *population average*) of Y_i with X_i held fixed. It is written as $\mathbb{E}\left(Y_i|X_i\right)$. For a specific value of $X_i = x$, we write $\mathbb{E}\left(Y_i|X_i = x\right)$.

- For continuous Y_i with conditional density $f_y(\cdot|X_i = x)$, the CEF is

$$\mathbb{E}\left(Y_i|X_i = x\right) = \int t f_y(t|X_i = x)\, dt \tag{1.1}$$

- For discrete Y_i with condition pmf $f_y(\cdot|X_i = x)$, the CEF is

$$\mathbb{E}\left(Y_i|X_i = x\right) = \sum_t t f_y(t|X_i = x) \tag{1.2}$$

Theorem 1.1 (Law of Iterated Expectations (LIE))**.**

$$\mathbb{E}\left(Y_i\right) = \mathbb{E}\left(\mathbb{E}\left(Y_i|X_i\right)\right) \tag{1.3}$$

In words: an unconditional expectation can be written as the population average of the CEF.

Theorem 1.2 (CEF Decomposition Property)**.** We can write

$$Y_i = \mathbb{E}\left(Y_i|X_i\right) + \varepsilon_i \tag{1.4}$$

where

(i) ε_i is mean-independent of X_i (that is, $\mathbb{E}\left(e_i|X_i\right) = 0$).

(ii) ε_i is uncorrelated with any function of X_i (and hence uncorrelated with the CEF, which is a function of X_i).

Interpretation. This theorem says any random variable Y_i can be decomposed into a piece that's explain by X_i (the CEF) and a piece left over which is orthogonal (i.e., uncorrelated) with any function of X_i.

Proof. Write

$$Y_i = \mathbb{E}\left(Y_i|X_i\right) + \varepsilon_i \tag{1.5}$$

We prove each item in turn:

(i) $\mathbb{E}\left(\varepsilon_i|X_i\right) = \mathbb{E}\left(Y_i - \mathbb{E}\left(Y_i|X_i\right)|X_i\right) = \mathbb{E}\left(Y_i|X_i\right) - \mathbb{E}\left(Y_i|X_i\right) = 0.$

- Note also that

$$\mathbb{E}\left(\varepsilon_i\right) = \mathbb{E}\left(\mathbb{E}\left(\varepsilon_i|X_i\right)\right) = 0 \tag{1.6}$$

(ii) Let h be a function of X_i. Then

$$\begin{aligned}\mathbb{E}\left(h(X_i)\varepsilon_i\right) &= \mathbb{E}\left(\mathbb{E}\left(h(X_i)\varepsilon\right)|X_i\right) && \text{(LIE)}\\ &= \mathbb{E}\left(h(X_i)\,\mathbb{E}\left(\varepsilon_i|X_i\right)\right)\end{aligned}$$

$$= 0$$

Then

$$\text{Cov}\left(h(X_i), \varepsilon_i\right) = \mathbb{E}\left(h(X_i)\varepsilon_i\right) - \mathbb{E}\left(h(X_i)\right)\mathbb{E}\left(\varepsilon_i\right) = 0 \tag{1.7}$$

since $\mathbb{E}\left(\varepsilon_i\right) = \mathbb{E}\left(h(X_i)\varepsilon_i\right) = 0$.

$\square$

Theorem 1.3 (CEF solves the MMSE prediction problem). Let $m(X_i)$ be any function of X_i. The CEF solves

$$\mathbb{E}\left(Y_i|X_i\right) = \underset{m(X_i)}{\text{argmin}}\, \mathbb{E}\left((Y_i - m(X_i))^2\right) \tag{1.8}$$

Proof. Notice that

$$\begin{aligned}(Y_i - m(X_i))^2 &= (Y_i - \mathbb{E}\left(Y_i|X_i\right) + \mathbb{E}\left(Y_i|X_i\right) - m(X_i))^2 \\ &= (Y_i - \mathbb{E}\left(Y_i|X_i\right))^2 + 2(Y_i - \mathbb{E}\left(Y_i|X_i\right))(\mathbb{E}\left(Y_i|X_i\right) - m(X_i)) + (\mathbb{E}\left(Y_i|X_i\right) - m(X_i))^2\end{aligned}$$

Note that

$$(Y_i - \mathbb{E}\left(Y_i|X_i\right))(\mathbb{E}\left(Y_i|X_i\right) - m(X_i)) = \varepsilon_i(\mathbb{E}\left(Y_i|X_i\right) - m(X_i)) = \varepsilon_i h(X_i) \tag{1.9}$$

where h is only a function of X_i and $\varepsilon_i = Y_i - \mathbb{E}\left(Y_i|X_i\right)$. We know that ε_i is orthogonal with any function of X_i. Therefore $\mathbb{E}\left(\varepsilon_i h(X_i)\right) = 0$. Therefore to minimize, $(Y_i - m(X_i))^2$, we need to minimize $(\mathbb{E}\left(Y_i|X_i\right) - m(X_i))^2$, so choose $m(X_i) = \mathbb{E}\left(Y_i|X_i\right)$. $\square$

Interpretation. The CEF is the best predictor of Y_i given X_i in that it solves the Minimum Mean Squared Error (MMSE) prediction problem.

Theorem 1.4 (ANOVA). The Analysis of Variance (ANOVA) Theorem states that

$$\text{Var}\left(Y_i\right) = \text{Var}\left(\mathbb{E}\left(Y_i|X_i\right)\right) + \mathbb{E}\left(\text{Var}\left(Y_i|X_i\right)\right) \tag{1.10}$$

Proof. By the CEF Decomposition Property, we have that

$$\begin{aligned}\text{Var}\left(Y_i\right) &= \text{Var}\left(\mathbb{E}\left(Y_i|X_i\right) + \varepsilon_i\right) \\ &= \text{Var}\left(\mathbb{E}\left(Y_i|X_i\right)\right) + \text{Var}\left(\varepsilon_i\right) && (\varepsilon_i \text{ and } \mathbb{E}\left(Y_i|X_i\right) \text{ uncorrelated})\end{aligned}$$

Then note that

$$\begin{aligned}\text{Var}\left(\varepsilon_i\right) &= \mathbb{E}\left(\varepsilon_i^2\right) - \mathbb{E}\left(\varepsilon_i\right)^2 \\ &= \mathbb{E}\left(\varepsilon_i^2\right) && (\varepsilon_i \text{ mean-zero}) \\ &= \mathbb{E}\left(\mathbb{E}\left(\varepsilon_i^2|X_i\right)\right) && (\text{LIE})\end{aligned}$$

