Problem Set 4

Luke Haner

2025-12-01

#Reading

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

A confounder is a variable such as Z that affects X and Y, while a collider is a variable such as Z that is caused both by X and Y. To address confounders in a model, you need to block its "backdoor path" (a non-causal association resulting from X and Y being affected by a confounding variable). To do this you can condition on it trough stratification, restriction, or adjustment. It is best to not adjust for colliders as if you do so, it would create a spurious association leading to a biased estimate of the effect of X on Y.

#2. How can conditioning on a collider create bias?

Conditioning on a collider creates bias because it introduces a non causal path between its causes (independent variable X and outcome Y), making them appear related even when they are not. By adjusting z which is caused both by X and Y, you are creating a spurious association that didn't exist before.

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

Statistical summaries or correlations are inefficient in telling us when to control a variable because they do not provide insight on the data generating process, which is essential to understanding the causal relationships at play.

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

A "kitchen sink" regression is one that is characterized by the selection of variables based on p-values or model-based information. This approach is problematic because it ignores underlying causal relationships which may introduce bias in ways such as ignoring a confounder or adjusting a collider.

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

A backdoor path is a non-causal pathway between the independent (X) and dependent variable (Y) through a confounding variable (Z) which creates a spurious connection, even if X does not actually cause Y. Multiple regression helps block these backdoor paths by conditioning on the confounder (Z). This holds Z constant in the model which allows us to get unbiased causal estimates of the effect of X on Y.

#Simulation

```r
#simulation on legislative professionalization on think tank emergence

#Setting see for predictability
set.seed(123)
n <- 5000

#Variables
C <- rnorm(n) #confounder of economy
Zt <- rnorm(n) # instrument affecting only X (legislator term limits)
```

```r
Zy <- rnorm(n) # variable only affecting y (outcome, number of think tanks that emerged per year over a

#Treatment: Legislative Professionalization
T <- 0.5*C + 0.7*Zt + rnorm(n) # treatment variable is a function of the confounder, instrument, and ra

#Mediator: Policy Complexity
M <- 0.4*T + 0.3*C + rnorm(n) #function of the treatment and confounder plus random noise

#Outcome: Think Tank Emergence
Y <- 1.0*T + 0.8*M + 0.5*C + 0.2*Zy + rnorm(n) #outcome Y is a function of the treatment, mediator, con

#Collider: Bill Passage Rate
K <- 0.6*T + 0.4*Y + rnorm(n) #Collider K is bill passage rate which is affected by the treatment and o

#loading simulated data into data frame
sim_data <- data.frame(T,Y,C,M,K,Zy,Zt)
```

#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?

```r
direct_model <- lm(Y ~ T + C + M, data = sim_data) #multiple regression that controls fro the confounde

summary(direct_model)
```

```
##
## Call:
## lm(formula = Y ~ T + C + M, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6080 -0.6841 -0.0288  0.6943  3.5131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003124   0.014326  -0.218    0.827
## T            0.988992   0.013014  75.993   <2e-16 ***
## C            0.504106   0.016138  31.238   <2e-16 ***
## M            0.806656   0.014194  56.831   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 4996 degrees of freedom
## Multiple R-squared:  0.8361, Adjusted R-squared:  0.836
## F-statistic:  8496 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```r
#The variables that are necessary to recover the direct effect of the treatment on the outcome variable
```

#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?

```r
total_effect_model <- lm(Y ~ T + C, data = sim_data) #multiple regression only controlling for confound

summary(total_effect_model)
```

```
##
## Call:
## lm(formula = Y ~ T + C, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5778 -0.8983  0.0097  0.8853  4.3341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01073    0.01838  -0.584    0.559
## T            1.31659    0.01497  87.947   <2e-16 ***
## C            0.76092    0.01988  38.282   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.299 on 4997 degrees of freedom
## Multiple R-squared:  0.7302, Adjusted R-squared:   0.73
## F-statistic:  6761 on 2 and 4997 DF,  p-value: < 2.2e-16
```

```r
#My model changes by removing the mediator from the model while still regressing Y on T while controlli
```

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

```r
#Controlling for collider (K)

model_collider <- lm(Y ~ T + C + K, data = sim_data)
summary(model_collider)
```

```
##
## Call:
## lm(formula = Y ~ T + C + K, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0638 -0.7807  0.0033  0.7855  4.1092
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01147    0.01621  -0.707    0.479
## T            0.70988    0.02079  34.139   <2e-16 ***
## C            0.60211    0.01803  33.396   <2e-16 ***
## K            0.54188    0.01435  37.770   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.146 on 4996 degrees of freedom
```

```
## Multiple R-squared:  0.7901, Adjusted R-squared:   0.79
## F-statistic:  6268 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```
#When controlling for the collider, the estimated effect of T drops from 1.32 (from the total effect mo

#controlling for the exogenous independent variable (Zy)

model_Zy <- lm(Y ~ T + C + Zy, data = sim_data)
summary(model_Zy)
```

```
##
## Call:
## lm(formula = Y ~ T + C + Zy, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1798 -0.8708  0.0105  0.8759  4.2922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01242    0.01812  -0.686    0.493
## T            1.31919    0.01476  89.357   <2e-16 ***
## C            0.75860    0.01960  38.703   <2e-16 ***
## Zy           0.21673    0.01811  11.970   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.281 on 4996 degrees of freedom
## Multiple R-squared:  0.7377, Adjusted R-squared:  0.7375
## F-statistic:  4683 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```
#When controlling for the exogenous independent variable ZY,the estimated effect of T on Y remains roug

#controlling for the instrument (Zt)

model_Zt <- lm(Y ~ T + C + Zt, data = sim_data)
summary(model_Zt)
```

```
##
## Call:
## lm(formula = Y ~ T + C + Zt, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5921 -0.8904  0.0132  0.8839  4.3241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01095    0.01838  -0.596    0.551
## T            1.30581    0.01834  71.201   <2e-16 ***
## C            0.76633    0.02058  37.242   <2e-16 ***
## Zt           0.02284    0.02245   1.017    0.309
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.299 on 4996 degrees of freedom
## Multiple R-squared:  0.7302, Adjusted R-squared:  0.7301
## F-statistic:  4507 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```
#When controlling for the instrument (Zt), the estimated effect of T on Y also remains roughly equal to
```

#4.Given the reading and simulation results, how should you choose which variables to include in a model?

You should choose variables to include in your model based on your theoretical mapping (primarily shown through your DAG). Confounders must be included as they can create bias in your estimates from the backdoor path which creates spurious correlation between X and Y. It is also important to decide how to handle mediators. If you are looking for direct causal effects they should be included, but if you are measuring total effects they should be excluded from your model. Additionally, the simulation shows that controlling for factors such as exogenous independent variables and instruments may be unnecessary since they are not part of any backdoor path. As such, including them does not bias the estimate, but also provides no added benefit. Overall, selecting variables should follow the causal logic you have constructed for a hypothesis rather than only statistics-based criteria.