POLS 201: Data Analysis and Politics

Professor Elena Llaudet

Lecture 14 | Controlling for Confounders Using Multiple Linear Regression

Plan for Today

- How Can We Estimate Causal Effects with Observational Data?
- Multiple Linear Regression Models
  - Interpretation of Coefficients
  - Interpretation of $\widehat{\beta_1}$ When X_1 Is the Treatment Variable and the Other X Variables Are All the Potential Confounding Variables
- What is the Effect of the Death of the Leader on the Level of Democracy?

How Can We Estimate Causal Effects with Observational Data?

- We cannot rely on random treatment assignment to eliminate potential confounders and make treatment and control groups comparable
- First, we must identify all potential confounding variables
  - variables that affect both (i) the likelihood of receiving the treatment and (ii) the outcome

$$\begin{array}{ccc} & Z & \\ \swarrow & & \searrow \\ X & \longrightarrow & Y \end{array}$$

- Then, we need to statistically control for them by fitting a multiple linear regression model

Multiple Linear Regression Models

Linear models with more than one X variable

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_{i1} + ... + \widehat{\beta}_p X_{ip}$$

where:

- $\widehat{Y}_i$ is the predicted value of Y for observation i
- $\widehat{\alpha}$ is the estimated intercept coefficient
- each $\widehat{\beta}_j$ (pronounced beta hat sub j) is the estimated coefficient for variable X_j (j=1, ..., p)
- each X_{ij} is the observed value of the variable X_j for observation i (j=1, ..., p)
- p is the total number of X variables in the model.

simple regression $\widehat{Y} = \widehat{\alpha} + \widehat{\beta} X$	multiple regression $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}_1 X_1 + ... + \widehat{\beta}_p X_p$
$\widehat{\alpha}$: $\widehat{Y}$ when X=0	$\widehat{\alpha}$: $\widehat{Y}$ when all X_j=0 (j=1, ..., p)
$\widehat{\beta}$: $\triangle\widehat{Y}$ associated with $\triangle X$=1	each $\widehat{\beta}_j$: $\triangle\widehat{Y}$ associated with $\triangle X_j$=1, while holding all other X variables constant or *ceteris paribus*

Interpretation of Coefficients in Multiple Linear Regression Models

- $\widehat{\alpha}$ is the $\widehat{Y}$ when *all* X_j=0
- Because there are multiple X variables, there are multiple $\widehat{\beta}$ coefficients (one for each X variable)
- Each $\widehat{\beta}_j$ is the $\triangle\widehat{Y}$ associated with $\triangle X_j$=1, *while holding all other X variables constant*

Interpretation of $\widehat{\beta_1}$ When X_1 Is the Treatment Variable and the Other X Variables Are All the Potential Confounding Variables

- Adding all confounders as controls in the model makes treatment and control groups comparable *after controls*
- As a result, we can interpret $\widehat{\beta_1}$ using **causal langauge**
- $\widehat{\beta_1}$ is the $\triangle\widehat{Y}$ *caused by* the presence of the treatment ($\triangle X_1$=1), while holding all confounders constant
- $\widehat{\beta_1}$ should be a valid estimate of the average treatment effect if all confounding variables are in the model

Does the Death of the Leader Increase the Level of Democracy?

(Based on Benjamin F. Jones and Benjamin A. Olken. 2009. "Hit or Miss? The Effect of Assassinations on Institutions and War." *American Economic Journal: Macroeconomics*, 1 (2): 55-87.)

- We will answer, by analyzing observational data
- Dataset on assassinations and assassination attempts against political leaders from 1875 to 2004
- To begin with, let's consider that, after an assassination attempt, the death a leader is close to random and, thus, leaders whose assassination attempt succeeded should be, on average, comparable to leaders whose assassination attempt failed
- If this is true, we can estimate the average causal effect of the death of the leader by computing the diffs-in-means estimator
- As we saw in the last class, we can compute the difference-in-means estimator by fitting a simple linear model where X is the treatment variable

The *leaders* dataset

variable	description
year	year of the assassination attempt
country	name of the country where the assassination attempt took place
leadername	name of the leader whose life was at risk in the assassination attempt
died	whether the leader died as a result of the assassination attempt: 1=yes, 0=no
politybefore	polity scores of the country before the assassination attempt (in points, in a scale from -10 to 10)
polityafter	polity scores of the country after the assassination attempt (in points, in a scale from -10 to 10)

In-Class Exercise: What is the Effect of the Death of the Leader on the Level of Democracy?

1. Open RStudio
2. Open **exercise_4.R** from within RStudio
3. Run steps 1 through 3

```
## STEP 1: Set the working directory to DSS folder
setwd("~/Desktop/DSS") #if Mac
setwd("C:/user/Desktop/DSS") #if Windows
```

```
## STEP 2: Load the dataset
leaders <- read.csv("leaders.csv") # reads and stores data
```

```
## STEP 3: Understand the data
head(leaders)# shows first observations
##   year     country      leadername died politybefore polityafter
## 1 1929 Afghanistan Habibullah Ghazi    0           -6          -6
## 2 1933 Afghanistan      Nadir Shah    1           -6          -7
## 3 1934 Afghanistan     Hashim Khan    0           -6          -8
## 4 1924     Albania            Zogu    0            0          -9
## 5 1931     Albania            Zogu    0           -9          -9
## 6 1968     Algeria     Boumedienne    0           -9          -9
```

- the treatment variable (X) is *died*
- the outcome variable (Y) is *polityafter*