Now note that

$$\begin{aligned}\mathbb{E}\left(\varepsilon_i^2|X_i\right) &= \mathbb{E}\left((Y_i - \mathbb{E}\left(Y_i|X_i\right))^2|X_i\right) && (\varepsilon_i \equiv Y_i - \mathbb{E}\left(Y_i|X_i\right)) \\ &= \mathbb{E}\left(\text{Var}\left(Y_i|X_i\right)\right)\end{aligned}$$

Therefore

$$\text{Var}\left(\varepsilon_i\right) = \mathbb{E}\left(\mathbb{E}\left(\varepsilon_i^2|X_i\right)\right) = \mathbb{E}\left(\text{Var}\left(Y_i|X_i\right)\right) \tag{1.11}$$

Putting these pieces together gives

$$\text{Var}\left(Y_i\right) = \text{Var}\left(\mathbb{E}\left(Y_i|X_i\right)\right) + \mathbb{E}\left(\text{Var}\left(Y_i|X_i\right)\right) \tag{1.12}$$

□

1.2 Linear Regression and the CEF

Define the $K \times 1$ regression coefficient vector β by solving

$$\beta = \underset{b}{\text{argmin}}\, \mathbb{E}\left((Y_i - X_i'b)^2\right) \tag{1.13}$$

The FOC is

$$\mathbb{E}\left(X_i(Y_i - X_i'\beta)\right) = 0 \tag{1.14}$$

Therefore

$$\beta = \mathbb{E}\left(X_i X_i'\right)^{-1} \mathbb{E}\left(X_i Y_i\right) \tag{1.15}$$

Define the **population residual** to be

$$e_i \equiv Y_i - X_i'\beta \tag{1.16}$$

Then the population residual is uncorrelated with the regressors X_i by the FOC

$$\mathbb{E}\left(X_i(Y_i - X_i'\beta)\right) = \mathbb{E}\left(X_i e_i\right) = 0 \tag{1.17}$$

Bivariate Regression Suppose there is only one regressor x_i and a constant:

$$Y_i = \alpha + \beta_1 x_i + e_i \tag{1.18}$$

Then the (population) regression coefficients are given by

$$\beta_1 = \frac{\text{Cov}\left(Y_i, x_i\right)}{\text{Var}\left(x_i\right)} \tag{1.19}$$

$$\alpha = \mathbb{E}\left(Y_i\right) - \beta_1 \mathbb{E}\left(x_i\right) \tag{1.20}$$

Multivariate Regression Suppose there are multiple non-constant regressors. The slope coefficient for the k-th regressor is given by

$$\beta_k = \frac{\text{Cov}\left(Y_i, \tilde{x}_{ki}\right)}{\text{Var}\left(\tilde{x}_{ki}\right)} \tag{1.21}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all the other covariates.

Interpretation. Each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after partialling out all the other variables in the model.

1.3 Regression Justification

Theorem 1.5 (The Linear CEF Theorem)**.** Suppose the CEF is linear. Then the population regression function is the CEF.

Theorem 1.6. Suppose the CEF is linear:

$$\mathbb{E}\left(Y_i|X_i\right) = X_i'\beta^* \tag{1.22}$$

where β^* is a $K \times 1$ vector of coefficients. We know that

$$\mathbb{E}\left(X_i(Y_i - \mathbb{E}\left(Y_i|X_i\right))\right) = 0 \tag{1.23}$$

by the CEF Decomposition Property. Then

$$\begin{aligned}
\beta^* &= \mathbb{E}\left(X_iX_i'\right)^{-1}\mathbb{E}\left(X_iY_i\right) && \text{(Linear Regression)}\\
&= \mathbb{E}\left(X_iX_i'\right)^{-1}\mathbb{E}\left(\mathbb{E}\left(X_iY_i|X_i\right)\right) && \text{(LIE)}\\
&= \mathbb{E}\left(X_iX_i'\right)^{-1}\mathbb{E}\left(X_i\,\mathbb{E}\left(Y_i|X_i\right)\right)\\
&= \mathbb{E}\left(X_iX_i'\right)^{-1}\mathbb{E}\left(X_iX_i'\beta\right) && \text{(CEF Linear)}\\
&= \beta
\end{aligned}$$

When is the CEF linear?

(i) Joint Normality: The vector (Y_i, X_i) has a multivariate normal distribution

(ii) **Regression model saturated**: that is, the saturated regression model has a separate parameter for every possible combination of values that the set of regressors can take on.

Theorem 1.7 (Best Linear Predictor Theorem)**.** The function $X_i'\beta$ where $\beta = \mathbb{E}\left(X_iX_i'\right)^{-1}\mathbb{E}\left(X_iY_i\right)$ is the best linear predictor of Y_i given X_i in a MMSE sense.

Interpretation. The CEF $\mathbb{E}\left(Y_i|X_i\right)$ is the best MMSE predictor of Y_i given X_i in the class of all functions of X_i. The population regression function is the best we can do in the class of linear functions.

Theorem 1.8 (The Regression-CEF Theorem)**.** The function $X_i'\beta$ provides the MMSE linear approximation to $\mathbb{E}\left(Y_i|X_i\right)$. More precisely,

$$\beta = \underset{b}{\operatorname{argmin}}\,\mathbb{E}\left((\mathbb{E}\left(Y_i|X_i\right) - X_i'b)^2\right) \tag{1.24}$$

Interpretation. Even if the CEF is nonlinear, regression provides the best linear approximation to it.

Corollary 1.9. Regression coefficients can be obtained by using $\mathbb{E}\left(Y_i|X_i\right)$ as a dependent variable instead of Y_i itself.

$$\beta = \mathbb{E}\left(X_iX_i'\right)^{-1}\mathbb{E}\left(X_iY_i\right) \qquad \text{(Linear Regression)}$$

$$
\begin{aligned}
&= \mathbb{E}\left(X_i X_i'\right)^{-1} \mathbb{E}\left(\mathbb{E}\left(X_i Y_i | X_i\right)\right) && \text{(LIE)} \\
&= \mathbb{E}\left(X_i X_i'\right)^{-1} \mathbb{E}\left(X_i\, \mathbb{E}\left(Y_i | X_i\right)\right)
\end{aligned}
$$

Remark. Useful for grouped-data regression.

