Lecture 18: Measurement Error and Potential Outcomes Framework

Pierre Biscaye

Fall 2022

Agenda

1. Measurement error in Y
2. Measurement error in X
3. Potential outcomes framework

Measurement error in the dependent variable

$$y^* = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \tag{1}$$

- y^* is the true measure of the outcome.
- But we observe $y = y^* + e_0$.
- What happens if we regress

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + v_i \tag{2}$$

Measurement error in the dependent variable

$$y_i = y_i^* + e_{0i} = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + u_i + e_{0i} \tag{3}$$

- So, if $E[u_i + e_{0i} | x_{1i}, ..., x_{ki}] = 0$, then $E[\hat{\beta}_j] = \beta_j$ for all j.
- In other words, if MLR 4 *would have held* and if the measurement error in y is *uncorrelated* with our x variables, then we still have unbiased estimators.
- But if the measurement error e_0 is correlated with some x_j, that will generate bias.
- We call measurement error that is random with respect to the true variable (y^*) *classical* measurement error; non-random measurement error is non-classical.
    - If e_0 is random with respect to y^*, it must also be uncorrelated with the x variables because otherwise y^* and e_0 would be correlated via the mutual correlation with x.

Example: reported Income

- Suppose that many people in our data earn income in cash (for example, people who receive tips, entrepreneurs, etc.).
- Then they may not remember the *exact* amount of earned income.
- We might want to estimate $income = \beta_0 + \beta_1 age + \beta_2 Exp + \beta_3 Ed + u$
- But have to settle for $rep.inc. = \beta_0 + \beta_1 age + \beta_2 Exp + \beta_3 Ed + u + e_0$
- As long as the mistakes people make in reporting are unrelated to our y^* variable (classical measurement error) and by extension to the x variables, the $\hat{\beta}$ estimates will be unbiased.
- Of course, if (say) education was related to mistakes in reporting...

Variance with random mistakes

- The good news: measurement error in the dependent variable isn't *too* much of a problem...
  - ...as long as it consists of random mistakes (classical measurement error) rather than systematic misreporting.
- Variance of our $\hat{\beta}$ estimates will still increase.
- If e_0 is uncorrelated with u, then $var(u + e_0) = \sigma_u^2 + \sigma_{e_0}^2 > \sigma_u^2$.
  - Recall this is the numerator for $var(\hat{\beta})$.
- $\Rightarrow$ Collecting better data is still valuable!

Variance with random mistakes

- The good news: measurement error in the dependent variable isn't *too* much of a problem...
  - ...as long as it consists of random mistakes (classical measurement error) rather than systematic misreporting.
- Variance of our $\hat{\beta}$ estimates will still increase.
- If e_0 is uncorrelated with u, then $var(u + e_0) = \sigma_u^2 + \sigma_{e_0}^2 > \sigma_u^2$.
  - Recall this is the numerator for $var(\hat{\beta})$.
- $\Rightarrow$ Collecting better data is still valuable!
- The bad news: non-classical measurement error can throw off your estimates in unpredictable ways.
- To Jupyter!

Measurement error in an independent variable

- Now, suppose $y = \beta_0 + \beta_1 x_1^* + u$
- Suppose SLR1-SLR4 all hold
- But, we don't observe x_1^*. Instead we observe $x_1 = x_1^* + e_1$.
- We then have $y = \beta_0 + \beta_1 x_1 + v = \beta_0 + \beta_1 x_1^* + u - \beta_1 e_1$.
- New SLR4: $E[u - \beta_1 e_1 | x_1] = 0$

Measurement error in an independent variable

- New SLR4: $E[u - \beta_1 e_1 | x_1] = E[u|x_1] - \beta_1 E[e_1|x_1] = 0$
- $E[u|x_1] = 0$ is fairly innocuous.
  - We know $E[u|x_1^*] = 0$.
  - $E[u|x_1] = E[u|x_1^*] + E[u|e]$.
  - So as long as the measurement error e is not correlated with the structural error u we are ok - no reason to believe it should be.
- $\beta_1 E[e_1|x_1] = 0$ is a harder assumption to get a handle on.
- In other words, is our measurement error uncorrelated with our measurement?
- We'll consider two special cases for $cov(e_1, x_1)$.

