PSY 503: Foundations of Psychological Methods

Lecture 18: Regression and Causality in the Absence of Random Assignment

Robin Gomila

Princeton

November 9, 2020

What happens when the treatment was not randomly assigned?

What happens when the treatment was not randomly assigned?

- $\widehat{ATE}$ is biased!

What happens when the treatment was not randomly assigned?

- $\widehat{ATE}$ is biased!
- This week's questions: What happens exactly? What can we do about it?

Meditation, stress, and coffee consumption

- Suppose that the following experimental protocol was deployed on many representative samples of U.S. students to test the effect of **meditation** on stress

Meditation, stress, and coffee consumption

- Suppose that the following experimental protocol was deployed on many representative samples of U.S. students to test the effect of **meditation** on stress
  - Treatment: Meditation (20 min on final exam day) vs. Placebo (e.g., taking a walk for 20 minute on final exam day)
  - DV: Students' blood pressure before a test

Meditation, stress, and coffee consumption

- Suppose that the following experimental protocol was deployed on many representative samples of U.S. students to test the effect of **meditation** on stress
  - Treatment: Meditation (20 min on final exam day) vs. Placebo (e.g., taking a walk for 20 minute on final exam day)
  - DV: Students' blood pressure before a test
- Suppose that taken together these studies suggest that

$$ATE = -5mmHg$$

Our plan to understand bias in non-experimental studies

Our plan to understand bias in non-experimental studies

1. Generate population potential outcomes in R

Our plan to understand bias in non-experimental studies

1. Generate population potential outcomes in R
2. Generate a variable called coffee, which returns 1 for students who consume more than 3 cups of coffee per day, 0 otherwise
   - This variable will be correlated with blood pressure
   - i.e., high coffee consumption associated correlated with high blood pressure

Our plan to understand bias in non-experimental studies

1. Generate population potential outcomes in R
2. Generate a variable called coffee, which returns 1 for students who consume more than 3 cups of coffee per day, 0 otherwise
   - This variable will be correlated with blood pressure
   - i.e., high coffee consumption associated correlated with high blood pressure
3. Look at what happens if we don't randomize and

Our plan to understand bias in non-experimental studies

1. Generate population potential outcomes in R
2. Generate a variable called coffee, which returns 1 for students who consume more than 3 cups of coffee per day, 0 otherwise
   - This variable will be correlated with blood pressure
   - i.e., high coffee consumption associated correlated with high blood pressure
3. Look at what happens if we don't randomize and students self-select into experimental conditions
   - We will assume that students who drink a lot of coffee self-select into the control (no meditation) condition

Our plan to understand bias in non-experimental studies

1. Generate population potential outcomes in R
2. Generate a variable called coffee, which returns 1 for students who consume more than 3 cups of coffee per day, 0 otherwise
   - This variable will be correlated with blood pressure
   - i.e., high coffee consumption associated correlated with high blood pressure
3. Look at what happens if we don't randomize and students self-select into experimental conditions
   - We will assume that students who drink a lot of coffee self-select into the control (no meditation) condition
4. Understand how regression "controls" / "adjustments" allow to correct for bias

Meditation, stress, and coffee consumption

Let's open R Studio!

Adding covariates to a regression **non-experimental settings**

Why add variables to a regression in non-experimental settings?

Why add variables to a regression in non-experimental settings?

- Basic definition and goal of regression for the bivariate case:
  - Estimate the CEF $\mathbb{E}[Y|X_1]$

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Why add variables to a regression in non-experimental settings?

- Basic definition and goal of regression for the bivariate case:
    - Estimate the CEF $\mathbb{E}[Y|X_1]$

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- When we add a third variable X_2 to the regression:

Why add variables to a regression in non-experimental settings?

- Basic definition and goal of regression for the bivariate case:
  - Estimate the CEF $\mathbb{E}[Y|X_1]$

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- When we add a third variable X_2 to the regression:
  - We estimate the relationship of two variables Y and X, conditional on a third variable Z

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- β's are the population parameters that we want to estimate

Why add variables to a regression in non-experimental settings?

- Causal inference when no random assignment
  - Block potential **confounding**, which are variables that are correlated with both Y and X. Omtting these variables introduces bias in the estimates of the **causal** effect of X on Y and often leads to incorrect causal inferences

Why add variables to a regression in non-experimental settings?