1.4 Fits and Residuals

Suppose α and $\beta_1, \ldots, \beta_k$ are the intercept and slope coefficients from a regression of Y_i on $X_{1i}, \ldots, X_{ki}$. The **fitted values** from this regression are

$$\hat{Y}_i = \alpha + \sum_{k=1}^{K} \beta_k X_{ki} \tag{1.25}$$

The **residuals** are

$$e_i = Y_i - \hat{Y}_i = Y_i - \alpha - \sum_{k=1}^{K} \beta_k X_{ki} \tag{1.26}$$

Properties Regression residuals

(i) Have expectation and sample mean 0:

$$\mathbb{E}\left(e_i\right) = \sum_{i=1}^{n} e_i = 0 \tag{1.27}$$

(ii) Are uncorrelated (in population *and* sample) with **all regressors that made them** and with the **corresponding fitted values.** That is,

- $\mathbb{E}\left(X_{ki} e_i\right) = \sum_{i=1}^{n} X_{ki} e_i = 0$
- $\mathbb{E}\left(\hat{Y}_i e_i\right) = \sum_{i=1}^{n} \hat{Y}_i e_i = 0$

Proof. Look at FOCs of minimization problem.

Finish.

□

1.5 Bivariate Regression with Dummy Regressor

Define

- $\mathbb{E}\left(Y_i | Z_i = 0\right) = \alpha$
- $\mathbb{E}\left(Y_i | Z_i = 1\right) - \mathbb{E}\left(Y_i | Z_i = 0\right) = \alpha + \beta$
  - Thus $\beta = \mathbb{E}\left(Y_i | Z_i = 1\right) - \mathbb{E}\left(Y_i | Z_i = 0\right)$

Then

$$
\begin{aligned}
\mathbb{E}\left(Y_i | Z_i\right) &= \mathbb{E}\left(Y_i | Z_i = 0\right) + \left(\mathbb{E}\left(Y_i | Z_i = 1\right) - \mathbb{E}\left(Y_i | Z_i = 0\right)\right) Z_i \\
&= \alpha + \beta Z_i
\end{aligned}
$$

Thus $\mathbb{E}\left(Y_i|Z_i\right)$ is a linear function of Z_i, with slope β and intercept α. Hence regression fits this CEF perfectly.

1.6 Asymptotic OLS Inference

Finish this.

1.7 Omitted Variables Bias

1.7.1 Short: Bivariate, Long: Multivariate

Long Regression

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l \tag{1.28}$$

Short Regression

$$Y_i = \alpha^s + \beta^s X_{1i} + e_i^s \tag{1.29}$$

Omitted on Included

$$X_{2i} = \pi_{20} + \pi_{21} X_{1i} + e_i^b \tag{1.30}$$

OVB Formula

$$\beta^s = \beta^l + \pi_{21}\gamma \tag{1.31}$$

Remark. "Short equals long plus the effect of omitted times the regression of omitted on included."

Derivation

$$\begin{aligned}
\beta^s &= \frac{\text{Cov}\left(Y_i, X_{1i}\right)}{\text{Var}\left(X_{1i}\right)} && \text{(regression coefficient in bivariate regression)} \\
&= \frac{\text{Cov}\left(\alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l, X_{1i}\right)}{\text{Var}\left(X_{1i}\right)} && \text{(subsititute long regression equation)} \\
&= \beta^l + \gamma\frac{\text{Cov}\left(X_{2i}, X_{1i}\right)}{\text{Var}\left(X_{1i}\right)} + \frac{\text{Cov}\left(e_i^l, X_{1i}\right)}{\text{Var}\left(X_{1i}\right)} \\
&= \beta^l + \gamma\frac{\text{Cov}\left(X_{2i}, X_{1i}\right)}{\text{Var}\left(X_{1i}\right)} \\
& \qquad \text{(e_i^l residual from a regression that includes X_{1i} as a regressor $\Rightarrow$ uncorrelated)} \\
&= \beta^l + \pi_{21}\gamma
\end{aligned}$$

1.7.2 Short, Long: Multivariate

Long Regression

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + \delta^l X_{3i} + e_i^l \tag{1.32}$$

Short Regression

$$Y_i = \alpha^s + \beta^s X_{1i} + \delta^s X_{3i} + e_i^s \tag{1.33}$$

Omitted on Included

$$X_{2i} = \pi_{20} + \pi_{21} X_{1i} + \pi_{23} X_{3i} + e_i^b \tag{1.34}$$

OVB Formula

$$\beta^s = \beta^l + \pi_{21}\gamma \tag{1.35}$$

Derivation

$$\begin{aligned}
\beta^s &= \frac{\text{Cov}\left(Y_i, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)} && \text{(regression anatomy: } \tilde{X}_{1i} \text{ residual of regression of } X_{1i} \text{ on } X_{3i}\text{)} \\
&= \frac{\text{Cov}\left(\alpha^l + \beta^l X_{1i} + \gamma X_{2i} + \delta^l X_{3i} + e_i^l, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)} && \text{(substitute long regression equation)} \\
&= \underbrace{\frac{\text{Cov}\left(\alpha^l, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)}}_{=0\text{, constant}} + \frac{\text{Cov}\left(\beta^l X_{1i}, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)} + \frac{\text{Cov}\left(\gamma X_{2i}, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)} + \underbrace{\frac{\text{Cov}\left(\delta^l X_{3i}, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)}}_{=0,\ \tilde{X}_{1i} \text{ residual, } X_{3i} \text{ regressor}} + \underbrace{\frac{\text{Cov}\left(e_i^l, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)}}_{=0,\ e_i^l \text{ residual, } \tilde{X}_{1i} \text{ made up of regressors}} \\
&= \beta^l \frac{\text{Cov}\left(X_{1i}, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)} + \gamma \frac{\text{Cov}\left(X_{2i}, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)} \\
&= \beta^l + \pi_{21}\gamma && \left(\text{Cov}\left(X_{1i}, \tilde{X}_{1i}\right) = \text{Var}\left(\tilde{X}_{1i}\right), \tfrac{\text{Cov}\left(X_{2i}, \tilde{X}_{1i}\right)}{\text{Var}\left(\tilde{X}_{1i}\right)} \text{ slope coefficient regression anatomy}\right)
\end{aligned}$$

1.8 Bad Controls

- Variables measure before the treatment variable was determine are generally *good controls*, because they can't be changed by the treatment.
- Control variables that are measured later may have been determined in part by the treatment – they are outcomes. These variables are *bad controls*. For example, consider the regression of wage on education and test scores as a proxy for ability. If the test score is SAT score, it can be thought as a part of outcomes from education. This should not be included to estimate return to education.

