Precept 8: Some review, heteroskedasticity, and causal inference

Soc 500: Applied Social Statistics

Alex Kindel

Princeton University

November 15, 2018

Learning Objectives

1. Review
    1. Calculating error variance
    2. Interaction terms (common support, main effects)
    3. Model interpretation ("increase", "intuitively")
    4. Heteroskedasticity
2. Causal inference with potential outcomes

[0] Thanks to Ian Lundberg and Xinyi Duan for material and ideas.

Calculating error variance

- We have some data: Y, X, Z.
- We think the correct model is $Y = X + Z + u$.
- We estimate this conditional expectation using OLS:
  $Y = \beta_0 + \beta_1 X + \beta_2 Z$
- We want to know the standard error of β_1.

Standard error of β_1

$SE(\hat{\beta}_j) = \sqrt{\frac{1}{1-R_j^2} \frac{\hat{\sigma_u^2}}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$, where R_j^2 is the R^2 of a regression of variable j on all others.

Question: What is $\hat{\sigma_u^2}$?

Calculating error variance

$$\hat{\sigma_u^2} = \frac{\sum_i \hat{u_i}^2}{DF_{resid}}$$

- You can adjust this in finite samples by $\hat{\bar{u}}$ (why?)

Interaction terms

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

- Assume $X \sim \mathcal{N}(?, ?)$ and $Z \in \{0, 1\}$

Interaction terms

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

Scenario 1

- When $Z = 0$, $X \sim \mathcal{N}(3, 4)$
- When $Z = 1$, $X \sim \mathcal{N}(-3, 2)$
- Do you think an interaction term is justified here?

Interaction terms

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

Scenario 2

- When $Z = 0$, $X \sim \mathcal{N}(0, 10)$
- Let $D \sim \text{Bern}(0.5)$
- When $Z = 1$ and $D = 0$, $X \sim \mathcal{N}(-3.1, 4)$
- When $Z = 1$ and $D = 1$, $X \sim \mathcal{N}(2.2, 5)$
- Do you think an interaction term is justified here?

Interaction terms

$$Y = \beta_0 + \beta_1 X + \beta_3 XZ$$

- What are we now assuming about the true relationship?

Interaction terms

$$Y = \beta_0 + \beta_1 X + \beta_3 XZ$$

- When $X = 0$, $E[Y|Z = 0] = E[Y|Z = 1]$
- $Cov[u, Z] = 0$ (why?)

Interpreting regression results

Be careful about making within-unit claims based on regression results.

- "When we see a 1 unit increase in X, we observe an increase of 12 percentage points in Y on average."

Interpreting regression results

Be careful about making within-unit claims based on regression results.

- "When we see a 1 unit increase in X, we observe an increase of 12 percentage points in Y on average."
- "For every 1 unit increase in X, we would expect Y to increase by 12 percentage points on average."

Interpreting regression results

Be careful about making within-unit claims based on regression results.

- "When we see a 1 unit increase in X, we observe an increase of 12 percentage points in Y on average."
- "For every 1 unit increase in X, we would expect Y to increase by 12 percentage points on average."
- "On average, we would expect units with 1 percentage point higher X to have 12 percentage points higher Y."

Interpreting regression results

Don't choose models based on substantive intuitions after you know the results.

- *The American Soldier* was one of the earliest pieces of rigorously statistical social science in the United States

Interpreting regression results

Don't choose models based on substantive intuitions after you know the results.

- *The American Soldier* was one of the earliest pieces of rigorously statistical social science in the United States
- Lazarsfeld (1949): Who adjusts better to military life, rural men or urban men?

Interpreting regression results

Don't choose models based on substantive intuitions after you know the results.

- *The American Soldier* was one of the earliest pieces of rigorously statistical social science in the United States
- Lazarsfeld (1949): Who adjusts better to military life, rural men or urban men?
  - Rural men, because they were more accustomed to austere living conditions

Interpreting regression results

Don't choose models based on substantive intuitions after you know the results.

- *The American Soldier* was one of the earliest pieces of rigorously statistical social science in the United States
- Lazarsfeld (1949): Who adjusts better to military life, rural men or urban men?
  - Rural men, because they were more accustomed to austere living conditions ...?

Interpreting regression results

Don't choose models based on substantive intuitions after you know the results.

- *The American Soldier* was one of the earliest pieces of rigorously statistical social science in the United States
- Lazarsfeld (1949): Who adjusts better to military life, rural men or urban men?
  - Rural men, because they were more accustomed to austere living conditions ...?
- A bait-and-switch: the answer is actually urban men! (they were more used to tight living quarters)

Interpreting regression results

Don't choose models based on substantive intuitions after you know the results.

