Problem Set 4 Simulation

Olivia Bogiages

2025-12-03

```
set.seed(345)
#population
N<-50000
```

Refer to figure 2 in Causal Inference Is Not Just a Statistics Problem to help with generating the data, or draw your own DAG to illustrate the causal relationships.

Start by generating random data for variables that are not causally affected by any others in your DAG (e.g. the confounder and exogenous variables). Then, generate the remaining variables as linear functions of the variables that causally affect them. Each linear function should have beta coe>icients that represent the true e>ect size, and a random error term.

Confounder: membership of an oppressed group, could bring closer to the violence and also closer to radical groups and recruitment

```
#Simulate a confounder
C_oppr <- rnorm(N, 0,1)

#An independent variable that has an exogenous effect on the treatment variable
#(this is also known as an instrument). Variable: urban/rural

Inst_Urban<-rbinom(n=50000,  size=1, prob=.7)

#Independent Variable, Proximity to violence
I_Prox <- rnorm(N)
#Dependent variable, Radicalization
D_Rad <- rnorm(N)+C_oppr+Inst_Urban
```

Mediator:Displacement, violence causes loss of home, pushes someone towards extreme options

```
#Simulate a mediator
M_Displ = .2*I_Prox + rnorm(N, 0,1)
```

Collider: Extra police or military presence, more present state force where there is violence and also in communities considered radical

```
#A collider
Ex_Pol=.4*I_Prox+.3*D_Rad+rnorm(N, 0,1)

#An independent variable that has an exogenous effect on the outcome variable (i.e.
#it affects Y but does not affect any other variable in your DAG), #outreach by/recruitment by radical

Ex_Recr=rnorm(N,0,1)
```

```
#Data Frame
Proximity_Data<- data.frame(
```

```
  Proximity_to_Violence=c(I_Prox),
  Radicalization=c(D_Rad),
  Membership_of_Oppressed_Group= c(C_oppr),
  Extra_Police= c(Ex_Pol),
  Exogenous_Recruitment=c(Ex_Recr),
  Displacement=c(M_Displ),
  Urban_Rural=c(Inst_Urban)
)
```

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?

It is necessary to include both the confounder and the mediator to recover the direct effect of the indpendent variable on the dependent vairable. This is to ensure that there is no bias and that your coefficients are not being over-attributed with a causal relationship that does not exist in order to isolate the direct relationship of the vairables you are looking at.

```
#Direct effect of proximity on radicalization
direct_effect<-lm(D_Rad~I_Prox+C_oppr+M_Displ, data=Proximity_Data)
summary(direct_effect)
```

```
##
## Call:
## lm(formula = D_Rad ~ I_Prox + C_oppr + M_Displ, data = Proximity_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6077 -0.7329  0.0168  0.7563  4.2229
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.704663   0.004934 142.826   <2e-16 ***
## I_Prox      -0.002409   0.005017  -0.480    0.631
## C_oppr       1.003361   0.004960 202.301   <2e-16 ***
## M_Displ      0.003608   0.004934   0.731    0.465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.103 on 49996 degrees of freedom
## Multiple R-squared:  0.4501, Adjusted R-squared:  0.4501
## F-statistic: 1.364e+04 on 3 and 49996 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? The total effect has a lower coefficient of .0028 compared to the direct effect coefficient of .0035.

```
#Remove the mediator but keep the confounder to look at total effect
total_effect<-lm(D_Rad~I_Prox+C_oppr, data=Proximity_Data)
summary(total_effect)
```

```
##
## Call:
## lm(formula = D_Rad ~ I_Prox + C_oppr, data = Proximity_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -5.6106 -0.7339  0.0161  0.7559  4.2301
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.704650   0.004934 142.825   <2e-16 ***
## I_Prox      -0.001689   0.004920  -0.343    0.731
## C_oppr       1.003354   0.004960 202.301   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.103 on 49997 degrees of freedom
## Multiple R-squared:  0.4501, Adjusted R-squared:  0.4501
## F-statistic: 2.046e+04 on 2 and 49997 DF,  p-value: < 2.2e-16
```

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

When I control for the collider the coefficient and intercept increases, this opens up a backdoor. When I control for the exogenous independent variable, the coefficient is much higher at .01, than the intial model. Controlling for the instrument yields a slightly lower coefficient of .0031 but it impacts the intercept.

```
#control for the collider
collider_control<-lm(D_Rad~I_Prox+Ex_Pol)
summary(collider_control)
```

```
##
## Call:
## lm(formula = D_Rad ~ I_Prox + Ex_Pol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0290 -0.9151  0.0171  0.9195  5.5094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.590967   0.006187   95.51   <2e-16 ***
## I_Prox      -0.213464   0.006437  -33.16   <2e-16 ***
## Ex_Pol       0.551672   0.005541   99.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.359 on 49997 degrees of freedom
## Multiple R-squared:  0.1655, Adjusted R-squared:  0.1654
## F-statistic:  4956 on 2 and 49997 DF,  p-value: < 2.2e-16
```

```
#control for the exogenous independent variable
exog_control<-lm(D_Rad~I_Prox+Ex_Recr)
summary(exog_control)
```

```
##
## Call:
## lm(formula = D_Rad ~ I_Prox + Ex_Recr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1022 -1.0036  0.0127  1.0128  6.2447
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.706223   0.006653 106.146   <2e-16 ***
## I_Prox      0.002402   0.006634   0.362    0.717
## Ex_Recr     0.001611   0.006672   0.241    0.809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.488 on 49997 degrees of freedom
## Multiple R-squared:  3.809e-06,  Adjusted R-squared:  -3.619e-05
## F-statistic: 0.09523 on 2 and 49997 DF,  p-value: 0.9092
```

```
#control for the instrument
inst_control<-lm(D_Rad~I_Prox+Inst_Urban)
summary(inst_control)
```

```
##
## Call:
## lm(formula = D_Rad ~ I_Prox + Inst_Urban)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3984 -0.9641  0.0047  0.9501  6.0981
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.004700   0.011572   0.406    0.685
## I_Prox      0.003181   0.006311   0.504    0.614
## Inst_Urban  1.001045   0.013823  72.420   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 49997 degrees of freedom
## Multiple R-squared:  0.09494,    Adjusted R-squared:  0.09491
## F-statistic:  2622 on 2 and 49997 DF,  p-value: < 2.2e-16
```

Given the reading and simulation results, how should you choose which variables to include in a model? You should first assess what kind of variables you have, and determine based on your hypothesis and theory exactly what role they play in the model. You should always include a confounder because failing to do so would bias your model. You should assess if you have colliders and if so, you should never include the collider. You should include the mediator only if you want the direct affect.