- Causal inference when no random assignment
  - Block potential **confounding**, which are variables that are correlated with both Y and X. Omtting these variables introduces bias in the estimates of the **causal** effect of X on Y and often leads to incorrect causal inferences
  - Relies on strong modeling assumptions! No way to know for sure what the confounding variables are in real world settings!

Why add variables to a regression in non-experimental settings?

- Causal inference when no random assignment
  - Block potential **confounding**, which are variables that are correlated with both Y and X. Omtting these variables introduces bias in the estimates of the **causal** effect of X on Y and often leads to incorrect causal inferences
  - Relies on strong modeling assumptions! No way to know for sure what the confounding variables are in real world settings!
- Descriptive
  - Get a sense for the relationships in the data
  - Describe more precisely our quantity of interest

Why add variables to a regression in non-experimental settings?

- Causal inference when no random assignment
  - Block potential **confounding**, which are variables that are correlated with both Y and X. Omtting these variables introduces bias in the estimates of the **causal** effect of X on Y and often leads to incorrect causal inferences
  - Relies on strong modeling assumptions! No way to know for sure what the confounding variables are in real world settings!
- Descriptive
  - Get a sense for the relationships in the data
  - Describe more precisely our quantity of interest
- Predictive
  - We can usually make better predictions about the dependent variables with more information on independent variables

Why add variables to a regression in non-experimental settings?

- To understand further the basics of multiple regression, let's forget about causality for now and exclusively focus on prediction (& description)

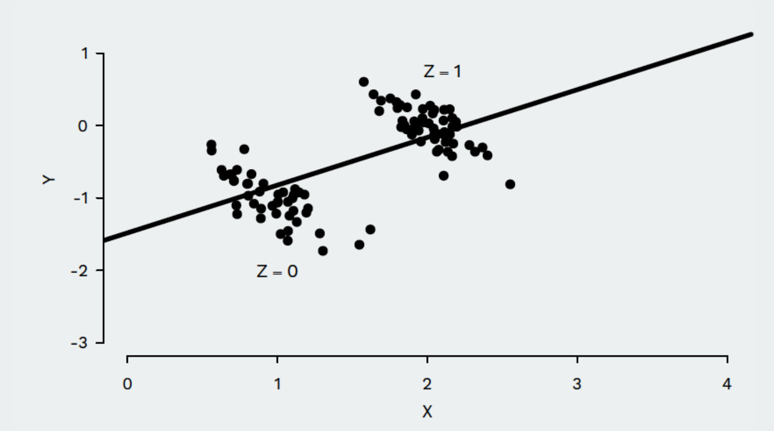
Why add variables to a regression in non-experimental settings?

- To understand further the basics of multiple regression, let's forget about causality for now and exclusively focus on prediction (& description)
  - No random assignment and no causal inference in the following slides!

Why add variables to a regression in non-experimental settings?

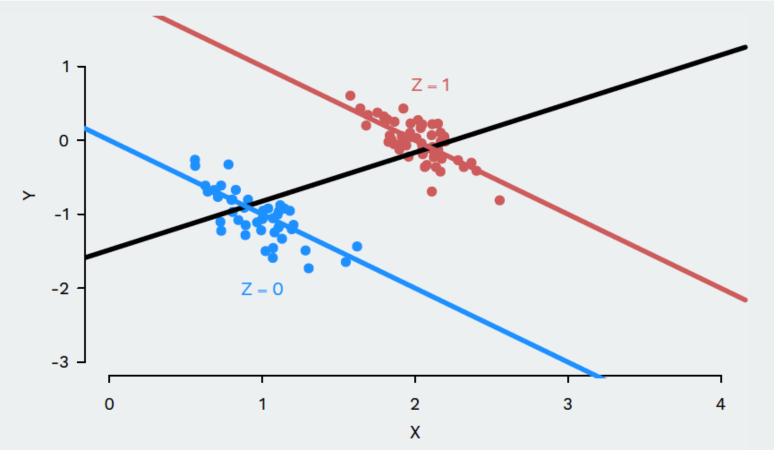
- To understand further the basics of multiple regression, let's forget about causality for now and exclusively focus on prediction (& description)
  - No random assignment and no causal inference in the following slides!
- To understand the relevance of controlling for variables for description / prediction, let's consider Simpson's paradox

Illustration: Simpson's paradox



- Overall a positive relationship between Y and X

Illustration: Simpson's paradox



- Overall a positive relationship between Y and X
- But within levels of Z_i, the opposite

Simpson's paradox: Example

- Cochran (1968) sought to compare cigarette to cigar smoking. He found that cigar smokers had higher mortality rates than cigarette smokers, but at any age level, cigarette smokers had higher mortality than cigar smokers.

