Dis 7: Internal vs. External Validity

1 Overview

- Definitions:
  - **Internal validity**: Inferences about the causal effects are valid for the population being studied.
  - **External validity**: Inferences and conclusions generated from this one study can be generalized to other populations and settings.
- When does internal validity NOT hold?
  - **Problem with $\hat{\beta}$**: OLS estimator $\hat{\beta}$ is biased ($E[\hat{\beta}] \neq \beta$) or inconsistent ($\hat{\beta} \overset{p}{\nrightarrow} \beta$ when sample size goes to infinity).

    How this type of problem can manifest:
    - **Omitted variable bias**
    - **Model functional form misspecification**
      (Example: Quadratic term of x_1, x_1^2, should have entered the regression model, but a misspecified model excluded x_1^2.)
    - **Errors in variables (include measurement error)**
      (more on this in section 2)
    - **Missing data and sample selection bias**: If the sample isn't randomly selected, or if missing data points aren't really missing "at random", but rather missing with some sort of pattern, then we wouldn't expect the $\hat{\beta}$ estimate to come close to the true β value.
    - **Simultaneous causality**: X causes Y, but Y also causes X
      (Example: Let Y be quantity sold of certain good, and X be its price. Price determines how many units this good can be sold, but quantity also affects price – if some good isn't selling well, price might be reduced to boost demand.)
  - **Problem with $se(\hat{\beta})$**: Hypothesis tests and confidence intervals cannot be trusted when standard error of $\hat{\beta}$ is incorrect.

    How this type of problem can manifest:
    - **Incorrectly specified error term**
      (homoskedasticity vs. heteroskedasticity from last week's section)
    - **Violation of i.i.d. data**: If the variables are not independently distributed across observations (most commonly, there's serial correlation across observations, which can arise in panel and time series data), then some adjustment in standard error is needed.
- When does external validity NOT hold?
  - **Differences in populations**
    (Example: A study done on college students probably won't generalize to older adults; A study done in the U.S. probably won't generalize to other countries.)
  - **Differences in settings**
    (Example: A study on what factors affect economic student's GPA probably won't generalize across universities. Even if the college population is similar between schools, course offering & level of GPA inflation might still differ across colleges.)

2 Errors in variables

- Suppose we are interested in the relationship between weight and calorie intake:

  $$\text{weight}_i = \beta_0 + \beta_1 \text{calories eaten}_i + u_i$$

  Both calories eaten (x) and weight (y) could be measured incorrectly.

- **Classical measurement error**

  If the measurement error is independent from x, y, and the unobserved model error u, then this type of error is called **classical measurement error**.

  - **Classical measurement error in independent variable**:

    $$\underbrace{e_i}_{\text{measurement error}} = \underbrace{\widetilde{x}_i}_{\text{observed value}} - \underbrace{x_i}_{\text{true value}} \qquad \Rightarrow \qquad \widetilde{x}_i = x_i + e_i$$

    This means that if we regress the true y on the observed $\widetilde{x}$, our estimated model becomes

    $$\begin{aligned} y &= \beta_0 + \beta_1 \widetilde{x} + u_{\text{est}} \\ &= \beta_0 + \beta_1 (x + e) + u_{\text{est}} \\ &= \beta_0 + \beta_1 x + \underbrace{(\beta_1 e + u_{\text{est}})}_{u_{\text{true}}} \end{aligned} \tag{1}$$

    Recall from Dis 4 on how bias is measured:

    $$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\rho_{xu} \frac{\sigma_u}{\sigma_x}}_{\text{bias}} \tag{2}$$

    where x in equation (2) is the explanatory variable included in the model with β_1 being its coefficient, and u is the error term in this model.

    Based on (1), the explantory variable to use here is $\widetilde{x}$, and the error term in the model is u_{est}. So the bias becomes

    $$\begin{aligned} \text{bias} = \rho_{\widetilde{x} u_{\text{est}}} \frac{\sigma_{u_{\text{est}}}}{\sigma_{\widetilde{x}}} &= \frac{Cov(\widetilde{x}, u_{\text{est}})}{Var(\widetilde{x})} \\ &= \frac{Cov(x + e, u_{\text{true}} - \beta_1 e)}{Var(x + e)} \\ &= \frac{Cov(x, u_{\text{true}}) - \beta_1 Cov(x, e) + Cov(e, u_{\text{true}}) - \beta_1 Cov(e, e)}{Var(x) + Var(e) + 2Cov(x, e)} \\ &= \frac{0 - \beta_1 \times 0 + 0 - \beta_1 Var(e)}{Var(x) + Var(e) + 0} \\ &= -\beta_1 \frac{\sigma_e^2}{\sigma_x^2 + \sigma_e^2} \end{aligned}$$

Plug this back into equation (2):

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \frac{\sigma_e^2}{\sigma_x^2 + \sigma_e^2} = \left(1 - \frac{\sigma_e^2}{\sigma_x^2 + \sigma_e^2}\right)\beta_1 = \underbrace{\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}\right)}_{\text{attenuation bias}}\beta_1$$

How to interpret the attenuation bias:

- Consider σ_x^2 as the **signal**, and σ_e^2 as the **noise**.
- If there's no noise in the observed data, then no bias exists.
- If the observed data is full of noise, then $\hat{\beta}_1 \xrightarrow{p} 0$ (⇒ signal does not help identify the effect from x onto y anymore).