STEP 4: Compute difference-in-means estimator

- To fit the simple linear model where $\widehat{\beta}$ is equivalent to the difference-in-means estimator, we run:

```
lm(leaders$polityafter ~ leaders$died) # or
```

```
lm(polityafter ~ died, data=leaders)
##
## Call:
## lm(formula = polityafter ~ died, data = leaders)
##
## Coefficients:
## (Intercept)         died
##      -1.895        1.132
```

- Fitted model: $\widehat{polityafter}$ = -1.90 + 1.13 *died*

- Interpretation of $\widehat{\beta}$?
  - definition: $\widehat{\beta}$ is the $\triangle\widehat{Y}$ associated with $\triangle X$=1
  - here: $\widehat{\beta}$ = 1.13 is the $\triangle\widehat{polityafter}$ associated with $\triangle died$=1
  - in words: the death of the leader (i.e., an increase in *died* of 1 by going from died=0 to died=1) is associated with a predicted increase in polity scores after the assassination attempt of 1.13 points, on average
- unit of measurement of $\widehat{\beta}$? same as $\triangle\overline{Y}$; here, Y is nonbinary and measured in points so $\triangle\overline{Y}$ is measured in points and so is $\widehat{\beta}$

- Interpretation of $\widehat{\beta}$? (continuation)
  - Since here X is the treatment variable and Y is the outcome variable of interest, $\widehat{\beta}$ is equivalent to the difference-in-means estimator so we should interpret $\widehat{\beta}$ using **causal langauge**
  - Causal language: We estimate that the death of the leader ***increases*** polity scores after the assassination attempt by 1.13 points, on average
- This should be a valid estimate of the average treatment effect if the assassination attempts where the leader died are comparable to those where the leader did not die
- Is this true? Let's see how the two groups compare to each other in terms of *politybefore* (a pre-treatment characteristic)

STEP 5: Identify potential confounding variables

- Calculate the average *politybefore* for the two groups:

```
mean(leaders$politybefore[leaders$died==1]) #treatment
## [1] -0.7037037
mean(leaders$politybefore[leaders$died==0]) # control
## [1] -1.743197
```

- Assassination attempts where the leader ended up dying were more democratic to begin with (their average *politybefore* was less negative)
  - *politybefore* might be a confounding variable:

$$\begin{tikzcd} & politybefore \arrow[dl] \arrow[dr] & \\ died \arrow[rr] & & polityafter \end{tikzcd}$$

STEP 6: Estimate average causal effect while controlling for confounders

- To estimate the average causal effect of the death of the leader while controlling for initial levels of democracy, we need to fit the following multiple regression linear model:

$$\widehat{polityafter} = \widehat{\alpha} + \widehat{\beta}_1\ died + \widehat{\beta}_2\ politybefore$$

- To fit the model, we use the function `lm()`
  - but now we specify as the main argument a formula of the type $Y \sim X_1 + X_2$

```
lm(leaders$polityafter ~ leaders$died + leaders$politybefore) # or
```

```
lm(polityafter ~ died + politybefore, data=leaders)
##
## Call:
## lm(formula = polityafter ~ died + politybefore, data = leaders)
##
## Coefficients:
## (Intercept)         died  politybefore
##     -0.4346       0.2616        0.8375
```

- Fitted model:

$$\widehat{polityafter} = -0.43 + 0.26\ died + 0.84\ politybefore$$

- Interpretation of $\widehat{\beta}_1$?
  - definition: $\widehat{\beta}_1$ is the $\triangle\widehat{Y}$ associated with $\triangle X_1$=1, while holding all other X variables constant
  - here: $\widehat{\beta}_1 = 0.26$ is the $\triangle\widehat{\textit{polityafter}}$ associated with $\triangle \textit{died}$=1, while holding *politybefore* constant
  - in words: the death of the leader is associated with a predicted increase in polity scores after the assassination attempt of 0.26 points, on average, while holding polity scores before constant
- unit of measurement of $\widehat{\beta}_1$? same as $\triangle\overline{Y}$; here, Y is nonbinary and measured in points so $\triangle\overline{Y}$ is measured in points and so is $\widehat{\beta}_1$

- Note that once we control for *politybefore* the effect size decreases substantially (it goes from 1.13 to 0.26)
- Based on this analysis, the death of the leader increases the level of democracy of a country but by a very small amount
  - more on this later in the semester

Today's Class
- How to Use Multiple Linear Regression Models to Control for Confounders and Estimate Average Treatment Effects Using Observational Data

Next Class
- Internal vs. External Validity
- **No computers needed**

- Interpretation of $\widehat{\beta}_1$? (continuation)
  - Since here X_1 is the treatment variable, Y is the outcome variable of interest, and X_2 is the confounder we are worried about, we can interpret $\widehat{\beta}_1$ using **causal langauge**
  - Causal language: We estimate that the death of the leader ***increases*** polity scores after the assassination attempt by 0.26 points, on average, when holding polity scores before the assassination attempt constant
- This should be a valid estimate of the average treatment effect if *politybefore* is the only confounder

ESTIMATING AVERAGE CAUSAL EFFECTS USING OBSERVATIONAL DATA AND MULTIPLE LINEAR REGRESSION MODELS. If, in the multiple linear regression model where X_1 is the treatment variable, we control for *all* potential confounders by including them in the model as additional X variables, then we can interpret $\widehat{\beta}_1$ as a valid estimate of the average causal effect of X on Y.