Section 5: Measurement Error, Bad Controls

Valentine Gilbert

October 7, 2021

Overview

Classical Measurement Error Results

In lecture we discussed a number of results regarding **classical measurement error**:

1. Mismeasurement in the **independent** variable ($X_i = X_i^* + m_i$) leads to attenuation bias
   - Intuition: Imagine if you had so much mismeasurement that your independent variable was essentially a random number
2. **Controls** exacerbate attenuation bias due to mismeasurement in the independent variable
   - Intuition: Residualizing X_i absorbs variation in X_i^* but not in m_i
3. Mismeasurement in the **dependent** variable ($Y_i = Y_i^* + m_i$) doesn't lead to bias but leads to larger standard errors
   - Intuition: The CEF is the same ($E[Y_i|X_i] = E[Y_i^*|X_i]$), so we don't get bias, but mismeasurement creates additional sampling variability
4. Mismeasurement in a control variable ($W_i = W_i^* + m_i$) leads to **contamination bias** in the coefficient on the independent variable
   - Intuition: Imagine if the control variable was pure noise – then the coefficient on X_i^* would suffer from OVB

Controls Exacerbate Attenuation Bias: Setup

Let's prove that **attenuation bias is exacerbated by controls**:

- We'd like to estimate the following regression: $Y_i = \beta_0 + \beta_1 X_i^* + \beta_2 W_i + u_i$
- Under classical measurement error, we observe $X_i = X_i^* + m_i$, where

$$cov(X_i^*, m_i) = 0 \qquad cov(u_i, m_i) = 0 \qquad cov(W_i, m_i) = 0 \qquad cov(Y_i, m_i) = 0$$

- Due to mismeasurement, we can only estimate $Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 W_i + v_i$
- How does α_1 relate to β_1? We'll use the regression anatomy theorem to show that:

$$\alpha_1 = \beta_1 \frac{var(\tilde{X}_i^*)}{var(\tilde{X}_i^*) + var(m_i)}$$

Controls Exacerbate Attenuation Bias: Residual Regressions

Consider the following two **residualizing regressions**:

$$X_i = \pi_0 + \pi_1 W_i + \tilde{X}_i$$
$$X_i^* = \pi_0 + \pi_1 W_i + \tilde{X}_i^*$$

Question: Why do these two regressions have the same coefficients?

- Because we now have classical measurement error in the dependent variable, which doesn't bias our regression estimates

Notice that $\tilde{X}_i = \tilde{X}_i^* + m_i$

- We'll use this identity on the next slide to complete the proof

Controls Exacerbate Attenuation Bias: Proof

We know from the regression anatomy theorem that

$$\beta_1 = \frac{cov(Y_i, \tilde{X}_i^*)}{var(\tilde{X}_i^*)} \qquad \alpha_1 = \frac{cov(Y_i, \tilde{X}_i)}{var(\tilde{X}_i)}$$

Using the identity $\tilde{X}_i = \tilde{X}_i^* + m_i$ from the previous slide, we get

$$\alpha_1 = \frac{cov(Y_i, \tilde{X}_i^* + m_i)}{var(\tilde{X}_i^*) + var(m_i)} = \frac{cov(Y_i, \tilde{X}_i^*)}{var(\tilde{X}_i^*) + var(m_i)}$$

Question: How can I make β_1 appear? Multiply by $1 = var(\tilde{X}_i^*)/var(\tilde{X}_i^*)$

$$\alpha_1 = \beta_1 \frac{var(\tilde{X}_i^*)}{var(\tilde{X}_i^*) + var(m_i)}$$

Bad Controls 1/3: Example

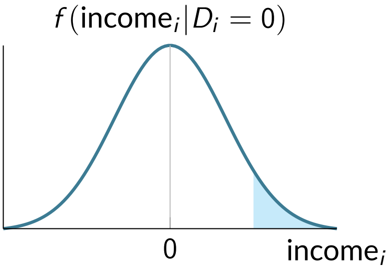
Let's consider the Tennessee STAR experiment again

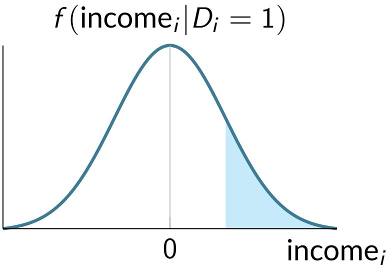
- Recall that students were randomized to large and small classrooms within school
- Why does this help us with causal inference?
  - Now treatment is independent of student characteristics, including unobservable characteristics like ability
- We can estimate the effect of classroom size on various outcomes
  - Test scores
  - College attendance
  - Earnings as an adult
- What if we want to know how much of the effect on earnings is due to increased college attendance?
- Could we regress earnings on classroom size and control for college attendance to understand the mechanism by which classroom size affects earnings?

