Problem Set 4

Fernanda Valdez

2025-11-22

PART 1: READING

1. What is the difference between a confounder and a collider? How should you address each in your models?

A confounder is when Z causes both X and Y. In your model, you need to control for Z. If you do not control for Z, then Z will also affect Y through X. This will bias your result and that is what we do not want to happen in our model. It will obscure the relationship between X and Y.

A collider is in the opposite direction when X and Y both affect Z. In contrast to confounders, one does not control for Z because this could cause a biased effect on Y.

2. How can conditioning on a collider create bias?

As previously noted, conditioning on a collider can create a correlation between X and Y which is also called collider-stratification bias when there is no correlation between X and Y. Essentially, conditioning on a collider can create a spurious relationship between X and Y.

3. Why can't statistical summaries or correlations alone tell us whether to control for a variable?

Statistical summaries cannot tell you when a variable is a mediator, for example. By using a DAG, you can see when to remove or include a variable in your model. Essentially, statistical summaries only show you what you tell them. So, if you don't tell your model what variable is a confounder or what to control for, then it is just going to compute what you instruct it and could result in a spurious relationship.

4. What is meant by a "kitchen sink" regression, and what is wrong with this approach to modeling?

A kitchen sink regression is when you include many variables into your regression model which may or may not be relevant to your outcome of interest. If you include variables that are not relevant to your outcome of interest, then you risk might obscure the direction of the variables' relationships, have no useful causal interpretations, inflate errors in hypothesis testing, and risk overfitting your model.

5. What is a "backdoor path" and how does multiple regression help block these paths?

This when the pathway that runs from the confounder, z, to X. Multiple regression allows you to include confounders and close those "backdoor paths" in order to isolate X's effect on Y.

PART 2: SIMULATION

Here, I simulate a social causal relationship between educational attainment and voter turnout. The following are my variables: Social causal relationship: education and voter turnout X: Education (x) Y: Increase voter turnout (y) Confounder: age (z) Mediator: civic engagement (m) Collider: social pressure (c) Instrument: extra credit (i) Exogenous effect on Y: poll hack (e)

Generate random data for variables that are not causally affected by any others in your DAG (confounder and exogenous variables).

```r
set.seed(769)
# sample size
n <- 5000
```

```r
# confounder: age
z <- rnorm(n)
# exogenous effect on Y: polls are hacked
e <- rnorm(n)
# instrumental variable: extra credit (i)
i <- rnorm(n)
```

Generate the remaining variables as linear functions of the variables that causally affect them. Each linear function should have beta coefficients that represent the true effect size, and a random error term.

```r
# Error term
e_t <- rnorm(n)

#Linear function for treatment variable (education)
x <- 0.4*z + rnorm(n)

#Linear function for mediator (civic engagement)
m <- 0.5*x + rnorm(n)

#Linear function for outcome variable (voter turnout)
y <- 0.6*x + 0.7*m + 0.8*z + rnorm(n) + e_t

#Linear function for collider (social pressure)
c <- 0.9*x + 0.10*y + rnorm(n)
```

1. Fit a model that recovers the direct effect of the treatment on the outcome variable. Which variables are necessary to recover the direct effect?

To recover the direct effect, I would need to include a confounder and mediator which in this case is age and civic engagement, respectively.

```r
# Data frame for variables
df <- data.frame(education = c(x),
                 voter_turnout = c(y),
                 age = c(z),
                 social_pressure = c(c),
                 civic_engagement = c(m),
                 extra_credit = c(i),
                 poll_hack = c(e))

# Model for direct effect

model <- lm (voter_turnout ~ education + age + civic_engagement, data=df)
summary(model)
```

```
##
## Call:
## lm(formula = voter_turnout ~ education + age + civic_engagement,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8809 -0.9357 -0.0068  0.9560  5.1117
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.006752   0.020028   0.337    0.736
## education        0.598935   0.022301  26.857   <2e-16 ***
## age              0.795405   0.021769  36.538   <2e-16 ***
## civic_engagement 0.724895   0.019657  36.878   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 4996 degrees of freedom
## Multiple R-squared:  0.5841, Adjusted R-squared:  0.5838
## F-statistic:  2339 on 3 and 4996 DF,  p-value: < 2.2e-16
```

