Problem Set 4

Neeraj Tom Savio

2025-12-11

Part 1: Reading

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

A: Confounders and colliders differ in their directional effects. Confounders affect both the treatment and outcome whereas colliders are affected by both the treatment and outcome. Confounders should be controlled for and colliders shoud not be included in the models.

2. How can conditioning on a collider create bias?

A: Conditioning on a collider opens up a causal path between the treatment and the outcome where one doesn't exist as both X and Y have an effect on Z, so conditioning on Z would introduce bias.

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

A: Statistical summaries and correlations can't tell us about the direction of the relationships between variables. They can't tell us if a variable is a mediator or confounder or collider, therefore we don't know if a variable should be controlled for or not.

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

A: A kitchen sink regression means one where we include all possible predictors in order to explain variation in the model. This is not good because rather than using data to test some theory, we are using the data to fit the theory and thus isn't good science.

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

A: A backdoor path essentially means the non-causal relationship that is created between X and Y that is created due to a confounder. Multiple relationship helps block these paths by giving us the effect of X on Y conditioned on a constant value of the confounder therefore breaking the non-causal relationship.

Part 2: Simulation

```
## Introductory Code

rm(list=ls()) #Clearing the environment
setwd("~/Desktop/POLS 602/R Files") #Setting up the working directory
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

For the social causal relationship, I simulate a relationship between relationship between no. of years in education and net worth. I assume all my variables have been converted to standard normal distributions.

```
set.seed(6767)

z <- rnorm(1000, 0, 1) #Confounder (Parental Income)
it <- rnorm(1000, 0, 1) #Exogenous effect on treatment (rainful in cubic meters)
x <- rnorm(1000, 0, 1)  + z + it  #Treatment (Education)
m <- rnorm(1000, 0, 1) + x #Mediator (Yearly income)
io <- rnorm(1000, 0, 1) #Exogenous effect on outcome (Inheritance in dollars)
y <- rnorm(1000, 0, 1) + 0.4*x + 0.06*z + 0.01*m + 0.02*io #Outcome variable (Net worth)
c <- rnorm(1000, 0, 1) + 6*x + 7*y  #Collider (No. of dependents)

pop <- data.frame(
  years_education = x,
  net_worth = y,
  parental_income = z,
  yearly_income = m,
  n_dependents = c,
  rainfall = it,
  inheritance = io)
```

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

```
model1 <- lm(y ~ x + m + z, data = pop) #Fitting the linear model
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x + m + z, data = pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5045 -0.7238  0.0002  0.7015  2.7684
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0009506  0.0326139  -0.029    0.977
## x            0.3930617  0.0392618  10.011   <2e-16 ***
## m            0.0296113  0.0320098   0.925    0.355
## z            0.0757238  0.0397313   1.906    0.057 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 996 degrees of freedom
## Multiple R-squared:  0.3721, Adjusted R-squared:  0.3702
## F-statistic: 196.7 on 3 and 996 DF,  p-value: < 2.2e-16
```

In order to recover the direct effect, we need to include the mediator and confounder.

Question b. 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?

```
model2 <- lm(y ~ x + z, data = pop)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x + z, data = pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5135 -0.7237  0.0054  0.6937  2.7769
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002148   0.032586  -0.066   0.9475
## x            0.422644   0.022777  18.556   <2e-16 ***
## z            0.074366   0.039701   1.873   0.0613 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 997 degrees of freedom
## Multiple R-squared:  0.3715, Adjusted R-squared:  0.3703
## F-statistic: 294.7 on 2 and 997 DF,  p-value: < 2.2e-16
```

In order to recover the total effect of the treatment on the outcome variable, we need to drop the mediator from our linear model since the treatment affects the outcome through the mediator.

Dropping the mediator from the model increases the effect attributed to x.

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

```
model3 <- lm(y ~ x + io, data = pop) # Controlling for the exogenous..
#..independent variable
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x + io, data = pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5770 -0.7321  0.0149  0.7111  2.7219
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002566   0.032639  -0.079    0.937
## x            0.446351   0.018553  24.059   <2e-16 ***
## io           0.032919   0.033082   0.995    0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.031 on 997 degrees of freedom
## Multiple R-squared:   0.37,  Adjusted R-squared:  0.3687
## F-statistic: 292.7 on 2 and 997 DF,  p-value: < 2.2e-16
```

Controlling for the exogenous independent variable does not have a lot of effect on the coefficient for x as it is exogenously related to y.

```
model4 <- lm(y ~ x + it, data = pop) #Controlling for the instrumental variable
summary(model4)
```

```
##
## Call:
## lm(formula = y ~ x + it, data = pop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5808 -0.7383 -0.0054  0.6818  2.7609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004768   0.032592  -0.146    0.884
## x            0.469426   0.022774  20.613   <2e-16 ***
## it          -0.066641   0.040490  -1.646    0.100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 997 degrees of freedom
## Multiple R-squared:  0.371,  Adjusted R-squared:  0.3698
## F-statistic: 294.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

Controlling for the instrumental variable increased the avolsute value of the intercept whereas the coefficient for x was relatively unaffected.

```
model5 <- lm(y ~ x + c, data = pop) # Controlling for the collider
summary(model5)
```

```
##
## Call:
## lm(formula = y ~ x + c, data = pop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39227 -0.09697  0.00192  0.09988  0.47144
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.0004355  0.0044785    0.097    0.923
```

```
## x           -0.8355065  0.0061775 -135.249   <2e-16 ***
## c            0.1404444  0.0006163  227.888   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1416 on 997 degrees of freedom
## Multiple R-squared:  0.9881, Adjusted R-squared:  0.9881
## F-statistic: 4.147e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

Controlling for the collider increased the absolute value coefficient of x considerably, implying a much greater value for the average value of x on y than actually exists.

Question d. Given the reading and simulation results, how should you choose which variables to include in

a model?
I would always include confounders. I would include mediators if I want to observe the direct effect of the treatment on the outcome but I would never include a collider.