Case 1: $cov(e_1, x_1) = 0$

- The measurement error is uncorrelated with the measured x.
- In this case, true variable x_1^* is the sum of measurement x_1 and a random variable e_1.
- Then $E[e_1|x_1] = 0$ so $E[u - \beta_1 e_1|x_1] = 0$ and SLR4 holds
- $\Rightarrow E[\hat{\beta_1}] = \beta_1$ (we already assumed SLR1-SLR3)
- This is the case with proxy x variables.
  - For example $Ability_i = \delta_0 + \delta_1 IQ_i + v_i$; by definition v is the part of $Ability$ not correlated with IQ.
  - In this case $Ability = x^*$, $IQ = x$.
- As with proxy variables, variance (and therefore SEs) increases:
- $var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2 > \sigma_u^2$

Case 2: $cov(e_1, x_1^*) = 0$

- This is *classical* measurement error: the error is unrelated to the true x_1^*.
- By construction, cov$(e_i, x_i) > 0$ (Why?)
- We therefore have an omitted variable in our error term.
- Recall formula for sign of OVB: $E[\hat{\beta}_1] = \beta_1 + \beta_2\delta_1$
  - Here $\beta_2 = -\beta_1$ (relationship between e_1 and y)
  - 0$< \delta_1 = \frac{cov(e_i, x_i)}{var(x_i)} < 1$
- $\implies E[\hat{\beta}_1] = \beta_1(1 - \delta_1)$
- The estimate is biased towards 0.
  - Formal calculation of this bias in a bit.

Interpreting $cov(e_1, x_1)$ and $cov(e_1, x_1^*)$

- $y = \beta_0 + \beta_1 x_1^* + u, \quad x_1 = x_1^* + e_1$
- Suppose $cov(x_1, e_1) = 0$
  - $cov(x_1^*, e_1) = cov(x_1 - e_1, e_1) = -\sigma^2_{\mathbf{e_1}} < 0$
  - So either x_1 or x_1^* *must* be correlated with e_1
- Alternatively, suppose $cov(x_1^*, e_1) = 0$
  - $cov(x_1^*, e_1) = cov(x_1 - e_1, e_1) = cov(x_1, e_1) - \sigma^2_{\mathbf{e_1}} = 0$
  - $\Rightarrow cov(x_1, e_1) = \sigma^2_{\mathbf{e_1}} > 0$
  - Often called "Classic Errors in Variables Problem"

The two cases in words

- Consider reported and actual income.
- If $cov(rep.inc, e_1) = 0$ then actual income is reported income + a random unreported bonus.
    - $cov(income, e_1) \neq 0$: if the unreported bonus is larger (more positive), actual income must also be larger.
- If $cov(income, e_1) = 0$ then reported income is actual income + a random mistake.
    - $cov(rep.inc, e_1) \neq 0$: if the random mistake is larger (more positive), reported income must also be larger.

Classical measurement error in x

$$y = \beta_0 + \beta_1 x_1 + u - \beta_1 e_1 \tag{4}$$

$$x_1 = x_1^* + e_1 \tag{5}$$

$$cov(x_1^*, e_1) = 0 \Rightarrow cov(x_1, e_1) = \sigma_{e_1}^2 \tag{6}$$

- The "Classic Errors in Variables Problem".
- We've skipped asymptotic results in the book.
- Instead, treat this as an omitted variables problem.

Classical measurement error as omitted variable

Represent the relationship between e_1 and x_1 in the usual way for an omitted variable.

$$y = \beta_0 + \beta_1 x_1 + u - \beta_1 e_1 \tag{7}$$

$$e_1 = \delta_0 + \delta_1 x_1 + v \tag{8}$$

$$\delta_1 = \frac{cov(x_1, e_1)}{var(x_1)} = \frac{\sigma_{e_1}^2}{\sigma_{x_1}^2} = \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \tag{9}$$

Then we can calculate the bias in our estimate:

$$E[\hat{\beta_1}] = \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}$$

$$E[\hat{\beta_1}] = \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right)$$

$$E[\hat{\beta_1}] = \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right)$$

Attenuation bias

- Since $0 < \frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} < 1$ we estimate a $\hat{\beta_1}$ which is closer to 0 than β_1.
- This is called *attenuation bias*
- In most cases, this means we are less likely to reject the null hypothesis: this type of measurement error is *conservative*.

To Jupyter!

Other cases

- If neither $\text{cov}(e_i, x_i) = 0$ nor $\text{cov}(e_i, x_i^*) = 0$ (most likely cases) can't say what will happen.
- Same goes with complex correlations between e_i and multiple x_i variables.
- One thing is certain: even when we get unbiased estimates, standard errors increase.
- We'll focus for intuition on the two special cases.

