Research Design

Wei Yu, PhD

January 30, 2019

(Slides were prepared by Dr. Christine Pal Chee)

Health Services Research

- Many questions in health services research aim to establish causality
  - Does the adoption of electronic medical records reduce health care costs or improve quality of care?
  - Did the transition to Patient Aligned Care Teams (PACT) improve quality of care and health outcomes?
  - What effect will the Affordable Care Act (ACA) have on the demand for VHA services?
- Ideally studied through randomized controlled trials
- When can regression analysis of observational data answer these questions?

Poll: Familiarity with Regressions

- How would you describe your familiarity with regression analysis?
  - Regression is my middle name.
  - I’ve run a few regressions and get the gist of how they work.
  - I took a statistics class many years ago.
  - What is a regression?

Objectives

- Provide a conceptual framework for research design
- Review the linear regression model
- Define exogeneity and endogeneity
- Discuss three forms of endogeneity
  - Omitted variable bias
  - Sample selection
  - Simultaneous causality

Research Question

- Start with a research question:
  - What is the effect of X on Y?
- For example:
  - What effect does exercise have on health?

Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

- Y: outcome variable of interest
- X_1: explanatory variable of interest
- X_2: control variable
- e: error term
  - e is the difference between the observed and predicted values of Y
  - e contains all other factors besides X_1 and X_2 that determine the value of Y
- β_1: the change in Y associated with a unit change in X_1, holding constant X_2
  - $\hat{\beta}_1$ is our estimate of β_1

- Model specifies all meaningful determinants of Y

Linear Regression Model (2)

- In our example:

$$health_i = \beta_0 + \beta_1 exercise_i + e_i$$

  - $health$: dependent variable
  - $exercise$: independent variable
  - e: error term
    - e contains all other factors besides exercise that determine health
  - β_1: the change in health associated with an increase in exercise

- When does $\hat{\beta}_1$ estimate the *causal* effect of exercise on health?

Exogeneity

- Assumption: $E(e_i|X_i) = 0$
  - Conditional mean of e_i given X_i is zero
    - Conditional mean independence
  - X is "exogenous"

- Knowing X_i does not help us predict e_i
  - e_i is the difference between the observed and predicted values of Y_i
  - e_i contains other factors besides X_i that determine the value of Y_i
  - Information other than X_i does not tell us anything more about Y_i

- Implies that X_i and e_i **cannot** be correlated

Exogeneity (2)

- In the context of a randomized controlled trial:

$$outcome_i = \beta_0 + \beta_1 treatment_i + e_i$$

  - e_i can include things like age, gender, pre-existing conditions, income, education, etc.

- Because treatment is randomly assigned, $treatment$ and e are independent
  - This implies $treatment$ is exogenous

- In observational studies, $treatment$ is not randomly assigned
  - The best we can do is to control for variables using statistics.

Exogeneity (3)

- In our example:

$$health_i = \beta_0 + \beta_1 exercise_i + e_i$$

- In order for $\hat{\beta}_1$ to estimate the causal effect of exercise on health, $exercise$ must be exogenous
  - All factors other than exercise do not tell us anything more about health
- In the context of a randomized controlled trial, $exercise$ is exogenous
  - Is the same true in the context of observational studies?

Endogeneity

- Violation of the exogeneity assumption
  - X is endogenous
  - Always true when X_i is correlated with e_i
- $\hat{\beta}_1$ is biased
  - $\hat{\beta}_1$ is unbiased if the expected value of $\hat{\beta}_1$ is equal to the true value of β_1
- $\hat{\beta}_1$ will not estimate a causal effect of X on Y
  - $\hat{\beta}_1$ is a measure of the correlation between X and Y
  - Correlation does not imply causation

Poll: Familiarity with endogeneity

Which of the following can cause endogeneity?

A. Omitted variable bias

B. Sample selection

C. Simultaneous causality

D. All of the above

E. A and C only

Forms of Endogeneity

- Omitted variable bias
- Sample selection
- Simultaneous causality

Omitted Variable Bias

- Arises when:
  - A variable omitted from the regression model is a determinant of the dependent variable, *Y*
  - The omitted variable is correlated with the regressor, *X*
- Leads $\hat{\beta}_1$ to be biased
  - $\hat{\beta}_1$ also captures the correlation between the omitted variable and the dependent variable

Omitted Variable Bias (2)

- Regression model: $Y_i = \beta_0 + \beta_1 X_i + e_i$
- Say another factor, W_i, determines Y_i
  - W_i is included in the error term, e_i
- If X_i and W_i are correlated
  - X_i and e_i are correlated
- X_i is endogenous
  - $\hat{\beta}_1$ is biased
    - $\hat{\beta}_1$ also captures the correlation between W_i and Y_i

Omitted Variable Bias: Example

- In our example:

$$health_i = \beta_0 + \beta_1 exercise_i + e_i$$

- Two questions:
  - Besides exercise, do any other factors determine health?
  - Are those factors correlated with exercise?