- **Classical measurement error in dependent variable**:

$$\underbrace{e_i}_{\text{measurement error}} = \underbrace{\widetilde{y}_i}_{\text{observed value}} - \underbrace{y_i}_{\text{true value}} \qquad \Rightarrow \qquad \widetilde{y}_i = y_i + e_i$$

This means that if we regress the observed $\widetilde{y}$ on the true x, our estimated model becomes

$$\begin{aligned} \widetilde{y} &= \beta_0 + \beta_1 x + u_{\text{est}} \\ y + e &= \beta_0 + \beta_1 x + u_{\text{est}} \\ y &= \beta_0 + \beta_1 x + \underbrace{(u_{\text{est}} - e)}_{u_{\text{true}}} \end{aligned}$$

So the bias becomes

$$\begin{aligned} \text{bias} = \rho_{xu_{\text{est}}} \frac{\sigma_{u_{\text{est}}}}{\sigma_x} &= \frac{Cov(x, u_{\text{est}})}{Var(x)} \\ &= \frac{Cov(x, u_{\text{true}} + e)}{Var(x)} \\ &= \frac{Cov(x, u_{\text{true}}) + Cov(x, e)}{Var(x)} \\ &= \frac{0 + Cov(x, e)}{Var(x)} \end{aligned}$$

As long as this is truly classical error (⇒ e is independent from x), then bias $= 0$.

Intuitively, think about this as y is measured randomly incorrectly, and the randomness part is captured by the original error term u term anyway, so this has no impact on the coefficient estimate.

- **Other types of errors in variables**

Classical measurement error is the most common type of errors in variables, but there are other ways where your variables can be erroneously recorded.

This week's problem set (PS 6) question 2 deals with one scenario: some data is randomly scrambled.

Say that 20% of x is randomly scrambled – that is, for 20% of x, the position of the data has changed (notice that the actual points aren't lost though).

Denote the observed x as $\widetilde{x}$, then

$$\widetilde{x}_i = \begin{cases} x_i & \text{for 80\% of } i \\ x_j & \text{for 20\% of } i \text{ and } j \neq i \end{cases}$$

This means for n number of observations in total, $.8n$ of $\widetilde{x}$ records the correct x_i observations, but $.2n$ of the $\widetilde{x}$ records the incorrectly positioned x_j.

So the β_1 (coefficient on the scrambled $\widetilde{x}$) estimate is

$$\begin{align}
\hat{\beta}_1 &= \frac{\widehat{Cov}(\widetilde{x}_i, y_i)}{\widehat{Var}(\widetilde{x}_i)} \notag \\
&= \frac{\widehat{Cov}(\widetilde{x}_i, \beta_0 + \beta_1 x_i + u_i)}{\widehat{Var}(\widetilde{x}_i)} && \text{(since } y_i = \beta_0 + \beta_1 x_i + u_i \text{ from the true model)} \notag \\
&= \beta_1 \frac{\widehat{Cov}(\widetilde{x}_i, x_i)}{\widehat{Var}(\widetilde{x}_i)} + \frac{\widehat{Cov}(\widetilde{x}_i, u_i)}{\widehat{Var}(\widetilde{x}_i)} \notag \\
&= \beta_1 \frac{\widehat{Cov}(\widetilde{x}_i, x_i)}{\widehat{Var}(\widetilde{x}_i)} + 0 = \beta_1 \frac{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_i x_i) - \overline{\widetilde{x}}\overline{x}}{\widehat{Var}(\widetilde{x}_i)} && \text{(each } \widetilde{x}_i \text{ is still uncorrelated with the true error)} \notag \\
&= \beta_1 \frac{\frac{1}{n}\left[\sum_{i=1}^{.8n}(x_i x_i) + \sum_{i=1}^{.2n}(x_j x_i)\right] - \frac{1}{n}\left[\sum_{i=1}^{.8n} x_i + \sum_{i=1}^{.2n} x_j\right]\overline{x}}{\widehat{Var}(\widetilde{x})} \notag \\
&= \beta_1 \frac{\frac{1}{n}\sum_{i=1}^{.8n}(x_i x_i) + \frac{1}{n}\sum_{i=1}^{.2n}(x_j x_i) - \overline{x}\frac{1}{n}\sum_{i=1}^{.8n} x_i - \overline{x}\frac{1}{n}\sum_{i=1}^{.2n} x_j}{\widehat{Var}(\widetilde{x})} \notag \\
&= \beta_1 \frac{.8\frac{1}{.8n}\sum_{i=1}^{.8n}(x_i x_i) + .2\frac{1}{.2n}\sum_{i=1}^{.2n}(x_j x_i) - \overline{x}.8\frac{1}{.8n}\sum_{i=1}^{.8n} x_i - \overline{x}.2\frac{1}{.2n}\sum_{i=1}^{.2n} x_j}{\widehat{Var}(\widetilde{x})} \notag \\
&= \beta_1 \frac{.8\left[\frac{1}{.8n}\sum_{i=1}^{.8n}(x_i x_i) - \overline{x}\frac{1}{.8n}\sum_{i=1}^{.8n} x_i\right] + .2\left[\frac{1}{.2n}\sum_{i=1}^{.2n}(x_j x_i) - \overline{x}\frac{1}{.2n}\sum_{i=1}^{.2n} x_j\right]}{\widehat{Var}(\widetilde{x})} \notag \\
&= \beta_1 \frac{.8\widehat{Cov}(x_i, x_i) + .2\widehat{Cov}(x_j, x_i)}{\widehat{Var}(\widetilde{x})} \notag \\
&= \beta_1 \frac{.8\widehat{Var}(x) + .2 \times 0}{\widehat{Var}(x)} \tag{3} \\
&= .8\beta_1 \notag
\end{align}$$