1.9 Measurement Error

1.9.1 Bivariate

We want to run the regression

$$Y_i = \alpha + \beta S_i^* + e_i \tag{1.36}$$

but rather than observing S_i^* we observe

$$S_i = S_i^* + m_i \tag{1.37}$$

Classical Measurement Error Assumptions. Assume the errors average to zero and are uncorrelated with S_i^* and e_i:

(i) $\mathbb{E}\left(m_i\right) = 0$

(ii) $\text{Cov}\left(S_i^*, m_i\right) = \text{Cov}\left(e_i, m_i\right) = 0$

The regression coefficient we want, β, is given by

$$\beta = \frac{\text{Cov}\left(Y_i, S_i^*\right)}{\text{Var}\left(S_i^*\right)} \tag{1.38}$$

Instead we calculate β_b (the "biased" coefficient) as

$$\begin{aligned}
\beta_b &= \frac{\text{Cov}\left(Y_i, S_i\right)}{\text{Var}\left(S_i\right)} \\
&= \frac{\text{Cov}\left(\alpha+\beta S_i^*+e_i, S_i^*+m_i\right)}{\text{Var}\left(S_i\right)} && \text{(subsitute regression equation and ME)} \\
&= \beta\frac{\text{Var}\left(S_i^*\right)}{\text{Var}\left(S_i\right)} && \text{(m_i is uncorrelated with either S_i^* or e_i)} \\
&= \beta\frac{\text{Var}\left(S_i^*\right)}{\text{Var}\left(S_i^*\right)+\text{Var}\left(m_i\right)}
\end{aligned}$$

Note that

$$r \equiv \frac{\text{Var}\left(S_i^*\right)}{\text{Var}\left(S_i^*\right)+\text{Var}\left(m_i\right)} \in [0,1] \tag{1.39}$$

Interpretation. r is the proportion of variation in S_i that is unrelated to mistakes (measurement error) and is called the **reliability** of S_i. The attenuation bias in β_b is

$$\beta_b - \beta = -(1-r)\beta \tag{1.40}$$

so the absolute value of β_b is smaller than β, which we call *attenuation bias*.

1.9.2 IV Eliminates Measurement Error in Bivariate Regression

Suppose we want to run the regression

$$Y_i = \alpha + \beta S_i^* + e_i \tag{1.41}$$

but rather than observing S_i^* we observe

$$S_i = S_i^* + m_i \tag{1.42}$$

We also have an instrument Z_i for S_i. In particular, we have that $\text{Cov}\left(e_i, Z_i\right) = 0$, which is *exclusion restriction*.

Classical Measurement Error Assumptions. Assume the errors average to zero and are uncorrelated with S_i^*, e_i, and Z_i.

(i) $\mathbb{E}\left(m_i\right) = 0$

(ii) $\text{Cov}\left(S_i^*, m_i\right) = \text{Cov}\left(e_i, m_i\right) = \text{Cov}\left(Z_i, m_i\right) = 0$

The IV estimator for β is

$$\begin{aligned}
\beta_{IV} &= \frac{\text{Cov}\left(Y_i, Z_i\right)}{\text{Cov}\left(S_i, Z_i\right)} \\
&= \frac{\text{Cov}\left(\alpha+\beta S_i^*+e_i, Z_i\right)}{\text{Cov}\left(S_i^*+m_i, Z_i\right)} \\
&= \frac{\beta\text{Cov}\left(S_i^*, Z_i\right)+\text{Cov}\left(e_i, Z_i\right)}{\text{Cov}\left(S_i^*, Z_i\right)+\text{Cov}\left(m_i, Z_i\right)} \\
&= \beta && \text{($\text{Cov}\left(Z_i, m_i\right)=0$ by CME, $\text{Cov}\left(e_i, Z_i\right)=0$ by exclusion)}
\end{aligned}$$

1.9.3 Multivariate

We want to run the regression

$$Y_i = \alpha + \beta S_i^* + \gamma X_i + e_i \tag{1.43}$$

but rather than observing S_i^* we observe

$$S_i = S_i^* + m_i \tag{1.44}$$

The regression coefficient we want, β, is given by

$$\beta = \frac{\text{Cov}\left(Y_i, \tilde{S}_i^*\right)}{\text{Var}\left(\tilde{S}_i^*\right)} \tag{1.45}$$

where

- $\tilde{S}_i^*$ is the residual from a regression of S_i^* on X_i.

The regression coefficient we calculate is

$$\beta_b = \frac{\text{Cov}\left(Y_i, \tilde{S}_i\right)}{\text{Var}\left(\tilde{S}_i\right)} \tag{1.46}$$

where

- $\tilde{S}_i$ is the residual from a regression of S_i on X_i.

Classical Measurement Error Assumptions. Assume the errors average to zero and are uncorrelated with S_i^*, e_i, and X_i:

(i) $\mathbb{E}\left(m_i\right) = 0$

(ii) $\text{Cov}\left(S_i^*, m_i\right) = \text{Cov}\left(e_i, m_i\right) = \text{Cov}\left(X_i, m_i\right) = 0$

The coefficient from a regression of S_i on X_i is the same as the coefficient from a regression of S_i^* on X_i:

$$\frac{\text{Cov}\left(S_i, X_i\right)}{\text{Var}\left(X_i\right)} = \frac{\text{Cov}\left(S_i^* + m_i, X_i\right)}{\text{Var}\left(X_i\right)} = \frac{\text{Cov}\left(S_i^*, X_i\right)}{\text{Var}\left(X_i\right)} \tag{1.47}$$

Therefore

$$\tilde{S}_i = \tilde{S}_i^* + m_i \tag{1.48}$$

To see this, note that

$$\begin{aligned}
S_i^* &= \pi X_i + \tilde{S}_i^* \\
S_i &= \pi X_i + \tilde{S}_i \\
\Rightarrow \tilde{S}_i &= \tilde{S}_i^* + \left(S_i - S_i^*\right) \\
&= \tilde{S}_i^* + m_i
\end{aligned}$$

where $\tilde{S}_i^*$ and m_i are uncorrelated, since $\tilde{S}_i^*$ is the residual of a regression of S_i^* on X_i, both of which are assumed to be uncorrelated with m_i.

