

Research Design

Christine Pal Chee

October 9, 2013

Health Services Research

- Many questions in health services research aim to establish causality
 - Does the adoption of electronic medical records reduce health care costs?
 - 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?
 - A. Regression is my middle name.
 - B. I've run a few regressions and get the gist of how they work.
 - C. I took a statistics class many years ago.
 - D. 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 receiving antiretroviral therapy (ARVs) have on substance use?

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 : additional control variable
 - e : error term
 - e contains all other factors besides X_1 and X_2 that determine the value of Y
 - e is the difference between the observed and predicted values 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:

$$substance_i = \beta_0 + \beta_1 ARV_i + e_i$$

- *substance*: dependent variable
- *ARV*: independent variable
- *e*: error term
 - *e* contains all other factors besides receiving ARVs that determine substance use
- β_1 : the change in the likelihood of substance use associated with receiving ARVs
- When does β_1 estimate the *causal* effect of receiving ARVs on substance use?

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 : the difference between the observed and predicted values of Y_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 hope for is that *treatment* is **as if** randomly assigned

Exogeneity (3)

- In our example:

$$substance_i = \beta_0 + \beta_1 ARV_i + e_i$$

- In order for $\hat{\beta}_1$ to estimate the causal effect of ARVs on substance use, ARV must be exogenous
 - All factors other than receiving ARVs do not tell us anything more about substance use
- In the context of a randomized controlled trial, ARV 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

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
 - $\hat{\beta}_1$ captures the correlation between W_i and Y_i

Omitted Variable Bias: Example

- In our example:

$$substance_i = \beta_0 + \beta_1 ARV_i + e_i$$

- Two questions:

- Besides receiving ARVs, do any other factors determine substance use?
- Are those factors correlated with receiving ARVs?

- Consider two factors: education and health

Omitted Variable Bias: Example (2)

■ Education

- Individuals who are more highly educated are less likely to engage in substance use
- Individuals who are more highly educated are more likely to receive ARVs

■ Health

- Individuals who are sick are less likely to engage in substance use
- Individuals who are sick are more likely to receive ARVs

Omitted Variable Bias: Example (3)

VARIABLES	(1) substance	(2) substance	(3) substance
ARV	-0.0254	-0.0324	-0.0131
Controls			
Education		X	
Health			X

- A change in $\hat{\beta}_1$ suggests omitted variable bias

Omitted Variable Bias: Solutions

- Randomized controlled trial
- Multiple linear regression
 - Include all relevant factors in the regression model so that we have conditional mean independence
 - What if it is not possible to include omitted variables in the regression?

Omitted Variable Bias: Solutions (2)

- Utilize panel data (same observational unit observed at different points in time)
 - Fixed effects regression: Possible to control for unobserved omitted variables as long as those omitted variables do not change over time
 - For more information: Stock and Watson, Chapter 10
- Instrumental variables regression
 - Utilize an instrument 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 Oct 30

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: Example

- Classic examples from economics:
 - The effect of union membership on wages
 - $wages_i = \beta_0 + \beta_1 union_i + e_i$
 - Problem: a worker's decision to join a union is endogenous; it may depend on many factors, including the difference between his union vs. non-union wages
 - The effect of college education on wages
 - $wages_i = \beta_0 + \beta_1 college_i + e_i$
 - Problem: a person's decision to go to college is endogenous; it may depend on many factors, including her anticipated return to college education

Sample Selection: Example (2)

- In the VA:
 - Want to evaluate the effect of a new tobacco dependence treatment program on quitting
 - $quit_i = \beta_0 + \beta_1 treatment_i + e_i$
 - Individuals who participate in the program may be more likely to quit to begin with
 - Want to evaluate the effect of home-based primary care on nursing home use
 - $nursing_i = \beta_0 + \beta_1 HBPC_i + e_i$
 - Individuals who use home-based primary care may be more plugged in and likely to know about other available programs
 - Facilities that adopt home-based primary care may already have other supportive programs to begin with

Sample Selection: Solutions

- Randomized controlled trial
- 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 Oct 30

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 receiving ARVs on substance use

- Regression model:

$$substance_i = \beta_0 + \beta_1 ARV_i + e_i$$

- If substance use also affects the likelihood of receiving ARVs:

$$ARV_i = \gamma_0 + \gamma_1 substance_i + \varepsilon_i$$

- Both equations are necessary to understand the relationship between ARVs and substance use

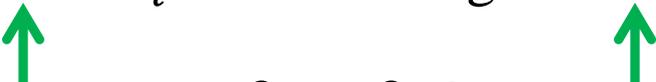
Simultaneous Causality: Example (2)

- We now have two simultaneous equations:

$$substance_i = \beta_0 + \beta_1 ARV_i + e_i \quad (1)$$

$$ARV_i = \gamma_0 + \gamma_1 substance_i + \varepsilon_i \quad (2)$$

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


$$substance_i = \beta_0 + \beta_1 ARV_i + e_i \quad (1)$$

- A higher value of $substance_i$ leads to a higher value of ARV_i


$$ARV_i = \gamma_0 + \gamma_1 substance_i + \varepsilon_i \quad (2)$$

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

Simultaneous Causality: Solutions

- Randomized controlled trial where the reverse causality channel is eliminated
 - Instrumental variables regression
 - Utilize an instrumental variable that is correlated with X (determines ARV) but is uncorrelated with the error term (does not otherwise determine Y , substance use)
 - More on this in the Instrumental Variables Regression lecture on Oct 30
-

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, 7th edition (2012)
- Wooldridge, Econometric Analysis of Cross Section and Panel Data, 2nd edition (2010)