Problem set 4

Nathaniel Williams

2025-12-02

Part 1 Reading

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

- A confounder is something that influences both the treatment and outcome variable of your model ie X->Y but Z->X & Z->Y
- You should adjust your model to account for a confounder, otherwise you will have a biased result. We condition on the collider to cut off the backdoor path
- A Collider is something that is impacted by both your treatment and outcome variable ie X->Collider & Y-> Collider
- You should not condition on a Collider, doing so will cause a biased result that introduces correlation to X and Y that is not there.

2. How can conditioning on a collider create bias?

- The two variables are not correlated on the pathway to the collider but conditoning on it causes them to be correlated. It causes some spurious association between two otherwise non related variables.

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

You first need to understanding the type of relationship you wish to access prediction or explanation. Then you need to thoughtfully access the directional assocation of variables. This requires common sense and consideration on the part of the researcher as they design their model. Throwing every possible variable into your model can open many pathways that are not actually present and misrepresent both the direction and the size of effects.

4. What is meant by a kitchen sink regression, and what is wrong with it.

A kitchen sink regression is simply throwing all possible variables you can think of into your model, regardless of considering whether they are or arent actually related to your outcome. Which is then followed by some variable selection strategy to impact you p-value and posit a significant relationship. Other studies that then rely on the findings of these papers further propogate the spurious arguments made.

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

A backdoor path is a non causal path that creates a correlation between two variables even if there was not one. Conditioning on this variable in a multiple regression allows us to "block" the pathway and get an unbiased effect estimate (assuming all else potential confounders are controlled for).

Part 2: Simulation

Need:

Treatment = X = Costly Weather events = CWE

Outcome = Y = Votes for green party = GreenVote

Confounder = State Infrastructure = SI

Mediator = Population/demographics = demo

Collider = Environmental Policy = EP

ExogX(instrument) = Temperature = temp

ExogY = Electoral System = ES

The theory here is that voters who experience costly weather events will be more likely to consider a green party vote. State infrastructure/weather prep can effect how costly a weather event is. A Snow storm in east Texas has more economic effects than one in Michigan where it is common and expected, and thus prepared for. But better weather pre or state infrastructure may be caused by environmentally conscious voters that prefer to vote green party. Environmental policy changes can be caused by a recent costly weather event that results in change, or by the election of green party candidates who push that policy. The electoral system may have an effect on green party vote because it decides whether it is a viable party, green parties get more votes in multi-party governments (more than 2) or proportional representation systems.

Demographics are a mediator in that the occurrence of costly weather events may impact peoples decision to move somewhere, and these people may have certain political preferences.

Temperature is the instrument of choice in that above average or below average temperatures during election months may force voters to think about current environmental conditions. Particularity hot days might cause memories of droughts or particularly cold days of recent damaging snow events, high humidity with flooding and so forth.

```
set.seed(9115)
n<- 500
# First make the unaffected variables
#confounder
SI<-rnorm(n, mean=20, sd=1)
#instrument
temp<-rnorm(n, mean=50, sd=1)
#Exog Y
ES<-rnorm(n, mean=85, sd=1)
#Make the x, it is affected bu the confounder and instrument
CWE<- temp*.3 + SI*.5 + rnorm(n)
#Mediator Is caused by X
demo<- CWE*.7 + rnorm(n)
# Y is impacted by confounder, X, exgoenous variable and the mediator
GreenVote<- CWE*.4 + SI*.2 + ES* .3 + demo*.15 + rnorm(n)
#Collider is impacted by both x and y
EP<- CWE*.3 + GreenVote*.4 + rnorm(n)
```

```
# 1. Fit a model that recovers the direct affect of the treatment on the outcome
ModelT<-lm(GreenVote~CWE+SI+demo+temp)
summary(ModelT)
```

```
##
## Call:
## lm(formula = GreenVote ~ CWE + SI + demo + temp)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9728 -0.7351 -0.0337  0.7069  3.3204
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.51298    2.55831   9.191  < 2e-16 ***
## CWE          0.40941    0.05997   6.827 2.54e-11 ***
## SI           0.19575    0.05642   3.470 0.000567 ***
## demo         0.18003    0.04789   3.760 0.000191 ***
## temp         0.02541    0.04963   0.512 0.608849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.084 on 495 degrees of freedom
## Multiple R-squared:  0.3356, Adjusted R-squared:  0.3302
## F-statistic: 62.51 on 4 and 495 DF,  p-value: < 2.2e-16
```