Thus

$$\text{Var}\left(\tilde{S}_i\right) = \text{Var}\left(\tilde{S}_i^*\right) + \text{Var}\left(m_i\right) \tag{1.49}$$

Therefore

$$\beta_b = \frac{\text{Cov}\left(Y_i, \tilde{S}_i\right)}{\text{Var}\left(\tilde{S}_i\right)}$$

$$
\begin{aligned}
&= \frac{\text{Cov}\left(\alpha + \beta S_i^* + \gamma X_i + e_i, \tilde{S}_i\right)}{\text{Var}\left(\tilde{S}_i\right)} \\
&= \frac{\text{Cov}\left(S_i^*, \tilde{S}_i\right)}{\text{Var}\left(\tilde{S}_i\right)} \beta \\
&= \frac{\text{Var}\left(\tilde{S}_i^*\right)}{\text{Var}\left(\tilde{S}_i^*\right) + \text{Var}\left(m_i\right)} \beta \\
&\equiv \tilde{r}\beta
\end{aligned}
$$

Note that $\text{Var}\left(\tilde{S}_i^*\right) \leq \text{Var}\left(S_i^*\right)$ since $\text{Var}\left(\tilde{S}_i^*\right)$ is the variance of a residual form a regression in which S_i^* is the dependent variable. This implies that

$$
\tilde{r} = \frac{\text{Var}\left(\tilde{S}_i^*\right)}{\text{Var}\left(\tilde{S}_i^*\right) + \text{Var}\left(m_i\right)} \leq \frac{\text{Var}\left(S_i^*\right)}{\text{Var}\left(S_i^*\right) + \text{Var}\left(m_i\right)} = r \tag{1.50}
$$

Interpretation. Adding covariates to a model with mismeasured schooling aggravates attenuation bias in estimates of the returns to schooling. Covariates are correlated with accurately measured schooling while being unrelated to the measurement error.

1.10 Limited dependent variables

Finish

2 Conditional Independence Assumptions

Claim 2.1.

$$
\text{Cov}\left(X, Y\right) = \mathbb{E}\left(\left(X - \mathbb{E}\left(X\right)\right)Y\right) \tag{2.1}
$$

Interpretation. You only need to difference *one* of the random variables when calculating covariance.

2.1 Matching meets regression

Matching Conditional on X_i, treatment D_i is as good as randomly assigned

$$
\{\text{Y}_{0i}, \text{Y}_{1i}\} \perp \text{D}_i | \text{X}_i \tag{2.2}
$$

Then TOT(Treatment effect On Treated) is

$$
\begin{aligned}
\delta_{TOT} &\equiv \mathbb{E}\left(Y_{1i} - Y_{0i} | D_i = 1\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\text{Y}_{1i}|\text{X}_i, \text{D}_i = 1\right) - \mathbb{E}\left(\text{Y}_{0i}|\text{X}_i, \text{D}_i = 1\right) | \text{D}_i = 1\right) && \text{(by LIE on } X_i\text{)} \\
&= \mathbb{E}\left(\mathbb{E}\left(\text{Y}_{1i}|\text{X}_i, \text{D}_i = 1\right) - \mathbb{E}\left(\text{Y}_{0i}|\text{X}_i, \text{D}_i = 0\right) | \text{D}_i = 1\right) && \text{(by CIA)} \\
&= \mathbb{E}\left(\delta_X | D_i = 1\right) && \text{(by def. of } \delta_X\text{)} \\
&= \frac{\mathbb{E}\left(\mathbf{1}\left(D_i = 1\right)\delta_X\right)}{\mathbb{P}\left(D_i = 1\right)} && \text{(def of conditional expectation)} \\
&= \frac{\mathbb{E}\left(\mathbb{E}\left(\mathbf{1}\left(D_i = 1\right)\delta_X | X_i\right)\right)}{\mathbb{P}\left(D_i = 1\right)} && \text{(LIE on } X_i\text{)} \\
&= \frac{\mathbb{E}\left(P(D_i = 1|X_i)\delta_X\right)}{\mathbb{E}\left(D_i\right)} && (D_i \text{ an indicator, } \delta_X \text{ constant})
\end{aligned}
$$

where

$$\delta_X \equiv \mathbb{E}\left(\mathrm{Y}_{1i}|\mathrm{X}_i, \mathrm{D}_i = 1\right) - \mathbb{E}\left(\mathrm{Y}_{0i}|\mathrm{X}_i, \mathrm{D}_i = 0\right)$$

Interpretation. This means δ_{TOT} is a *treatment-probability (conditional probability of treatment) weighted average of* δ_X, whose sample analogue is observable- sample mean between the treated and the control.

Remark. To calculate the conditional expectation, we use

$$\mathrm{E}(X|H) = \frac{\mathrm{E}\left(1_H X\right)}{P(H)} = \int_{\mathcal{X}} x \mathrm{d}P(x|H) \tag{2.3}$$

Regression Meets Matching Think of regression estimates parameter δ_R in

$$\mathrm{Y}_i = \sum_x d_{ix}\beta_x + \delta_R \mathrm{D}_i + \varepsilon_i$$

Subtract mean from d tilde (CEF)?

Then

$$\begin{aligned}
\delta_R &= \frac{\mathrm{Cov}\left(\mathrm{Y}_i, \tilde{\mathrm{D}}_i\right)}{V\left(\tilde{\mathrm{D}}_i\right)} \\
&= \frac{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i|\mathrm{X}_i\right]\right)\mathrm{Y}_i\right]}{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i|\mathrm{X}_i\right]\right)^2\right]} \\
&= \frac{E\left\{\left(\mathrm{D}_i - E\left[\mathrm{D}_i|\mathrm{X}_i\right]\right)E\left[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i\right]\right\}}{E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i|\mathrm{X}_i\right]\right)^2\right]} \\
&= \frac{E\left[\sigma_D^2\left(\mathrm{X}_i\right)\delta_X\right]}{E\left[\sigma_D^2\left(\mathrm{X}_i\right)\right]}
\end{aligned}$$

where $\sigma_D^2 \equiv E\left[\left(\mathrm{D}_i - E\left[\mathrm{D}_i|\mathrm{X}_i\right]\right)^2|\mathrm{X}_i\right]$ is the conditional variance of D_i on X_i.

