Review Session 4

Measurement Error and Causal Inference

Ben Berger

February 16, 2023

Today

- Measurement Error
- Intro to Causal Inference
- Review Exercise

Measurement Error

$$\tilde{X}_i = X_i + v_i$$

Measurement error in a variable X_i is equal to X_i plus some amount v_i that may or may not be correlated with X_i.

For example, we may be interested in whether an individual has a Covid-19 infection $Covid_i$, but instead we observe rapid/PCR test results $C\tilde{o}vid_i$ which are measured with some error.

Consider a person with undetected Covid-19:

- $Covid_i = 1$
- $C\tilde{o}vid_i = Covid_i + v_i = 0$
- $v_i = -1$

Measurement Error

Classical measurement error

- $Corr(X_i, v_i) = 0$.
- Measurement error is uncorrelated with X_i.
- Inflates standard errors if error in Y, but still unbiased in this case.
- Biases coefficient if error in X.

Non-Classical Measurement Error

- $Corr(X_i, v_i) \neq 0$.
- Generally will bias coefficient estimates regardless of whether it is in the dependent or independent variable.
- If in the independent variable, we can use instrumental variables if we have any.

Classical Measurement Error

Does classical measurement error in the **dependent** variable bias $\hat{\beta}_1$? No. If measurement error in Y_i is uncorrelated with Y_i, then on average the relationship between X_i and Y_i will be the same.

Does classical measurement error in the **independent** variable bias $\hat{\beta}_1$? Yes. Measurement error will *attenuate* $\hat{\beta}_1$ towards zero because it captures a weighted average of the true relationship and the non-relationship between measurement error and the dependent variable.

$$\hat{\beta}_1 \underset{p}{\rightarrow} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2}\beta_1$$

As the error variance σ_v^2 increases, the estimate is biased more towards zero.

Click Here for a Visual Example

Measurement Error: Example 1

Suppose you are piloting a population health intervention. You want to measure participants' total cholesterol; however, you cannot observe participants' true cholesterol level. Instead you can use a test from one of two companies that tests the cholesterol in a sample of participants' blood. ***BigMed*** *supplies a test that is accurate on average, but the test has an error with a standard deviation of 5 points.* ***Theracorp*** *also supplies a test that is accurate on average, but it uses only a tiny sample of blood. Therefore, there is substantially more variation from sample-to-sample in the result of Theracorp's test. The error distributions of the tests do not vary systematically with individuals' true cholesterol.*

Measurement Error: Example 1

1. Is this measurement error classical or non-classical?

Error uncorrelated with cholesterol →classical

2. Your boss wants to go ahead with Theracorp's test because it is a fraction of the cost of BigMed's test. She also claims that bias isn't an issue. Is her decision wise?

In this case, your boss is correct that bias isn't an issue. Because cholesterol is the dependent variable, measurement error will not bias the coefficient estimate, it will only increase its standard error. However, because Theracorp's test increases the standard error relative to BigMed's test, the study will require potentially many more participants to be adequately powered to precisely identify the effect. Recruiting these participants may be much more expensive than just using BigMed's test.

Measurement Error: Example 2

> *Now suppose you want to determine whether individuals have Covid-19 or not, but the best you can do is to observe the results of a test that is sometimes incorrect. Let* $\tilde{Covid_i} = 1$ *if the test reports positive and* $\tilde{Covid_i} = 0$ *if negative.*

Recall we can express the test results like $\tilde{Covid_i} = Covid_i + v_i$ where v_i is measurement error.

1. What is v_i when $\tilde{Covid_i} = Covid_i$? Correctly measured, so $v_i = 0$!
2. What is v_i for a false positive? $\tilde{Covid_i} = 1$ and $Covid_i = 0$, therefore $v_i = 1$.
3. What is v_i for a false negative? $\tilde{Covid_i} = 0$ and $Covid_i = 1$, therefore $v_i = -1$
4. Is measurement error in $Covid_i$ classical? Definitely not! The error is negatively correlated with infection status.