Measurement error summary

- Proxy variables: if proxy explains all variation in true variable that is correlated with other predictors ("good" proxy), get unbiased estimates. Might still reduce bias if explain a lot of the variation.

Measurement error summary

- Proxy variables: if proxy explains all variation in true variable that is correlated with other predictors ("good" proxy), get unbiased estimates. Might still reduce bias if explain a lot of the variation.
- Measurement error in dependent variable: if error is classical (unrelated to y^*), get unbiased estimates. If not, bias could go different ways.

Measurement error summary

- Proxy variables: if proxy explains all variation in true variable that is correlated with other predictors ("good" proxy), get unbiased estimates. Might still reduce bias if explain a lot of the variation.
- Measurement error in dependent variable: if error is classical (unrelated to y^*), get unbiased estimates. If not, bias could go different ways.
- Measurement error in independent variable:
  - If error is classical (unrelated to true x^*), estimates are biased toward 0 (attenuated).
  - If error is unrelated to reported x (other special case), estimates are unbiased. This is the case with "good" proxy variables.
  - If in between, can't say what will happen. Need to think carefully about likely nature of measurement error.

Measurement error summary

- Proxy variables: if proxy explains all variation in true variable that is correlated with other predictors ("good" proxy), get unbiased estimates. Might still reduce bias if explain a lot of the variation.
- Measurement error in dependent variable: if error is classical (unrelated to y^*), get unbiased estimates. If not, bias could go different ways.
- Measurement error in independent variable:
  - If error is classical (unrelated to true x^*), estimates are biased toward 0 (attenuated).
  - If error is unrelated to reported x (other special case), estimates are unbiased. This is the case with "good" proxy variables.
  - If in between, can't say what will happen. Need to think carefully about likely nature of measurement error.
- In all cases, measurement error will increase error variance and make estimates less precise.
- All of these intuitions apply to multiple regression case, too.

Impact evaluation

- We will often want to evaluate the effects of a policy.
    - Does providing childcare subsidies increase female labor force participation?
    - Does increasing (or reducing) police resources reduce crime?
    - Does irrigation boost crop yields?
- Let's suppose we have a policy that we want to evaluate.
- For people who receive the policy (a "*treatment*"), $T_i = 1$; for people who do not $T_i = 0$.
- If we have an outcome y_i, we can regress $y_i = \beta_0 + \beta_1 T_i + u_i$.

Impact evaluation and potential outcomes

- When we do policy analysis, we want to estimate $E[y_i | T_i = 1] - E[y_i | T_i = 0]$.
- What we can estimate is $y_i = \beta_0 + \beta_1 T_i + u_i$.
    - T_i is binary ⇒ if SLR1-4 hold, $\hat{\beta_1}$ estimates $E[y_i | T_i = 1] - E[y_i | T_i = 0]$.
- But there's a big catch: we never observe the same person *i* at the same time both receiving treatment and not.
    - E.g., a person receives either the vaccine or the placebo.
    - E.g., a city either receives an increase in police funding or it does not.
- At any point in time we observe an outcome under treatment for someone who is in the treatment grouop, or an outcome under control for someone who is in the control group, but not both.

Potential outcomes framework

In the potential outcomes framework, we think of four possible conditions we could hope to observe:

$$E[Y_i^T | T_i = 1] \text{ Outcome under treatment for treated} \tag{10}$$

$$E[Y_i^T | T_i = 0] \text{ Outcome under control for treated} \tag{11}$$

$$E[Y_i^C | T_i = 0] \text{ Outcome under control for control} \tag{12}$$

$$E[Y_i^C | T_i = 1] \text{ Outcome under treatment for control} \tag{13}$$

We can observe (1) and (3) but never (2) or (4), the counterfactuals we would have expected in the treatment and control groups if they had not or had received the treatment, respectively.

The *counterfactual*

- The *counterfactual* for $E[Y_i^T | T_i = 1]$ is $E[Y_i^T | T_i = 0]$
- The *counterfactual* is what we should have expected to happen to the treatment group if they *were not* treated.
- We therefore want to observe $E[Y_i^T | T_i = 1] - E[Y_i^T | T_i = 0]$.
  - Could equivalently observe $E[Y_i^C | T_i = 1] - E[Y_i^C | T_i = 0]$..

The *counterfactual* and what we observe

- We want to observe $E[Y_i^T | T_i = 1] - E[Y_i^T | T_i = 0]$.
- We can estimate $y_i = \beta_0 + \beta_1 T_i + u_i$.
- $\hat{\beta_1}$ estimates $E[Y_i^T | T_i = 1] - E[Y_i^C | T_i = 0]$.
- When can we be confident that $E[Y_i^C | T_i = 0]$ is similar to the *counterfactual* $E[Y_i^T | T_i = 0]$?