Interpretation. This means δ_R is a *conditional-variance weighted (condition variance of treatment) average of* δ_X, whose sample analogue is observable- sample mean between the treated and the control. In particular,

$$\sigma_D^2 = P(D_i = 1|X_i)[1 - P(D_i = 1|X_i)]$$

Regression is a matchmaker!

2.2 Propensity Score

Propensity-Score Theorem Remember

$$\delta_{TOT} = \mathbb{E}\left(\delta_X|D_i = 1\right) \tag{2.4}$$

What if X_i is many/multi continuous? Then, cell specific δ_X does not make sense, which requires grouping or parametric assumptions.

Theorem 2.1 (Propensity-Score Theorem)**.** Suppose the CIA holds for Y_{ji}. That is $\{Y_{0i}, Y_{1i}\} \perp D_i|X_i$. Then $\{Y_{0i}, Y_{1i}\} \perp D_i|p(X_i)$.

Interpretation. The Propensity-Score Theorem says you only need to control for covariates that affect the *probability of treatment*. Alternatively, the only covariate you need to control for is the probability of treatment itself.

$$
\begin{aligned}
\mathbb{E}\left(\frac{Y_i D_i}{p(X_i)}\right) &= \mathbb{E}\left(\frac{(Y_{0i} + (Y_{1i} - Y_{0i})D_i)D_i}{p(X_i)}\right) \\
&= \mathbb{E}\left(\frac{Y_{1i} D_i}{p(X_i)}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\frac{Y_{1i} D_i}{p(X_i)} | X_i\right)\right) \\
&= \mathbb{E}\left(Y_{1i}\right)
\end{aligned}
$$

3 Instrument Variables

3.1 Wald Estimator

Setup. Use binary instrument Z_i to estimate model with one endogenous regressor and no covariates.

Causal Regression Model.

$$Y_i = \rho s_i + \eta_i \tag{3.1}$$

If S_i is an endogenous regressor, then η_i and S_i may be correlated. Suppose Z_i is a binary instrument that equals 1 with probability p.

Claim 3.1 (Wald Estimator). In the causal regression model above, we have that

$$\rho = \frac{\mathbb{E}\left(Y_i | Z_i = 1\right) - \mathbb{E}\left(Y_i | Z_i = 0\right)}{\mathbb{E}\left(S_i | Z_i = 1\right) - \mathbb{E}\left(S_i | Z_i = 0\right)} \tag{3.2}$$

We drop the i subscripts and using lowercase for ease of notation.

Proof by direct calculation. We have that

$$
\begin{aligned}
\operatorname{Cov}(y, z) &= \mathbb{E}(yz) - \mathbb{E}(y)\,\mathbb{E}(z) \\
&= p\,\mathbb{E}(yz|z=1) + (1-p)\,\mathbb{E}(yz|z=0) - p(p\,\mathbb{E}(y|z=1) + (1-p)\,\mathbb{E}(y|z=0)) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(law of total expectation)} \\
&= p\,\mathbb{E}(y_1) - p^2\,\mathbb{E}(y_1) + p(1-p)\,\mathbb{E}(y_0) \\
&= p(1-p)\left(\mathbb{E}(y_1) - \mathbb{E}(y_0)\right)
\end{aligned}
$$

Similarly

$$\operatorname{Cov}(s, z) = p(1-p)\left(\mathbb{E}(s_1) - \mathbb{E}(s_0)\right)$$

Then

$$
\begin{aligned}
\operatorname{Cov}(y, z) &= \operatorname{Cov}(\rho s + \eta, z) \\
&= \rho \operatorname{Cov}(s, z)
\end{aligned}
$$

Therefore

$$\rho = \frac{\operatorname{Cov}(y, z)}{\operatorname{Cov}(s, z)} = \frac{\mathbb{E}(y_1) - \mathbb{E}(y_0)}{\mathbb{E}(s_1) - \mathbb{E}(s_0)} \tag{3.3}$$

□

Proof using exclusion restriction. The exclusion restriction gives us that

$$\mathbb{E}\left(\eta|z=0\right) \tag{3.4}$$

Therefore

$$\begin{aligned}\mathbb{E}\left(y|z\right) &= \rho\,\mathbb{E}\left(s|z\right)+\mathbb{E}\left(\eta|z\right)\\ &= \rho\,\mathbb{E}\left(s|z\right)\end{aligned}$$

So that we obtain the same ρ as above.

Does this proof work with an intercept?

□

3.2 2SLS

$$\begin{aligned}S_i &= X_i'\pi_{10}+\pi_{11}Z_i+\xi_{1i} && \text{(First-Stage)}\\ Y_i &= X_i'\pi_{20}+\pi_{21}Z_i+\xi_{2i} && \text{(Reduced Form)}\end{aligned}$$

$$Y_i=\alpha' X_i+\rho S_i+\eta_i \qquad \text{(Causal Relation of Interest (ILS))}$$

Substitute the *first stage* into the *causal relation of interest* to get

$$\begin{aligned}Y_i &= \alpha' X_i+\rho\left(X_i'\pi_{10}+\pi_{11}Z_i+\xi_{1i}\right)+\eta_i\\ &= X_i'\left(\alpha+\rho\pi_{10}\right)+\rho\pi_{11}Z_i+\left(\rho\xi_{1i}+\eta_i\right)\\ &= X_i'\pi_{20}+\pi_{21}Z_i+\xi_{2i}\end{aligned}$$

First Stage:

$$\begin{aligned}S_i &= X_i'\pi_{10}+\pi_{11}'Z_i+\xi_{1i}\\ \hat{S}_i &= X_i'\hat{\pi}_{10}+\hat{\pi}_{11}'Z_i && \text{(OLS fitted values)}\end{aligned}$$

Second Stage: The coefficient on $\hat{S}_i$ in the regression of Y_i on X_i and $\hat{S}_i$ is th 2SLS estimator of ρ.

$$Y_i=\alpha' X_i+\rho\hat{S}_i+\left[\eta_i+\rho\left(S_i-\hat{S}_i\right)\right] \tag{3.5}$$

The estimator for ρ is consistent since

- The first-stage are consistent (OLS).
- The covariates X_i and instruments Z_i are uncorrelated with η_i and $S_i-\hat{S}_i$ (by properties of regression residuals: regression residuals are uncorrelated with the regressors that made them).