Bad Controls 2/3: Example

No! College attendance is a **bad control** because it is affected by the treatment

- How would our estimate from the long regression be biased?
- Suppose the treatment allows lower income students to attend college:





- The blue shaded region shows which students go to college.
- How do treated and untreated students compare, conditional on college attendance?

Bad Controls 3/3: Potential Outcomes

We can use potential outcomes to understand the **bias** introduced by bad controls

- Let D_i indicate treatment status, C_i indicate college attendance, and Y_i be earnings
- Let $\{C_i(1), C_i(0)\}$ and $\{Y_i(1), Y_i(0)\}$ be potential outcomes
- Consider the difference in earnings for treated and untreated students among those who went to college:

$$
\begin{aligned}
& E[Y_i|D_i = 1, C_i = 1] - E[Y_i|D_i = 0, C_i = 1] \\
& = E[Y_i(1)|D_i = 1, C_i(1) = 1] - E[Y_i(0)|D_i = 0, C_i(0) = 1] \\
& = E[Y_i(1)|C_i(1) = 1] - E[Y_i(0)|C_i(0) = 1] \\
& = \underbrace{E[Y_i(1) - Y_i(0)|C_i(1) = 1]}_{\text{ATE among college attendees}} + \underbrace{E[Y_i(0)|C_i(1) = 1] - E[Y_i(0)|C_i(0) = 1]}_{\text{selection bias}}
\end{aligned}
$$

- In what direction do you think the selection bias goes?

Inference with Linear Combinations of Coefficients 1/2: Polynomials

Consider this regression equation: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$

- Why might you include polynomials in a regression?
  - If you think the relationship between Y_i and X_i is non-linear
  - If you think mean independence only holds conditional on a non-linear function of control variables
- What is the estimated effect of X_i on Y_i if $X_i = x$?
- $\partial Y_i / \partial X_i \big|_{X_i = x} = \hat{\beta}_1 + 2\hat{\beta}_2 x$
- What is the standard error of this estimated effect?
- $\sqrt{var(\hat{\beta}_1 + 2\hat{\beta}_2 x)} = \sqrt{var(\hat{\beta}_1) + 4x^2 var(\hat{\beta}_2) + 4x cov(\hat{\beta}_1, \hat{\beta}_2)}$
- Why do we need the covariance term? What's the intuition?

Inference with Linear Combinations of Coefficients 2/2: Interaction Effects

- Often you may be interested in heterogeneity of treatment effects
- You may wonder, for example, if an intervention has different effects for men and women. How could you investigate this?
- With interaction effects: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 \text{Male}_i + \beta_3 D_i \times \text{Male}_i + u_i$
- How do I interpret each coefficient?
- What is the standard error of the estimated treatment effect on men?
  - $\sqrt{var(\hat{\beta}_1 + \hat{\beta}_3)} = \sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_3) + 2cov(\hat{\beta}_1, \hat{\beta}_3)}$
- Intuition for covariance: Random samples that happen to give large estimates of β_1 may also tend to give large (or small) estimates of β_3

OVB Practice

The most important formula in econometrics
- Bruich, Chamberlain, Feldstein

- What's a causal relationship of interest?
- What's the naïve (short) regression?
- What would we like to control for?
- How will our naïve estimate be biased?

Regression Anatomy

Consider the regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

- What is the regression anatomy result?
- $\beta_1 = \mathrm{Cov}(Y_i, \tilde{X}_i) / \mathrm{Var}(\tilde{X}_i)$
- What does this say in words?
- What is the matrix extension of this result?
- If $\boldsymbol{y} = \boldsymbol{X}_1 \beta_1 + \boldsymbol{X}_2 \beta_2 + \boldsymbol{u}$ then $\beta_1 = (\boldsymbol{X}'_{1\perp 2} \boldsymbol{X}_{1\perp 2})^{-1} \boldsymbol{X}'_{1\perp 2} \boldsymbol{y}$

Question 5

- “Show how to derive the heteroskedasticity robust standard error formula for a regression with an intercept and a binary independent variable.”
- $Y_i = \beta_0 + \beta_1 X_i + u_i$
- What is the interpretation of β_1?
- $\beta_1 = \mathrm{E}(Y_i|X_i = 1) - \mathrm{E}(Y_i|X_i = 0)$
- How do we estimate β_1?
- $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$, where $\bar{Y}_j = \frac{1}{N_j}\sum_{i:X_i=j} Y_i$
- What is the heteroskedasticity robust standard error of $\hat{\beta}_1$?

$$
\begin{aligned}
SE(\hat{\beta}_1) &= \sqrt{\mathrm{Var}(\hat{\beta}_1)} \\
&= \sqrt{\mathrm{Var}(\bar{Y}_1) + \mathrm{Var}(\bar{Y}_0)} \qquad \text{(What assumption did I make?)} \\
&= \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}
\end{aligned}
$$

Questions 21 and 22

- "With discrete regressors, least-squares regression is just a bunch of sample means." —Discuss.
- "With discrete regressors, functional form is not a problem." —Discuss.