2. Fit a model that recovers the total effect of the treatment on the outcome variable. How does your model change to estimate the total effect? To recover the total effect of the treatment on the outcome variable, you should not control for the mediator (social pressure) in your model.

```r
model <- lm (voter_turnout ~ education + age, data=df)
summary(model)
```

```
##
## Call:
## lm(formula = voter_turnout ~ education + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1319 -1.0807  0.0001  1.0999  5.8139
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01518    0.02258  -0.672    0.501
## education    0.95850    0.02262  42.375   <2e-16 ***
## age          0.79328    0.02455  32.311   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.596 on 4997 degrees of freedom
## Multiple R-squared:  0.4709, Adjusted R-squared:  0.4707
## F-statistic:  2223 on 2 and 4997 DF,  p-value: < 2.2e-16
```

3. How do your results change when you control for the collider, the exogenous independent variable, or the instrument (individually, not all simultaneously)?

```r
#collider
model <- lm (voter_turnout ~ education + age + social_pressure, data=df)
summary(model)
```

```
##
## Call:
## lm(formula = voter_turnout ~ education + age + social_pressure,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1978 -1.0583 -0.0001  1.0968  5.8708
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -0.01524    0.02223  -0.686    0.493    
## education        0.67980    0.03133  21.701   <2e-16 ***
## age              0.77234    0.02423  31.880   <2e-16 ***
## social_pressure  0.27976    0.02212  12.649   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.571 on 4996 degrees of freedom
## Multiple R-squared:  0.4873, Adjusted R-squared:  0.487 
## F-statistic:  1583 on 3 and 4996 DF,  p-value: < 2.2e-16
```

If I control for the collider, then I will open the backdoors to my coefficient for education and it becomes biased.

```
#exogenous independent variable
model <- lm (voter_turnout ~ education + age + poll_hack, data=df)
summary(model)
```

```
## 
## Call:
## lm(formula = voter_turnout ~ education + age + poll_hack, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.1038 -1.0798  0.0012  1.1029  5.7826 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.01541    0.02257  -0.683   0.4948    
## education    0.95979    0.02263  42.421   <2e-16 ***
## age          0.79322    0.02455  32.316   <2e-16 ***
## poll_hack    0.04103    0.02274   1.805   0.0712 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.595 on 4996 degrees of freedom
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.4709 
## F-statistic:  1484 on 3 and 4996 DF,  p-value: < 2.2e-16
```

If I control for my exogenous variable, then there won't be a lot of change in my model because it is an exogenous variable only for the outcome variable (y) which in this case is voter turnout.

```
#instrumental variable
model <- lm (voter_turnout ~ education + age + extra_credit, data=df)
summary(model)
```

```
## 
## Call:
## lm(formula = voter_turnout ~ education + age + extra_credit, 
##     data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.1511 -1.0781  0.0015  1.1024  5.8180 
## 
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.015210   0.022580  -0.674    0.501
## education     0.958560   0.022623  42.372   <2e-16 ***
## age           0.793382   0.024558  32.307   <2e-16 ***
## extra_credit  0.005628   0.022846   0.246    0.805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.596 on 4996 degrees of freedom
## Multiple R-squared:  0.4709, Adjusted R-squared:  0.4706
## F-statistic:  1482 on 3 and 4996 DF,  p-value: < 2.2e-16
```

If I control for my instrumental variable (extra credit), then this can help isolate my variation in my treatment variable.

4. Given the reading and simulation results, how should you choose which variable to include in a model? In a model, I would always include confounders. I would not include colliders, but I would control for mediators (if I want to measure a direct effect).