Theorem 3.1. 2SLS is the same as IV, where the instrument is $\hat{S}_i^*$ (the residual from a regression of $\hat{S}_i$ on X_i).

Proof. The IV estimator of ρ is

$$\rho_{IV}=\frac{\operatorname{Cov}\left(Y_i,\hat{S}_i^*\right)}{\operatorname{Cov}\left(S_i,\hat{S}_i^*\right)} \tag{3.6}$$

□

We can simplify the denominator as

$$\operatorname{Cov}\left(S_i,\hat{S}_i^*\right)=\operatorname{Cov}\left(\hat{S}_i+\xi_{1i},\hat{S}_i^*\right) \qquad \left(S_i=\hat{S}_i+\xi_{1i}\right)$$

$$
\begin{aligned}
&= \text{Cov}\left(X_i'\pi^* + \hat{S}_i^* + \xi_{1i}, \hat{S}_i^*\right) && (\hat{S}_i = X_i'\pi^* + \hat{S}_i^*) \\
&= \underbrace{\text{Cov}\left(X_i'\pi^*, \hat{S}_i^*\right)}_{=0} + \text{Cov}\left(\hat{S}_i^*, \hat{S}_i^*\right) + \underbrace{\text{Cov}\left(\xi_{1i}, \hat{S}_i^*\right)}_{=0} \\
&= \text{Var}\left(\hat{S}_i^*\right)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\rho_{IV} &= \frac{\text{Cov}\left(Y_i, \hat{S}_i^*\right)}{\text{Cov}\left(S_i, \hat{S}_i^*\right)} \\
&= \frac{\text{Cov}\left(Y_i, \hat{S}_i^*\right)}{\text{Var}\left(\hat{S}_i^*\right)} \\
&= \rho_{2SLS}
\end{aligned}
$$

because ρ_{2SLS} is the regression coefficient of $\hat{S}_i$ in a regression of Y_i on $\hat{S}_i$:

$$
Y_i = \alpha' X_i + \rho \hat{S}_i + \eta_i \tag{3.7}
$$

Theorem 3.2. One-instrument 2SLS equals IV, where the instrument is $\tilde{Z}_i$ (the residual from a regression of Z_i on the covariates X_i).

Proof. We have that

$$
S_i = X_i'\pi_{10} + \pi_{11} Z_i + \xi_{1i} \qquad \text{(First-Stage)}
$$

By regression anatomy (Frisch-Waugh), we can estimate π_{11} by regression S_i with X_i partialled out (that is, $\hat{S}_i^*$) on Z_i with the X_i partialled out (that is, $\tilde{Z}_i$). This gives that

$$
\hat{S}_i^* = \tilde{Z}_i \hat{\pi}_{11} \tag{3.8}
$$

□

3.3 Grouped Data and 2SLS

(i) 2SLS using dummy instruments is the same thing as GLS on a set of group means.

(ii) GLS in turn can be understood as a linear combination of all the Wald estimators that can be constructed from pairs of means.

3.4 Heterogeneous Effects

Theorem 3.3 (LATE Theorem)**.** Suppose, for all i,

(i) Independence: $\{Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}\} \perp Z_i$

(ii) Exclusion: $Y_i(d, 0) = Y_i(d, 1) \equiv Y_{di}$ for $d = 0, 1$

(iii) First-Stage: $\mathbb{E}\left(D_{1i} - D_{0i}\right) \neq 0$

(iv) Monotonicity: $D_{1i} - D_{0i} \geq 0$ or vice versa

Then

$$
\frac{\mathbb{E}\left(Y_i | Z_i = 1\right) - \mathbb{E}\left(Y_i | Z_i = 0\right)}{\mathbb{E}\left(D_i | Z_i = 1\right) - \mathbb{E}\left(D_i | Z_i = 0\right)} = \mathbb{E}\left(Y_{1i} - Y_{0i} | D_{1i} > D_{0i}\right) = \mathbb{E}\left(\rho_i | \pi_{1i} > 0\right) \tag{3.9}
$$

Proof. **Step 1: Write Y_i in terms of potential outcomes.** The Exclusion Restriction allows us to define potential outcomes indexed against treatment status using the single-index notation (Y_{1i}, Y_{0i}). We have that

$$Y_{1i} \equiv Y_i(1,1) = Y_i(1,0)$$
$$Y_{0i} \equiv Y_i(0,1) = Y_i(0,0)$$

Therefore Y_i can be written in terms of potential outcomes by

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

Step 2: Use this equation for Y_i, D_i to rewrite the Wald estimator.
We can simplify

$$\begin{aligned} \mathbb{E}\left(Y_i|Z_i = 1\right) &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 1\right) && \text{(Exclusion Restriction)} \\ &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_{1i}|Z_i = 1\right) \\ &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_{1i}\right) && \text{(Independence)} \\ \mathbb{E}\left(Y_i|Z_i = 0\right) &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_{0i}\right) \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{E}\left(Y_i|Z_i = 1\right) - \mathbb{E}\left(Y_i|Z_i = 0\right) &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_{1i}\right) - \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_{0i}\right) \\ &= \mathbb{E}\left((Y_{1i} - Y_{0i})(D_{1i} - D_{0i})\right) \end{aligned}$$

By monotonicity, we have that $D_{1i} - D_{0i}$ equals 1 (that is $D_{1i} > D_{0i}$) or equals 0 (that is, $D_{1i} = D_{0i}$). Therefore, by the law of total expectation,

$$\mathbb{E}\left((Y_{1i} - Y_{0i})(D_{1i} - D_{0i})\right) = \mathbb{E}\left(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}\right)\mathbb{P}\left(D_{1i} > D_{0i}\right) \tag{3.10}$$

Step 3: Use independence to simplify the Wald denominator.

$$\begin{aligned} \mathbb{E}\left(D_i|Z_i = 1\right) - \mathbb{E}\left(D_i|Z_i = 0\right) &= \mathbb{E}\left(D_{1i}|Z_i = 1\right) - \mathbb{E}\left(D_{0i}|Z_i = 0\right) \\ &= \mathbb{E}\left(D_{1i} - D_{0i}\right) && \text{(Independence)} \\ &= \mathbb{P}\left(D_{1i} > D_{0i}\right) && (D_{1i} - D_{0i} \text{ either 1 or 0 by monotonicity}) \end{aligned}$$

Step 4: Combine the Pieces.

$$\begin{aligned} \frac{\mathbb{E}\left(Y_i|Z_i = 1\right) - \mathbb{E}\left(Y_i|Z_i = 0\right)}{\mathbb{E}\left(D_i|Z_i = 1\right) - \mathbb{E}\left(D_i|Z_i = 0\right)} &= \frac{\mathbb{E}\left(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}\right)\mathbb{P}\left(D_{1i} > D_{0i}\right)}{\mathbb{P}\left(D_{1i} > D_{0i}\right)} \\ &= \mathbb{E}\left(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}\right) \\ &= \mathbb{E}\left(\rho_i|\pi_{1i} > 0\right) \end{aligned}$$

$\square$

Interpretation. An instrument which is

(i) As good as randomly assigned

(ii) Affects the outcome through a single known channel

(iii) Has a first-stage

(iv) Affects the causal channel of interest in only one direction can be used to estimate the average causal effect on the affected group.