Simpson's paradox: Example

- Cochran (1968) sought to compare cigarette to cigar smoking. He found that cigar smokers had higher mortality rates than cigarette smokers, but at any age level, cigarette smokers had higher mortality than cigar smokers.
- Instance of a more general problem called the **ecological inference fallacy**

Regression with Two Explanatory Variables

- Suppose we are interested in the relationship between income and one's propensity to donate to a charity

Regression with Two Explanatory Variables

- Suppose we are interested in the relationship between income and one's propensity to donate to a charity
- Variables of interest
  - Y : Measure of donation in the past year
  - X_1: Measure of income
  - X_2: Children

Regression with Two Explanatory Variables

- Suppose we are interested in the relationship between income and one's propensity to donate to a charity
- Variables of interest
  - Y : Measure of donation in the past year
  - X_1: Measure of income
  - X_2: Children
- With one predictor, we ask: Does income (X_1) predict or explain donation (Y)?

Regression with Two Explanatory Variables

- Suppose we are interested in the relationship between income and one's propensity to donate to a charity
- Variables of interest
  - Y : Measure of donation in the past year
  - X_1: Measure of income
  - X_2: Children
- With one predictor, we ask: Does income (X_1) predict or explain donation (Y)?
- With two predictors, we ask questions like: Does income (X_1) predict or explain donation (Y), once we "control" for children (X_2)?

Regression with Two Explanatory Variables

- Suppose we are interested in the relationship between income and one's propensity to donate to a charity
- Variables of interest
  - Y : Measure of donation in the past year
  - X_1: Measure of income
  - X_2: Children
- With one predictor, we ask: Does income (X_1) predict or explain donation (Y)?
- With two predictors, we ask questions like: Does income (X_1) predict or explain donation (Y), once we "control" for children (X_2)?
- Let's explore what is meant by **controlling for another variable** with linear regression

Simple regression of Donation on Income

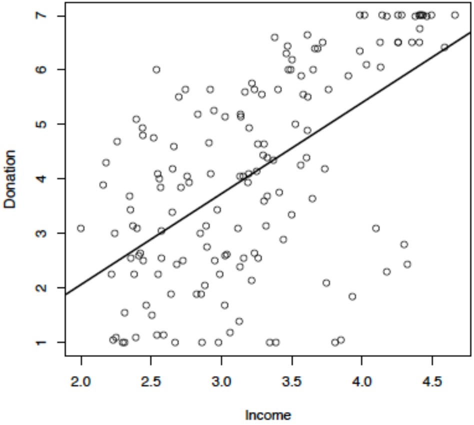
- Let's look at the bivariate regression of donation on income

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$$

Simple regression of Donation on Income

- Let's look at the bivariate regression of donation on income

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$$



Simple regression of Donation on Income

- We observe:

$$\widehat{Donation} = -1.26 + 1.6\, income$$

- Interpretation:

Simple regression of Donation on Income

- We observe:

$$\widehat{Donation} = -1.26 + 1.6\, income$$

- Interpretation: A one point increase in income is associated with a 1.6 point increase in donation

Simple regression of Donation on Income

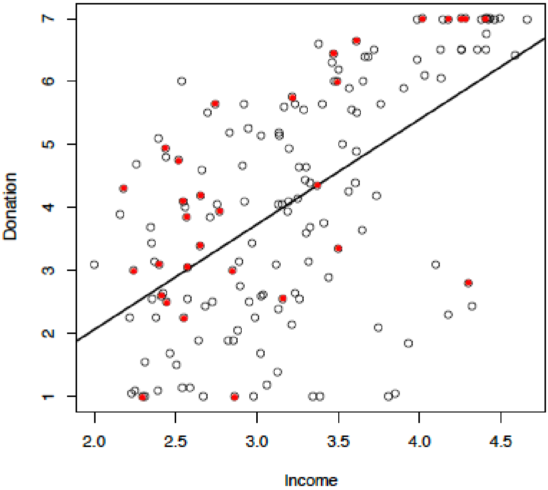
- We observe:

$$\widehat{Donation} = -1.26 + 1.6\, income$$

- Interpretation: A one point increase in income is associated with a 1.6 point increase in donation
- But we can use more information in our prediction equation
  - For example, some observations come from individuals who have children whereas others come from individuals who do not have children
  - And it may be the case that these individuals have different levels of donation