The *counterfactual* and what we observe

- We want to observe $E[Y_i^T | T_i = 1] - E[Y_i^T | T_i = 0]$.
- We can estimate $y_i = \beta_0 + \beta_1 T_i + u_i$.
- $\hat{\beta_1}$ estimates $E[Y_i^T | T_i = 1] - E[Y_i^C | T_i = 0]$.
- When can we be confident that $E[Y_i^C | T_i = 0]$ is similar to the *counterfactual* $E[Y_i^T | T_i = 0]$?
  - If SLR4 holds: no variables affecting Y are correlated with T_i, meaning the treatment and control groups are equivalent in expectation.
- If SLR4 does not hold, then something in u is correlated with T_i: treatment and control groups have differences that affect Y.
  - E.g., if the treatment is more funding for police, this might be correlated with having a Democrat- or Republican-run city government.
  - E.g., if the treatment is irrigation, this might be correlated with farmer education or proximity to a water source.

Randomization

- If treatment is randomized, and our sample is large enough

$$E[Y_i^T | T_i = 1] = E[Y_i | T_i = 1] = E[Y_i^C | T_i = 1] \quad (14)$$
$$E[Y_i^T | T_i = 0] = E[Y_i | T_i = 0] = E[Y_i^C | T_i = 0] \quad (15)$$

- Effectively, randomization of T ensures $E[u|T] = 0$, so treatment and control groups are equivalent in expectation.
- Thus, what we observe is equal to the counterfactual.
- Under a randomized controlled trial (RCT), we can estimate the unbiased effects of policy by

$$\bar{Y_i}^T - \bar{Y_i}^C \quad or \quad (16)$$
$$y_i = \beta_0 + \beta_1 T_i + u_i \quad (17)$$

Estimating impacts via RCT requires planning

- T_i needs to be randomly assigned.
- This is unlikely in observational data.
- Instead, we need to implement this randomization:
  1. Need to know in advance the population eligible for the program.
  2. Need to work with program implementers to select a random sample of this population, and then randomly assign treatment status to units in the sample.
  3. Need to collect data at least after implementation.

Example: effects of web-based job portals

- Web-based job portals are increasingly popular throughout the world.
  - In many contexts, share information about jobs through sms messages.
- Does enrollment in a web-based portal improve employment outcomes?
- In principle, we could collect data on a sample of job-seekers and compare employment outcomes for those who are on the portal to those who are not.
- Any concerns with this approach?

Job portal RCT in India

1. A project collected a sample frame of recent vocational institute graduates in India.
2. At random, enrolled some of them in a web-based portal that shares job information.
3. At random, made sure that the portal gave some of the enrolled individuals information about *a lot* of jobs.
    - Randomization done by computer program.
4. Collected data three times:
    - Baseline before treatment assignment
    - Midline 6 months after treatment
    - Endline 12-18 months after treatment

We have two treatment groups

1. Group T has “normal” access to the portal.
2. Group TP has “priority” access to the portal.
    - Group TP receives a lot of sms messages about jobs: this is from the second randomization
3. The control group has no access to the portal: $T = 0$ and $TP = 0$.

We have random assignment to T and TP

- So $E[Y_i^T | T_i = 0] = E[Y_i^C | T_i = 0] = E[Y_i | T_i = 0]$
- So we can examine $E[Y_i^T | T_i = 1] - E[Y_i^C | T_i = 0]$ to estimate the effects of the portal.
    - And similarly for the effects of the priority portal access.
- We focus on employment as an outcome.

2 primary specifications

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 TP_i + u_i \tag{18}$$

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 TP_i + \gamma_1 x_{1i} + ... + \gamma_k x_{ki} + u_i \tag{19}$$

- Why add controls?

To Jupyter!

2 primary specifications

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 TP_i + u_i \tag{18}$$

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 TP_i + \gamma_1 x_{1i} + ... + \gamma_k x_{ki} + u_i \tag{19}$$

- Why add controls?
- Controls in randomization: rarely concerned about omitted variables.
  - After all, treatment assignment is randomized.
- But, controls can still reduce σ_u^2: increase estimate precision.
  - Recall: $var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST(1-R_j^2)}$

To Jupyter!