- *The American Soldier* was one of the earliest pieces of rigorously statistical social science in the United States
- Lazarsfeld (1949): Who adjusts better to military life, rural men or urban men?
  - Rural men, because they were more accustomed to austere living conditions ...?
- A bait-and-switch: the answer is actually urban men! (they were more used to tight living quarters)
- **HARK**ing: "**H**ypothesizing **A**fter **R**esults are **K**nown"

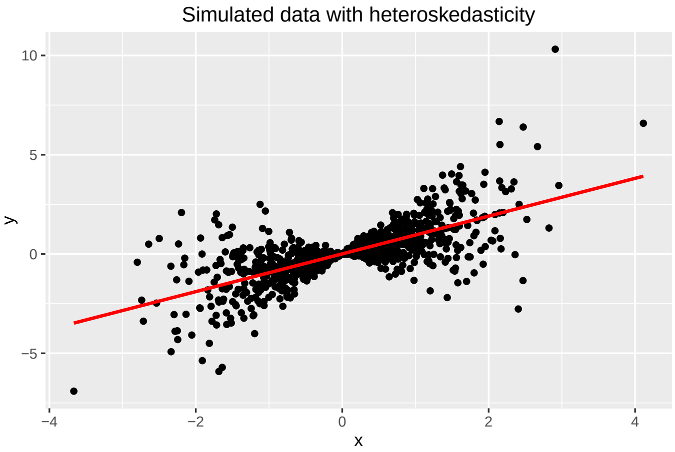
Residual plots for homoskedasticity

- If homoskedasticity is violated, the spread of the residuals *around the conditional mean* will vary with the predicted values.

Violation of homoskedasticity

I generated some data that violates homoskedasticity (the error variance depends on X):

$$Y = X + u, u \sim N(0, |X|)$$

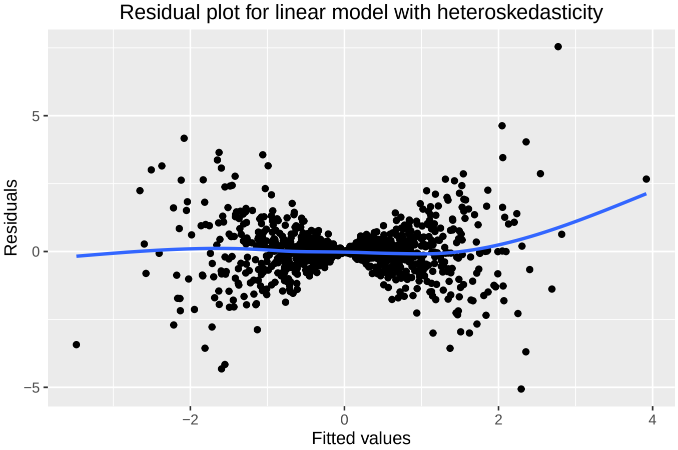


Violation of homoskedasticity

Then I fit a linear model.

$$Y = \beta_0 + \beta_1 X + v$$

The residual plot is below. The error variance is related to the fitted values - homoskedasticity is violated!



Exploring causal inference with a running example

Research question

Does college education cause higher earnings?

Exploring causal inference with a running example

Research question

Does college education cause higher earnings?

Potential (internal) problems:

- Fundamental problem: We cannot observe both states
- Selection bias: Higher ability people may select into college
- SUTVA violation: Your sister's college education might affect your earnings, or "college" might be multiple things
- Heterogeneity: Different effects on different people

Neyman-Rubin Model: Potential outcomes

(from lecture slides)

Two possible conditions:

- Treatment condition $T = 1$
- Control condition $T = 0$

Suppose that we have an individual i.

Key assumption: we can imagine a world where individual i is assigned to treatment and control conditions

Potential Outcomes: responses under each condition, $Y_i(T)$

- Response under treatment $Y_i(1)$
- Response under control $Y_i(0)$

Potential outcomes in our college-earnings example

In our example examining how college affects future earnings, what do the numbers in the table below represent?