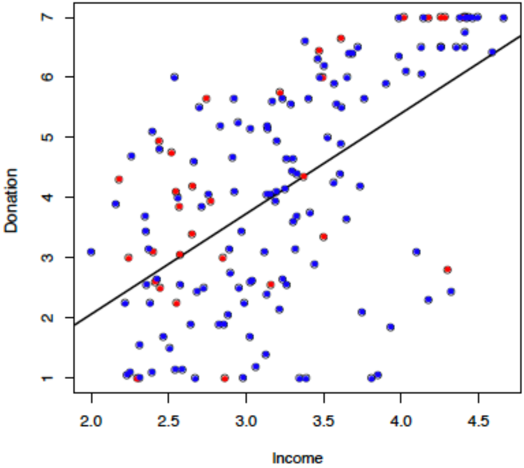
Simple regression of Donation on Income

- Individuals with children (in red) tend donate more



Simple regression of Donation on Income

- Individuals with children (in red) tend donate more
- Individuals without children (in blue) tend donate less



Adding a covariate

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2$$

- This implies that we want to predict Y using the information we have about X_1 and X_2
- In words:

$$\widehat{Donation} = \widehat{\beta}_0 + \widehat{\beta}_1 income + \widehat{\beta}_2 children$$

Interpreting a binary covariate

- Let X_{2_i} indicate whether individual i has children

Interpreting a binary covariate

- Let X_{2_i} indicate whether individual i has children
- When $X_2 = 0$, the model becomes:

$$\begin{aligned}\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1\end{aligned}$$

Interpreting a binary covariate

- Let X_{2_i} indicate whether individual i has children
- When $X_2 = 0$, the model becomes:

$$\begin{aligned}\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1\end{aligned}$$

- When $X_2 = 1$, the model becomes:

$$\begin{aligned}\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_1\end{aligned}$$

Interpreting a binary covariate

- Let X_{2_i} indicate whether individual i has children
- When $X_2 = 0$, the model becomes:

$$\begin{aligned} \widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 \end{aligned}$$

- When $X_2 = 1$, the model becomes:

$$\begin{aligned} \widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_1 \end{aligned}$$

- What does this mean?

Interpreting a binary covariate

- Let X_{2_i} indicate whether individual i has children
- When $X_2 = 0$, the model becomes:

$$\begin{aligned}\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1\end{aligned}$$

- When $X_2 = 1$, the model becomes:

$$\begin{aligned}\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_1\end{aligned}$$

- What does this mean? We are fitting two lines with the **same slope** but **different intercepts**

Regression of donation on income and children

- Suppose multiple regression model provides estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ such that:
  - $\widehat{\beta}_0 = -1.5060$
  - $\widehat{\beta}_1 = 1.7059$
  - $\widehat{\beta}_2 = 0.58$

Regression of donation on income and children

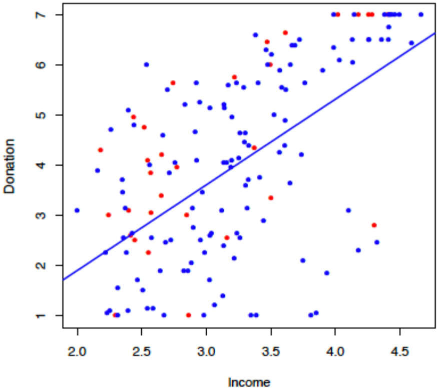
- Individuals without children:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$$
$$\widehat{Y} = -1.5 + 1.7X_1$$

Regression of donation on income and children

- Individuals without children:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$$
$$\widehat{Y} = -1.5 + 1.7X_1$$



Regression of donation on income and children

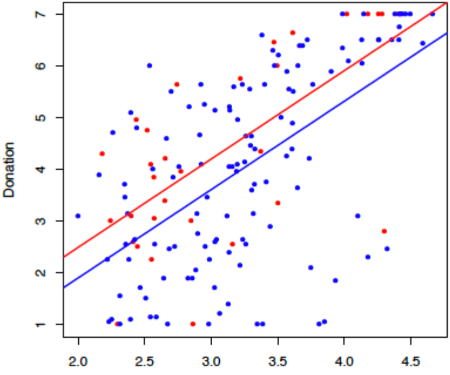
- Individuals with children:

$$\widehat{Y} = (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_1$$
$$\widehat{Y} = -.92 + 1.7X_1$$

Regression of donation on income and children

- Individuals with children:

$$\widehat{Y} = (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_1$$
$$\widehat{Y} = -.92 + 1.7X_1$$