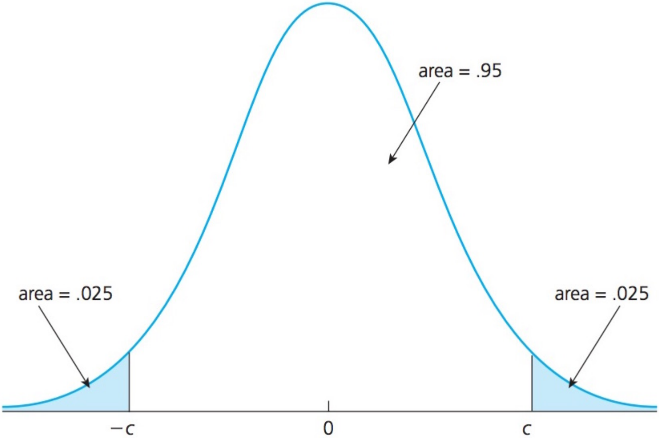
What could influence our estimates?

- We find that job portal access significantly *decreases* the probability of being employed at endline by 5 percentage points, ceteris paribus.
- Priority job portal access has no significant effect.
- Could there be omitted variables (violating MLR4)?

Randomization guarantees that $E[u|T] = 0$

- Even more than that, for any variable x, $E[x|T] = E[x]$.
- But in any *individual* sample, *some* variables will be correlated with T.
    - Randomization only guarantees independence in expectation.
- This is because our standard of proof of a statistical relationship is that it is less than α% likely that we would see a relationship due to chance.

Some improbable values occur naturally



Frequentist significance

- We can flip the α% around: it is exactly α% likely that *any* x variable will be correlated with T in our sample.
- If we test lots of variables, we will certainly find some:
  - Suppose we test 10 variables. The probability that none are correlated with T at the 5% level is $0.95^{10} = 0.6$.
  - So the probability that at least one is is 1-0.6= 0.4.
  - If we test 20 variables, the probability that at least one is correlated with T is $1 - 0.95^{20} = 0.65$
  - There are *lots* of variables that might be in u. *Some* be correlated with T.
- We can use baseline data (from before treatment) to detect which x variables are associated with T just due to chance.

Balance at baseline

Table 17: Balance

	(1) Control	(2) Treatment	(3) Priority Treatment	(4) (1) vs. (2), p-value	(5) (1) vs. (3), p-value	(6) (2) vs. (3), p-value	(7) Joint test
=1 if male	1.26	0.02	0.01	0.19	0.52	0.53	0.42
Age	26.31	-0.20	0.03	0.36	0.89	0.29	0.51
Married Y/N	1.46	-0.01	0.01	0.66	0.60	0.31	0.60
Religion=Hindu	0.10	0.02	0.03	0.07	0.05	0.82	0.10
Religion=Muslim	0.90	-0.02	-0.03	0.07	0.06	0.85	0.11
=1 if ST/SC caste	0.81	-0.02	-0.02	0.23	0.26	1.00	0.42
=1 if OBC caste	0.01	0.05	0.06	0.02	0.01	0.58	0.01
=1 if general caste	0.19	-0.03	-0.04	0.17	0.07	0.56	0.17
Father's education>0	0.04	0.03	0.02	0.23	0.47	0.66	0.48
Mother's education>0	1.05	0.03	-0.02	0.34	0.39	0.06	0.18
=1 if live in village	-0.17	-0.01	0.01	0.61	0.74	0.38	0.67
Received formal skills training	1.94	0.03	0.02	0.17	0.40	0.64	0.39
=1 if currently employed	0.87	0.03	0.02	0.16	0.35	0.69	0.37
=1 if looking for job	0.33	0.02	0.01	0.40	0.80	0.57	0.69
Access to Internet Y/N (clean)	0.67	0.04	0.02	0.03	0.28	0.33	0.10
Reservation wage (winsorized)	19376.31	-398.80	753.96	0.21	0.03	0.00	0.00

Standard errors are clustered at the respondent level. Asterisks indicate statistical significance at the 1% ***, 5% **, and 10% * levels.

What does this mean for MLR4?

- Randomization guaranteed that on average, there would be no x variables in u correlated with T.
  - On average, $E[u|T] = 0$.
- In practice, we know that in any one sample, *some* variables will be correlated
  - What if these correlations influence $\widehat{\beta_{OLS}}$?
- One option: control for x variables that you know about that are correlated with T.
- $y_i = \beta_0 + \beta_1 T_i + \beta_2 TP_i + \gamma_1 x_{1i} + ... + \gamma_k x_{ki} + u_i$
  - In this case we would want to add religion dummies, access to internet, and reservation wage to the model along our existing controls.