In (3), $\widehat{Cov}(x_j, x_i) = 0$ since data is i.i.d. And since only the position of 20% of x have changed (for example, consider this as we swapping x_{10} with x_{15} – the position of the data changed, but the data point isn't lost), the sum of all xs should still be the same, meaning that

$$\overline{\widetilde{x}} = \overline{x} \quad \text{and} \quad \frac{1}{n}\sum_{i}^{n}\widetilde{x}_i^2 = \frac{1}{n}\sum_{i}^{n} x_i^2 \quad \Rightarrow \quad \widehat{Var}(\widetilde{x}) \equiv \frac{1}{n}\sum_{i}^{n}\widetilde{x}_i^2 - \overline{\widetilde{x}}^2 = \frac{1}{n}\sum_{i}^{n} x_i^2 - \overline{x}^2 \equiv \widehat{Var}(x)$$

This thus implies that $E[\hat{\beta}_1] = .8\beta_1$, meaning that $\hat{\beta}_1$ is no longer an unbiased estimator in this case.

3 Problems

1. (Direction of bias)

   Consider the following true data generating process that satisfies all linear regression assumptions:

   $$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

   Suppose, however, that our estimated model is the following:

   $$y_i = \beta_0 + \beta_1 x_{1i} + v_i$$

   (a) Given $\beta_2 > 0$, $\beta_3 < 0$, $Corr(x_1, x_2) < 0$, and $Corr(x_1, x_3) = 0$, is $\hat{\beta}_1$ biased upward or downward in asymptotic?

   Our true error is from the data generating process u_i, and the error term from the estimated model is v_i. Since the estimated model didn't include x_2 and x_3 as explanatory variables,

   $$v_i = \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

   Using the bias formula in section 2 of this handout, we have

   $$\begin{aligned}\text{bias} = \frac{Cov(x_1, v)}{Var(x_1)} &= \frac{Cov(x_1, \beta_2 x_2 + \beta_3 x_3 + u)}{Var(x_1)} \\ &= \frac{\beta_2 Cov(x_1, x_2) + \beta_3 Cov(x_1, x_3) + Cov(x_1, u)}{\sigma_{x_1}^2} \\ &= \beta_2 \frac{Cov(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}} \frac{\sigma_{x_2}}{\sigma_{x_1}} + \beta_3 \frac{Cov(x_1, x_3)}{\sigma_{x_1}\sigma_{x_3}} \frac{\sigma_{x_3}}{\sigma_{x_1}} + \frac{0}{\sigma_{x_1}^2} \\ &= \underbrace{\beta_2}_{>0} \underbrace{Corr(x_1, x_2)}_{<0} \underbrace{\frac{\sigma_{x_2}}{\sigma_{x_1}}}_{>0} + \underbrace{\beta_3}_{<0} \underbrace{Corr(x_1, x_3)}_{=0} \underbrace{\frac{\sigma_{x_3}}{\sigma_{x_1}}}_{>0} \\ &< 0\end{aligned}$$

   This means that

   $$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \text{bias} = \beta_1 + \text{something negative}$$

   resulting in $\hat{\beta}_1$ biased downward in asymptotic.