Causal Inference

- Causal inference is the process of establishing causal relationships between different variables.
- A (causal) effect reflects the difference between what happened and what could have happened.
- The fundamental issue of causal inference is missing data: we don't observe what would have happened in a counterfactual world.
- Examples:
  - How much R&D would firms do if we doubled the R&D tax credit?
  - How much property damage would Philadelphia have suffered had the Eagles won the Super Bowl?
- Challenges:
  - Validity of applied research rests on untestable assumptions.
  - No fun at parties.

Potential Outcomes and Counterfactuals

- Potential outcomes provide framework for thinking about causality using counterfactuals.
- Suppose we are interested in the effect of receiving the initial dose of the Pfizer vaccine on the probability of Covid-19 infection within some specified period of time.
  - $Y_{1i} = Covid_{1i}$: infection status if vaccinated.
  - $Y_{0i} = Covid_{0i}$: infection status if unvaccinated.
  - Treatment effect for individual i is $\tau_i = Covid_{1i} - Covid_{0i}$.
- Given that we observe vaccination status Vax_i and outcome $Covid_i$, can we identify τ_i?
  - No — we cannot identify individual treatment effects because we never observe an individual's outcome in both the treated an untreated state.
  - $Covid_i = Covid_{0i} + (Covid_{1i} - Covid_{0i})Vax_i$

Potential Outcomes and Counterfactuals

We can decompose the difference in means into the average treatment effect on the treated (ATT) and selection bias.

$$
\begin{aligned}
E[Covid_i|Vax=1]-E[Covid_i|Vax_i=0] = & \underbrace{E[Covid_{1i}-Covid_{0i}|Vax_i=1]}_{\text{ATT}} \\
& + \underbrace{E[Covid_{0i}|Vax_i=1]-E[Covid_{0i}|Vax_i=0]}_{\text{Selection Bias}}.
\end{aligned}
$$

For example, if the ATT is -0.01 but we estimate a difference in means of -0.02, then the vaccinated would have been 1 percentage points less likely to contract Covid-19 even if they had not been vaccinated.

Potential Outcomes and Counterfactuals

Suppose we have obtained a huge dataset including the vaccination and infection status of all individuals in the United States. We want to use this dataset to estimate average effects of vaccination on infection.

We estimate the bivariate regression: $Covid_i = \beta_0 + \beta_1 Vax_i + u_i$ by OLS. Is $\hat{\beta}_1$ an unbiased estimate of ATT?

Almost certainly not.

Potential Outcomes and Counterfactuals

$$Covid_i = \beta_0 + \beta_1 Vax_i + u_i$$

- β_1 captures the difference in infection rate between the vaccinated and unvaccinated.
- People who are vaccinated tend to be more concerned about Covid-19 and thus may limit their exposure to the virus more than the unvaccinated, reducing the vaccinated infection rate. This would tend to make $\hat{\beta}_1$ more negative.
- Some of the unvaccinated cannot be vaccinated (e.g. due to allergies) and thus may limit their exposure to the virus more than the vaccinated, reducing the unvaccinated infection rate. This would tend to make $\hat{\beta}_1$ more positive.

Randomized Control Trials

- Pros:
  - If properly administered, RCTs capture the average causal effect of a treatment.
- Cons:
  - Expen$ive.
  - Ethical dilemmas.
- How does random assignment guarantee causality?
  - It severs the connection between treatment and other determinants of the outcome, ridding us of selection bias.
  - For example, by randomly assigning vaccinations in a clinical trial, we make it so that vaccination is uncorrelated with unobserved factors like willingness to social distance.
  - This helps achieve ***internal validity*** so that estimates reflect a true causal relationship.

Review Exercise: Current Population Survey

For the remainder of the session, we will analyze regressions using data derived from the January 2023 Current Population Survey, a large survey of Americans.