Omitted Variable Bias: Example (2)

- Consider: diet
  - Does diet affect health?
    - Eating well likely improves health
  - Is diet correlated with exercise?
    - Individuals who eat well are probably more likely to exercise

Omitted Variable Bias: Example (3)

- Diet affects health and is correlated with exercise
  - Diet is an omitted variable
  - $\hat{\beta}_1$ will be biased
    - $\hat{\beta}_1$ also captures the relationship between diet and health

Omitted Variable Bias: Solutions

- Multiple linear regression
  - Include all relevant factors in the regression model so that we have conditional mean independence
  - Often not possible to include all omitted variables in the regression
- Randomized controlled trial
- Natural experiment
  - More on this in the Natural Experiments and Difference-in-Differences lecture on February 13

Omitted Variable Bias: Solutions (2)

- Utilize panel data (same observational unit observed at different points in time)
  - Fixed effects regression: control for unobserved omitted variables that do not change over time
  - More on this in the Fixed Effects and Random Effects lecture on March 6
- Instrumental variables regression
  - Utilize an instrumental variable that is correlated with the independent variable of interest but is uncorrelated with the omitted variables
  - More on this in the Instrumental Variables Regression lecture on February 27

Sample Selection

- Arises when:
  - A selection process influences the availability of data
  - The selection process is related to the dependent variable, Y, beyond depending on X
- Leads $\hat{\beta}_1$ to be biased

Sample Selection (2)

- Form of omitted variable bias
  - The selection process is captured by the error term
  - Induces correlation between the regressor, X, and the error term, e

Sample Selection: Examples

- Want to evaluate the effect of a new tobacco cessation program (offered to all patients) on quitting
  - $quit_i = \beta_0 + \beta_1 treatment_i + e_i$
  - Problem: Individuals who participate in the program may be more likely to quit to begin with
- Want to evaluate the effect of a new primary care model (rolled out for some patients at a facility) on patient satisfaction
  - $satisfaction_i = \beta_0 + \beta_1 model_i + e_i$
  - Problem: Patients who don't like the new program stop coming to the facility and receive their care elsewhere

Sample Selection: Solutions

- Randomized controlled trial
- Natural experiment
  - More on this in the Natural Experiments and Difference-in-Differences lecture on February 13
- Sample selection and treatment effect models
  - For more information:
    - Greene, 2000 Chapter 20
    - Wooldridge, 2010, Chapter 17
- Instrumental variables regression
  - More on this in the Instrumental Variables Regression lecture on February 27

Simultaneous Causality

- Arises when:
  - There is a causal link from X to Y
  - There is also a causal link from Y to X
- Also called simultaneous equations bias
- Leads $\hat{\beta}_1$ to be biased
  - Reverse causality leads $\hat{\beta}_1$ to pick up both effects

Simultaneous Causality: Example

- We want to estimate the effect of primary care visits on glucose levels

$$glucose_i = \beta_0 + \beta_1 pcvisits_i + e_i$$

- If there is a policy in place that increases primary care visits when someone has high glucose levels:

$$pcvisits_i = \gamma_0 + \gamma_1 glucose_i + \varepsilon_i$$

- Both equations are necessary to understand the relationship between primary care visits and glucose levels

Simultaneous Causality: Example (2)

- We now have two simultaneous equations:

$$glucose_i = \beta_0 + \beta_1 pcvisits_i + e_i \text{ (1)}$$
$$pcvisits_i = \gamma_0 + \gamma_1 glucose_i + \varepsilon_i \text{ (2)}$$

- Suppose a positive error e_i leads to a higher value of $glucose_i$

$$\overset{\uparrow}{glucose_i} = \beta_0 + \beta_1 pcvisits_i + \overset{\uparrow}{e_i} \text{ (1)}$$

- If $\gamma_1 > 0$, then a higher value of $glucose_i$ leads to a higher value of $pcvisits_i$

$$\overset{\uparrow}{pcvisits_i} = \gamma_0 + \overset{>0}{\gamma_1} \overset{\uparrow}{glucose_i} + \varepsilon_i \text{ (2)}$$

- Therefore, a positive error e_i leads to a higher value of $pcvisit_i$
  - $e_i \uparrow \rightarrow pcvisits_i \uparrow$
  - $pcvisits_i$ and e_i are correlated
  - $\hat{\beta}_1$ is biased

Simultaneous Causality: Solutions

- Randomized controlled trial where the reverse causality channel is eliminated
- Natural experiment
  - More on this in the Natural Experiments and Difference-in-Differences lecture on March 6
- Instrumental variables regression
  - Utilize an instrumental variable that is correlated with X but is uncorrelated with the error term (does not otherwise determine Y)
  - More on this in the Instrumental Variables Regression lecture on February 27

Summary

- Good research design requires an understanding of how the dependent variable is determined
- Need to ask: is the explanatory variable of interest exogenous?
  - Are there omitted variables?
  - Is there sample selection?
  - Is there simultaneous causality?
- Exogeneity is necessary for the estimation of a causal treatment effect
- Understanding sources of endogeneity can:
  - Help us understand what our regression estimates actually estimate and the limitations of our analyses
  - Can point us to appropriate methods to use to answer our research question

Resources

- Stock and Watson, Introduction to Econometrics, 3rd edition (2011)
- Green, Econometric Analysis, 8th edition (2018)
- Wooldridge, Econometric Analysis of Cross Section and Panel Data, 2nd edition (2010)