	Treatment ($Y_i(1)$)	Control ($Y_i(0)$)
Person 1	45,000	32,000
Person 2	54,000	45,000
Person 3	34,000	34,000

Potential outcomes in our college-earnings example

- $Y_i(1)$ is

Potential outcomes in our college-earnings example

- $Y_i(1)$ is the earnings person i would make if they attended college
- $Y_i(0)$ is

Potential outcomes in our college-earnings example

- $Y_i(1)$ is the earnings person i would make if they attended college
- $Y_i(0)$ is that person's potential earnings if they did not attend college
- The individual causal effect is $\tau_i =$

Potential outcomes in our college-earnings example

- $Y_i(1)$ is the earnings person i would make if they attended college
- $Y_i(0)$ is that person's potential earnings if they did not attend college
- The individual causal effect is $\tau_i = Y_i(1) - Y_i(0)$

Why can we never estimate τ_i?

Potential outcomes in our college-earnings example

- $Y_i(1)$ is the earnings person i would make if they attended college
- $Y_i(0)$ is that person's potential earnings if they did not attend college
- The individual causal effect is $\tau_i = Y_i(1) - Y_i(0)$

Why can we never estimate τ_i? Because we can never observe an individual in both the treatment and control states.
This is the fundamental problem of causal inference.

Fundamental Problem of Causal Inference

Holland 1986: Only one outcome can be observed

	Treatment ($Y_i(1)$)	Control ($Y_i(0)$)
Person 1	45,000	32,000
Person 2	54,000	45,000
Person 3	34,000	34,000

Fundamental Problem of Causal Inference

Holland 1986: Only one outcome can be observed

	Treatment ($Y_i(1)$)	Control ($Y_i(0)$)
Person 1	?	32,000
Person 2	54,000	?
Person 3	34,000	?

Causal inference with a strong assumption

$$T \rightarrow Y$$

- T is the treatment (college education)
- Y is the outcome (earnings)

To make causal inferences, we must assume ignorability:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$$

Causal inference with a strong assumption

$$T \rightarrow Y$$

- T is the treatment (college education)
- Y is the outcome (earnings)

To make causal inferences, we must assume ignorability:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$$

Then the population average treatment effect is

Causal inference with a strong assumption

$$T \rightarrow Y$$

- T is the treatment (college education)
- Y is the outcome (earnings)

To make causal inferences, we must assume ignorability:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$$

Then the population average treatment effect is

$$\begin{aligned}\bar{\tau} = E[Y_i(1) - Y_i(0)] &= E[Y_i(1)] - E[Y_i(0)] \\ &= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]\end{aligned}$$

Meaning,

Causal inference with a strong assumption

$$T \rightarrow Y$$

- T is the treatment (college education)
- Y is the outcome (earnings)

To make causal inferences, we must assume ignorability:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$$

Then the population average treatment effect is

$$\begin{aligned}\bar{\tau} = E[Y_i(1) - Y_i(0)] &= E[Y_i(1)] - E[Y_i(0)] \\ &= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]\end{aligned}$$

Meaning, if college education is ignorable with respect to potential earnings, then the average causal effect of college on earnings is just the average earnings difference between the two groups.

Identification and the role of assumptions

- A causal effect is identified if we could pin it down with an infinite amount of data.
- In any causal study we must state the assumptions under which a causal effect is identified.
- Law of Decreasing Credibility (Manski): The credibility of inference decreases with the strength of the assumptions maintained

Identification and the role of assumptions

- A causal effect is identified if we could pin it down with an infinite amount of data.
- In any causal study we must state the assumptions under which a causal effect is identified.
- Law of Decreasing Credibility (Manski): The credibility of inference decreases with the strength of the assumptions maintained

Assuming ignorability, the causal effect of college on earnings is identified by the mean difference between the two groups.
But this is a heroic assumption, so inference is not very credible!

Stable Unit Treatment Value Assumption (SUTVA)

1. There is only one version of the treatment, not T_1, T_2

Stable Unit Treatment Value Assumption (SUTVA)

1. There is only one version of the treatment, not T_1, T_2
2. Potential outcomes depend only on my treatment status ($Y(1)$, not $Y(1, 0, 0, 1, 0, ..., 0, 1)$ or $Y(\boldsymbol{T})$) (no interference)

Stable Unit Treatment Value Assumption (SUTVA)

1. There is only one version of the treatment, not T_1, T_2
2. Potential outcomes depend only on my treatment status ($Y(1)$, not $Y(1, 0, 0, 1, 0, ..., 0, 1)$ or $Y(\boldsymbol{T})$) (no interference)

In our college-earnings example, what would violate each of these assumptions?

Stable Unit Treatment Value Assumption (SUTVA)

1. There is only one version of the treatment, not T_1, T_2
2. Potential outcomes depend only on my treatment status ($Y(1)$, not $Y(1, 0, 0, 1, 0, ..., 0, 1)$ or $Y(\boldsymbol{T})$) (no interference)

In our college-earnings example, what would violate each of these assumptions?