We will focus on the following variables:

- `served`: 1 if the respondent ever served in the US Armed Forces, 0 otherwise
- `earnings`: weekly earnings in dollars, 0 if not working
- `foreign_born`: 1 if the respondent was born abroad, 0 if born in US
- `female`: 1 if the respondent is female, 0 otherwise,
- `age`: respondent's current age **minus 18**

Review Exercise: US Military Service

	A	B	C	D
Dependent Var.:	served	served	served	served
Constant	0.0747*** (0.0011)	0.1283*** (0.0019)	0.1377*** (0.0021)	0.0908*** (0.0022)
foreign_born	-0.0583*** (0.0016)	-0.0587*** (0.0017)	-0.1090*** (0.0031)	-0.1050*** (0.0030)
female		-0.1047*** (0.0019)	-0.1230*** (0.0022)	-0.1272*** (0.0022)
foreign_born × female			0.0989*** (0.0033)	0.0974*** (0.0033)
age				-0.0005** (0.0002)
age square				5.05e-5*** (3.05e-6)
Observations	78,301	78,301	78,301	78,301
R2	0.00858	0.05436	0.06054	0.10647
Adj. R2	0.00857	0.05434	0.06051	0.10641

Review Exercise: US Military Service

1. According to Regression B, what is the probability a foreign-born female has ever served in the US military?
2. According to Regression C, what is the probability a foreign-born female has ever served in the US military?
3. Interpret the coefficient on age in Regression D.

Review Exercise: US Military Service

1. According to Regression B, what is the probability a foreign-born female has ever served in the US military?
   - $0.1283 - 0.0587 - 0.1047 = -0.0351$
2. According to Regression C, what is the probability a foreign-born female has ever served in the US military?
   - $0.1377 - 0.1090 - 0.1230 + 0.0989 = 0.0046.$
3. Interpret the coefficient on age in Regression D.
   - The association of service with a one unit change in age is $\widehat{\beta}_1 + 2\widehat{\beta}_2\texttt{age}$.
   - age is coded as the respondent's actual age minus 18, so in this case $\texttt{age} = 0$.
   - For an 18-year old American, an additional year of age is associated with a 0.05 percent decrease in the probability of military service, controlling for gender, whether they were foreign-born, and their interaction.

Review Exercise: US Military Service

4. Now suppose you want to estimate the predicted difference in service rates for foreign- vs. US-born Americans controlling for gender and a quadratic in age. However, you suspect that the association of age with service varies differently by gender. Write a PRF you could estimate that allows for differential quadratic age trends by gender.

$$\begin{aligned}\texttt{service}_i = & \beta_0 + \beta_1 \texttt{foreign\_born}_i + \beta_2 \texttt{female}_i + \beta_3 \texttt{age}_i + \beta_4 \texttt{age}_i^2 \\ & + \beta_5 \texttt{female}_i \times \texttt{age}_i + \beta_6 \texttt{female}_i \times \texttt{age}_i^2\end{aligned}$$

Review Exercise: Earnings

	A	B	C
Dependent Var.:	earnings	earnings	log(earnings)
Constant	187.5***	1,317.8***	6.955***
	(3.523)	(13.57)	(0.0139)
foreign_born	11.92	-28.43	0.0075
	(8.839)	(32.41)	(0.0279)
female	-50.03***	-263.0***	-0.2410***
	(4.447)	(18.34)	(0.0193)
foreign_born × female	-26.79*	-6.784	-0.0711
	(10.85)	(45.05)	(0.0468)
Observations	78,301	10,377	10,377
R2	0.00319	0.02973	0.02570
Adj. R2	0.00315	0.02945	0.02542

Review Exercise: Earnings

1. Although Regressions A and B estimate the same coefficients, the estimates differ substantially between the two because Regression B omits individuals who aren't currently working. Why would this lead to a much smaller intercept in Regression A?
2. Why might the coefficient on `foreign_born:female` be large and significant in Regression A but not in Regression B?
3. Interpret the coefficient on `female` in Regression C.

Review Exercise: Earnings

1. Although Regressions A and B estimate the same coefficients, the estimates differ substantially between the two because Regression B omits individuals who aren't currently working. Why would this lead to a much smaller intercept in Regression A?

   - Individuals that aren't working have earnings of $0. Including them in the sample decreases the predicted earnings for all groups including US-born males whose average earnings are given by the intercept.

2. Why might the coefficient on `foreign_born:female` be large and significant in Regression A but not in Regression B?

   - If the difference between foreign-born females and males in the propensity to work is larger in magnitude than the difference between US-born females and males, this would tend to also be reflected by a greater difference in earnings. Regression B uses a sample of only workers, so this wouldn't have an impact on predicted earnings in that regression.

3. Interpret the coefficient on `female` in Regression C.

   - US-born females earn 24 percent less per week than US-born males.