   Side note: Notice that when $Corr(x_1, x_3) = 0$, the entire second term drops out, and the bias only depends on how x_2 affects y (sign on β_2), and how x_2 and x_1 is correlated (sign on $Corr(x_1, x_2)$). This is because when $Corr(x_1, x_3) = 0$, x_3 is no longer an omitted variable for x_1, so naturally it has no effect on β_1's estimate.

   (b) Continue to assume $\beta_2 > 0$, $\beta_3 < 0$, $Corr(x_1, x_2) < 0$, but now suppose that $Corr(x_1, x_3) > 0$. Is $\hat{\beta}_1$ now biased upward or downward in asymptotic?

   Using the bias formula we found in (a), except now $Corr(x_1, x_3) > 0$:

   $$\text{bias} = \underbrace{\beta_2}_{>0} \underbrace{Corr(x_1, x_2)}_{<0} \underbrace{\frac{\sigma_{x_2}}{\sigma_{x_1}}}_{>0} + \underbrace{\beta_3}_{<0} \underbrace{Corr(x_1, x_3)}_{>0} \underbrace{\frac{\sigma_{x_3}}{\sigma_{x_1}}}_{>0} < 0$$

With the bias being negative, $\hat{\beta}_1$ is still biased downward in asymptotic.

(c) Continue to assume $\beta_2 > 0$, $\beta_3 < 0$, $Corr(x_1, x_2) < 0$, but now suppose that $Corr(x_1, x_3) < 0$. Is $\hat{\beta}_1$ now biased upward or downward in asymptotic, or can you say anything about the bias?

Using the bias formula we found in (a), except now $Corr(x_1, x_3) > 0$:

$$\text{bias} = \underbrace{\beta_2}_{>0} \underbrace{Corr(x_1, x_2)}_{<0} \underbrace{\frac{\sigma_{x_2}}{\sigma_{x_1}}}_{>0} + \underbrace{\beta_3}_{<0} \underbrace{Corr(x_1, x_3)}_{<0} \underbrace{\frac{\sigma_{x_3}}{\sigma_{x_1}}}_{>0}$$

Now the first term in the bias formula is negative, but the second term in the bias formula is positive. Without knowing which term dominates the other, we cannot say anything about the bias now: $\hat{\beta}_1$ could be biased downward, upward, or has 0 bias. It all depends on how the negative and positive terms balance with each other.

2. (Classical measurement error in both x and y)

Consider the regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

It's determined that both x and y suffer from classical measurement error, and the true $\beta_1 < 0$. Would this bias $\hat{\beta}_1$ upward or downward in asymptotic?

Since both x and y suffer from classical measurement error, denote the observed values as $\widetilde{x}$ and $\widetilde{y}$ respectively, then

$$\widetilde{x}_i = x_i + a_i \quad \text{and} \quad \widetilde{y}_i = y_i + b_i$$

where a_i and b_i are independent of each other, and are independent of the true x_i, y_i, and u_i.

Our estimated model using the observed values is thus

$$\begin{aligned}
\widetilde{y}_i &= \beta_0 + \beta_1 \widetilde{x}_i + v_i \\
y_i + b_i &= \beta_0 + \beta_1 (x_i + a_i) + v_i \\
y_i &= \beta_0 + \beta_1 x_i + \underbrace{(v_i + \beta_1 a_i - b_i)}_{u_i}
\end{aligned}$$

Using the bias formula in section 2 of this handout, we have

$$\begin{aligned}
\text{bias} &= \frac{Cov(\widetilde{x}_i, v_i)}{Var(\widetilde{x}_i)} \\
&= \frac{Cov(x_i + a_i, u_i - \beta_1 a_i + b_i)}{Var(x_i + a_i)} \\
&= \frac{Cov(x_i, u_i) + Cov(x_i, -\beta_1 a_i) + Cov(x_i, b_i) + Cov(a_i, u_i) + Cov(a_i, -\beta_1 a_i) + Cov(a_i, b_i)}{Var(x_i) + Var(a_i) + 2Cov(x_i, a_i)} \\
&= \frac{0 + 0 + 0 + 0 + -\beta_1 Var(a_i) + 0}{Var(x_i) + Var(a_i)} \\
&= -\beta_1 \frac{\sigma_a^2}{\sigma_x^2 + \sigma_a^2}
\end{aligned}$$

which is identical to what we found in the case of only x having classical measurement error!

(Intuitively, this makes sense as the classical measurement error on y is naturally captured by the regression model's error term, so only the classical measurement error on x matters here.)

Now, since $\beta_1 < 0$, $-\beta_1 > 0$. And with $\frac{\sigma_a^2}{\sigma_x^2+\sigma_a^2} > 0$ as long as σ_a^2 is finite, then bias > 0, meaning that the estimator $\hat{\beta}_1$ is biased upward in asymptotic.