When Condition on X, the Confounder, the mediator and the experiment we get the proper direct effect .4

```
#2. Total effect
ModelTE<-lm(GreenVote~CWE+demo)
summary(ModelTE)
```

```
##
## Call:
## lm(formula = GreenVote ~ CWE + demo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1446 -0.7362 -0.0106  0.7722  3.4617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.56275    1.04531  25.411  < 2e-16 ***
## CWE          0.49250    0.05340   9.223  < 2e-16 ***
## demo         0.18385    0.04807   3.824 0.000148 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.095 on 497 degrees of freedom
## Multiple R-squared:  0.3194, Adjusted R-squared:  0.3167
## F-statistic: 116.6 on 2 and 497 DF,  p-value: < 2.2e-16
```

To get the total effect you only condition on the confounder, this is the average causal effect of X on Y for a one unit increase. Mediator and instrument were removed from the model

```
#3. How do results change when accounting for collider, exog iv, or the instrument (individually not al
ColliderModel<-lm(GreenVote~CWE+EP)
summary(ColliderModel)
```

```
##
## Call:
## lm(formula = GreenVote ~ CWE + EP)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1750 -0.7082 -0.0031  0.6925  2.8628
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.68890    1.08755   19.94   <2e-16 ***
## CWE          0.40356    0.04444    9.08   <2e-16 ***
## EP           0.42289    0.04250    9.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.015 on 497 degrees of freedom
## Multiple R-squared:  0.4158, Adjusted R-squared:  0.4134
## F-statistic: 176.9 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
ExogIvModel<-lm(GreenVote~CWE+ES)
summary(ExogIvModel)
```

```
##
## Call:
## lm(formula = GreenVote ~ CWE + ES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0466 -0.7211 -0.0359  0.8106  3.4064
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.21991    4.22007  -0.763    0.446
## CWE          0.61105    0.04040  15.123  < 2e-16 ***
## ES           0.35354    0.04856   7.280 1.31e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 497 degrees of freedom
## Multiple R-squared:  0.3669, Adjusted R-squared:  0.3644
## F-statistic:   144 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
InstrModel<-lm(GreenVote~CWE+temp)
summary(InstrModel)
```

```
##
## Call:
## lm(formula = GreenVote ~ CWE + temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0287 -0.7517 -0.0504  0.7927  3.3323
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.53005    2.44632  11.254   <2e-16 ***
## CWE          0.62408    0.04405  14.167   <2e-16 ***
## temp        -0.02080    0.04998  -0.416    0.678
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 497 degrees of freedom
## Multiple R-squared:  0.2997, Adjusted R-squared:  0.2968
## F-statistic: 106.3 on 2 and 497 DF,  p-value: < 2.2e-16
```

The Collider model suggest that there is a signifcant relationship where the collider (environmental policy) is causing .4 of the Green party vote. But in reality it is green party vote that is causing the policy. This is a false relationship.

The exogneous Y variable model causes the predicted relationship between weather events and green party vote to be far greater than it actually is (.6 vs .4)

Accounting for only weather events and the temperature again over inflates the relationship, but also suggest a negative relationship between vote and temperature, when we know that the instrument causes part of weather events (x) and therefore does have a positive relationship with green party vote (Y)

4. How to choose what variables to include in the model.

Choosing what variables to include depends on whether we are interested in the total effect of weather on voting or the direct effect. In general total effect is the standard and also what I am interested in. Therefore the model that utilizes the instrument and only controls for the observed confounder, and not the mediator. Is the model that I would use.