1. Elite college (Princeton) might be a different treatment from public universities. Only one treatment allowed.

Stable Unit Treatment Value Assumption (SUTVA)

1. There is only one version of the treatment, not T_1, T_2
2. Potential outcomes depend only on my treatment status ($Y(1)$, not $Y(1, 0, 0, 1, 0, ..., 0, 1)$ or $Y(\boldsymbol{T})$) (no interference)

In our college-earnings example, what would violate each of these assumptions?

1. Elite college (Princeton) might be a different treatment from public universities. Only one treatment allowed.
2. My sister's college education might affect my earnings. Interference not allowed.

Definitions of treatment effects

Average treatment effect (ATE): Average over the whole population

$$E[Y_i(1) - Y_i(0)]$$

Average treatment effect on the treated (ATT): Average among those who take the treatment

$$E[Y_i(1) - Y_i(0) \mid D_i = 1]$$

Average treatment effect on the control (ATC): Average among those who do not take the treatment

$$E[Y_i(1) - Y_i(0) \mid D_i = 0]$$

We could define an average treatment effect for any subpopulation of interest.

Definitions of treatment effects

In the case of college education and earnings,

- the ATT is the effect of college among those who attend
- the ATC is the effect among those who do not attend
- the ATE is the average effect over the whole population

In what situations might we care about each one?

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

1. Race cannot be manipulated

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

1. Race cannot be manipulated
   - without the capacity to manipulate the question is arguably ill-posed and the estimand is unidentified

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

1. Race cannot be manipulated
   - without the capacity to manipulate the question is arguably ill-posed and the estimand is unidentified
2. Everything else is post-treatment

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

1. Race cannot be manipulated
   - without the capacity to manipulate the question is arguably ill-posed and the estimand is unidentified
2. Everything else is post-treatment
   - everything else comes after race which is perhaps unsatisfying

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

1. Race cannot be manipulated
   - without the capacity to manipulate the question is arguably ill-posed and the estimand is unidentified
2. Everything else is post-treatment
   - everything else comes after race which is perhaps unsatisfying
   - this also presumes we are only interested in the total effect

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

1. Race cannot be manipulated
   - without the capacity to manipulate the question is arguably ill-posed and the estimand is unidentified
2. Everything else is post-treatment
   - everything else comes after race which is perhaps unsatisfying
   - this also presumes we are only interested in the total effect
3. Race is unstable

Immutable characteristics

from lecture slides

There are three problems with race as a treatment in the causal inference sense

1. Race cannot be manipulated
   - without the capacity to manipulate the question is arguably ill-posed and the estimand is unidentified
2. Everything else is post-treatment
   - everything else comes after race which is perhaps unsatisfying
   - this also presumes we are only interested in the total effect
3. Race is unstable
   - there is substantial variance across treatments which is a SUTVA violation

Objections to potential outcomes

(Morgan and Winship concluding chapter)

- Cannot be used for nonmanipulable causes

Objections to potential outcomes

(Morgan and Winship concluding chapter)

- Cannot be used for nonmanipulable causes
  - but we can get at related things other ways

Objections to potential outcomes

(Morgan and Winship concluding chapter)

- Cannot be used for nonmanipulable causes
  - but we can get at related things other ways
- Appropriate for effects of causes, not causes of effects

Objections to potential outcomes

(Morgan and Winship concluding chapter)

- Cannot be used for nonmanipulable causes
  - but we can get at related things other ways
- Appropriate for effects of causes, not causes of effects
  - but we have to start with small steps before reaching for grand claims

Objections to potential outcomes

(Morgan and Winship concluding chapter)

- Cannot be used for nonmanipulable causes
  - but we can get at related things other ways
- Appropriate for effects of causes, not causes of effects
  - but we have to start with small steps before reaching for grand claims
- Relies on metaphysical quantities that cannot be observed

Objections to potential outcomes

(Morgan and Winship concluding chapter)

- Cannot be used for nonmanipulable causes
  - but we can get at related things other ways
- Appropriate for effects of causes, not causes of effects
  - but we have to start with small steps before reaching for grand claims
- Relies on metaphysical quantities that cannot be observed
  - but we try to get at these with as minimal assumptions as possible
- People rarely have stable characteristics across treatment thresholds
  - e.g. graduate students, vampires; see Paul & Healy (2014), "Transformative Treatments"