3.4.1 Failures of Monotonicity

When the monotonicity assumption fails, the instrument pushes some people into treatment while pushing others out of treatment. Those who are pushed out of treatment are called **defiers**. Now, without monotonicity, $D_{1i} - D_{0i} \in \{1, 0, -1\}$.

Therefore, by the law of total expectation,

$$\begin{aligned}
\mathbb{E}\left((Y_{1i} - Y_{0i})(D_{1i} - D_{0i})\right) &= \mathbb{E}\left((Y_{1i} - Y_{0i})|D_{1i} > D_{0i}\right)\mathbb{P}\left(D_{1i} > D_{0i}\right) + \mathbb{E}\left((Y_{1i} - Y_{0i}) \times -1|D_{1i} < D_{0i}\right)\mathbb{P}\left(D_{1i} < D_{0i}\right) \\
&= \mathbb{E}\left(Y_{1i} - Y_{0i}|D_{1i} > D_{0i}\right)\mathbb{P}\left(D_{1i} > D_{0i}\right) - \mathbb{E}\left(Y_{1i} - Y_{0i}|D_{1i} < D_{0i}\right)\mathbb{P}\left(D_{1i} < D_{0i}\right)
\end{aligned}$$

Thus, it could be that the treatment effects are positive for everyone, but the reduced form is zero because the effects on compliers are canceled out by the effects on the defiers.

3.4.2 IV in Randomized Trials

Theorem 3.4 (Bloom Result: IV estimates the effect of the treatment on the treated in a randomized trial with one-sided non-compliance)**.** Suppose the assumption of the LATE Theorem (3.3) hold and $\mathbb{E}\left(D_i|Z_i = 0\right) = 0$ (that is, there is one-sided non-compliance). Then

$$\begin{aligned}
\frac{\mathbb{E}\left(Y_i|Z_i = 1\right) - \mathbb{E}\left(Y_i|Z_i = 0\right)}{\mathbb{E}\left(D_i|Z_i = 1\right) - \mathbb{E}\left(D_i|Z_i = 0\right)} &= \frac{\mathbb{E}\left(Y_i|Z_i = 1\right) - \mathbb{E}\left(Y_i|Z_i = 0\right)}{\mathbb{E}\left(D_i|Z_i = 1\right)} \\
&= \frac{\text{ITT Effect}}{\text{Compliance Rate}} \\
&= \mathbb{E}\left(Y_{1i} - Y_{0i}|D_i = 1\right)
\end{aligned}$$

Proof. As in the proof of the LATE theorem (with the additional simplification given by one-sided non-compliance)

$$\begin{aligned}
\mathbb{E}\left(Y_i|Z_i = 1\right) &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 1\right) \\
\mathbb{E}\left(Y_i|Z_i = 0\right) &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 0\right) \\
&= \mathbb{E}\left(Y_{0i}|Z_i = 0\right) && \text{(since } \mathbb{E}\left(D_i|Z_i = 0\right) = 0\text{)}
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}\left(Y_i|Z_i = 1\right) - \mathbb{E}\left(Y_i|Z_i = 0\right) &= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 1\right) - \mathbb{E}\left(Y_{0i}|Z_i = 0\right) \\
&= \mathbb{E}\left(Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 1\right) - \mathbb{E}\left(Y_{0i}|Z_i = 1\right) && \text{(Independence)} \\
&= \mathbb{E}\left((Y_{1i} - Y_{0i})D_i|Z_i = 1\right) \\
&= \mathbb{E}\left((Y_{1i} - Y_{0i})D_i|D_i = 1, Z_i = 1\right)\mathbb{P}\left(D_i = 1|Z_i = 1\right) \\
&= \mathbb{E}\left(Y_{1i} - Y_{0i}|D_i = 1\right)
\end{aligned}$$

since $\mathbb{E}\left(D_i|Z_i = 0\right) = 0$ means $D_i = 1$ implies $Z_i = 1$.

□

3.5 2SLS Mistakes

3.5.1 Manual 2SLS

Manual 2SLS proceeds as

(i) Estimate the first stage by OLS

(ii) Plug the fitted values into the second stage equation, and estimate by OLS

$$\begin{aligned} S_i &= X_i'\pi_{10} + \pi_{11}'Z_i + \xi_{1i} \\ Y_i &= \alpha' X_i + \rho \hat{s}_i + \left[\eta_i + \rho\left(S_i - \hat{S}_i\right)\right] \end{aligned} \tag{3.11}$$

where

(i) X_i is a set of covariates.

(ii) Z_i is a set of excluded instruments.

(iii) The first stage fitted values are $\hat{S}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11}'Z_i$.

The OLS residual variance is the variance of $\eta_i + \rho(S_i - \hat{S}_i)$. The proper 2SLS standard errors include the variance of η_i only.

3.5.2 Covariate Ambivalence

3.5.3 Forbidden Regressions

4 Regression Discontinuity Designs

4.1 Sharp RD

Sharp RD is used when treatment status is a *deterministic and discontinuous* function of a covariate x_i, called the *running variable*. For example, $D_i = \mathbf{1}\left(x_i \geq x_0\right)$.

Model (RD: Potential Outcomes with Linear, Constant-Effects)**.**

No OVB

4.2 Fuzzy RD is IV

Fuzzy RD exploits discontinuities in the probability or expected value of treatment conditional on a covariate. Then, the discontinuity becomes an instrumental variable for treatment status instead of deterministically switching treatment on or off.

5 Inference

6 Machine Labor