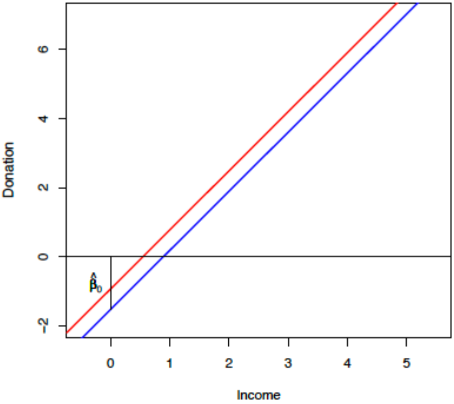


Regression of donation on income and children

- Our prediction equation is: $\widehat{Y} = -1.5 + 1.7X_1 + .58X_2$
  - Where do these quantities appear on the graph?

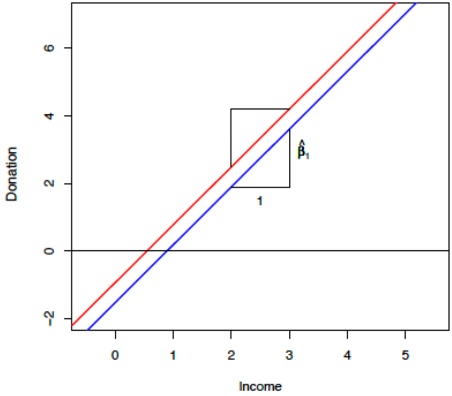
Regression of donation on income and children

- Our prediction equation is: $\widehat{Y} = -1.5 + 1.7X_1 + .58X_2$
  - $\widehat{\beta}_0 = -1.5$ is the intercept for the prediction line for individuals without children



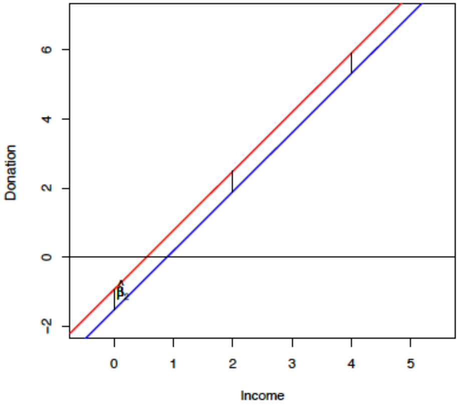
Regression of donation on income and children

- Our prediction equation is: $\widehat{Y} = -1.5 + 1.7X_1 + .58X_2$
  - $\widehat{\beta}_1 = 1.7$ is the slope for both lines



Regression of donation on income and children

- Our prediction equation is: $\widehat{Y} = -1.5 + 1.7X_1 + .58X_2$
  - $\widehat{\beta}_2 = .58$ is the vertical distance between two lines for individuals with and without children



Omitted variable bias

- In observational studies, causal claims about the relationship between two variables Y and X_1 can be made if we assume that confounding variables are "controlled for"
  - Confounding variables are variables correlated with both Y and X_1

Omitted variable bias

- In observational studies, causal claims about the relationship between two variables Y and X_1 can be made if we assume that confounding variables are "controlled for"
  - Confounding variables are variables correlated with both Y and X_1
- Not controlling for confounding variables introduces a specific type of bias in the coefficient of interest

Omitted variable bias

- In observational studies, causal claims about the relationship between two variables Y and X_1 can be made if we assume that confounding variables are "controlled for"
  - Confounding variables are variables correlated with both Y and X_1
- Not controlling for confounding variables introduces a specific type of bias in the coefficient of interest
- ... called **omitted variable bias**!

Omitted variable bias

- For one confounder X_2:

	$\text{cov}(X_1, X_2) > 0$	$\text{cov}(X_1, X_2) < 0$	$\text{cov}(X_1, X_2) = 0$
$\beta_2 > 0$	Positive bias	Negative Bias	No bias
$\beta_2 < 0$	Negative bias	Positive Bias	No bias

Omitted variable bias

- For one confounder X_2:

	$\text{cov}(X_1, X_2) > 0$	$\text{cov}(X_1, X_2) < 0$	$\text{cov}(X_1, X_2) = 0$
$\beta_2 > 0$	Positive bias	Negative Bias	No bias
$\beta_2 < 0$	Negative bias	Positive Bias	No bias

- In observational datasets, omitted variable bias is often generated by more than one variable. In this more general case, the direction of the bias is more difficult to discern. It depends on all the pairwise correlations.