Problem Set 4

Ellie Choe

2025-12-12

Part 1: Reading

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

A collider is influenced by both the exposure (X) and the outcome (Y), whereas a confounder influences both X and Y. We should not control for a collider because doing so induces bias in the estimate of the effect of X on Y. Conversely, we must control for a confounder; failing to do so will yield a biased result.

2. How can conditioning on a collider create bias?

Conditioning on a collider creates a spurious association between two variables that are otherwise independent. It opens a non-causal path between the exposure and the outcome, forcing a relationship where none exists or distorting the true relationship.

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

Statistical summaries and visualizations can be identical for datasets generated by completely different causal mechanisms. For instance, the correlation between X and Z can be the same regardless of whether Z is a confounder or a collider. Statistics alone cannot express the directionality of causal relationships, which is required to identify potential bias.

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

Kitchen sink regression refers to a multivariable regression procedure where all available variables—regardless of their association with the outcome—are entered into a model, followed by an automated or manual variable selection strategy based on p-values or model-based information criteria. This approach is problematic because it ignores the directionality of relationships, produces effect estimates with no meaningful causal interpretation, increases the alpha error rate due to multiple testing, leads to overfitting and model instability, and disregards common sense or domain expertise.

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

A backdoor path creates a non-causal association between the exposure and the outcome, even if there is no direct association between them. Multiple regression helps block these paths by conditioning on (or adjusting for) the variables that lie on the path, effectively holding them constant to isolate the true causal effect.

Part 2: Simulation

```
set.seed(111)
n <- 2000

# Generate random data for variables that are not causally a>ected by any others
C <- rnorm(n, 0, 1) # confounder: war
I <- rbinom(n, 1, 0.5) # instrument: EU refugee intake policy
U <- rnorm(n, 0, 1) # exogenous variable: local economic hardship

# Treatment X: refugee numbers
alpha0 <- 0
alpha_C_on_X <- 0.6 # war -> refugee numbers
alpha_I_on_X <- 0.9 # EU policy _> refugee numbers
errorterm_X <- rnorm(n, 0, 1)
X <- alpha0 + alpha_C_on_X * C + alpha_I_on_X * I + errorterm_X

# Mediator M: perceived security threat
beta_M_intercept <- 0
beta_X_on_M <- 0.8
beta_C_on_M <- 0.5
errorterm_M <- rnorm(n, 0, 1)
M <- beta_M_intercept + beta_X_on_M * X + beta_C_on_M * C + errorterm_M

# Outcome Y: far-right party support
beta0 <- 0
beta_X_on_Y <- 1.0
beta_M_on_Y <- 0.6
beta_C_on_Y <- 0.7
beta_U_on_Y <- 0.5
errorterm_Y <- rnorm(n, 0, 1)
Y <- beta0 + beta_X_on_Y * X + beta_M_on_Y * M + beta_C_on_Y * C + beta_U_on_Y * U + errorterm_Y

# Collider S: immigration policy support
gamma_X_on_S <- 0.7
gamma_Y_on_S <- 0.5
errorterm_S <- rnorm(n, 0, 1)
S <- gamma_X_on_S * X + gamma_Y_on_S * Y + errorterm_S

# Combine into a data frame
df <- data.frame(Y, X, M, C, I, U, S)
head(df)
```

```
##           Y          X          M          C I          U          S
```

```
## 1 -0.2696956  0.8984925 -0.27553004  0.2352207 1 -0.6757416  2.3764438
## 2  1.8673640  1.3542384  0.32125517 -0.3307359 1  0.7843683  3.4945056
## 3  2.0019271  1.1698156  0.04909233 -0.3116238 1  0.5942222  2.2767346
## 4 -6.9195425 -1.4701529 -2.58791947 -2.3023457 0 -1.0862305 -3.2092972
## 5  1.6455971  1.6221275  1.29437598 -0.1708760 0  0.6288367  2.4928274
## 6  0.6105494  0.1949980 -0.92721574  0.1402782 1 -0.3093528  0.1369854
```

```r
# 1. estimate X's direct effect
mod_direct <- lm(Y ~ X + M + C + U, data = df)
summary(mod_direct)
```

```
##
## Call:
## lm(formula = Y ~ X + M + C + U, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4185 -0.7021  0.0057  0.7006  3.4706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04325    0.02419   1.788    0.074 .
## X            0.99127    0.02696  36.766   <2e-16 ***
## M            0.60308    0.02215  27.231   <2e-16 ***
## C            0.64821    0.02791  23.222   <2e-16 ***
## U            0.53832    0.02233  24.111   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9951 on 1995 degrees of freedom
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.8708
## F-statistic:  3368 on 4 and 1995 DF,  p-value: < 2.2e-16
```

```r
# 2. estimate X's total effect
mod_total <- lm(Y ~ X + C + U, data = df)
summary(mod_total)
```

```
##
## Call:
## lm(formula = Y ~ X + C + U, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4407 -0.8145  0.0140  0.7949  4.1109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02651    0.02831   0.936    0.349
## X            1.47703    0.02367  62.397   <2e-16 ***
## C            0.94720    0.03005  31.522   <2e-16 ***
## U            0.51610    0.02612  19.755   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.165 on 1996 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8228
## F-statistic:  3095 on 3 and 1996 DF,  p-value: < 2.2e-16
```

```
# 3. Control for the collider, the exogenous independent variable, or the instrument.
mod_collider <- lm(Y ~ X + C + U + S, data = df)
summary(mod_collider)
```

```
##
## Call:
## lm(formula = Y ~ X + C + U + S, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9599 -0.6886  0.0131  0.6799  3.1963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02724    0.02484   1.097    0.273
## X            0.73260    0.03682  19.895   <2e-16 ***
## C            0.71486    0.02801  25.518   <2e-16 ***
## U            0.38213    0.02356  16.219   <2e-16 ***
## S            0.50875    0.02078  24.479   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 1995 degrees of freedom
## Multiple R-squared:  0.8639, Adjusted R-squared:  0.8637
## F-statistic:  3167 on 4 and 1995 DF,  p-value: < 2.2e-16
```

```
mod_noU <- lm(Y ~ X + C, data = df)
summary(mod_noU)
```

```
##
## Call:
## lm(formula = Y ~ X + C, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6910 -0.8507  0.0106  0.8587  4.2929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02076    0.03095   0.671    0.503
## X            1.49964    0.02585  58.023   <2e-16 ***
## C            0.93748    0.03284  28.544   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.274 on 1997 degrees of freedom
## Multiple R-squared:  0.7885, Adjusted R-squared:  0.7883
## F-statistic:  3722 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```r
mod_instrument <- lm(Y ~ X + C + U + I, data = df)
summary(mod_instrument)
```

```
##
## Call:
## lm(formula = Y ~ X + C + U + I, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5094 -0.8164  0.0138  0.7977  4.1458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004656   0.036828  -0.126    0.899
## X            1.462896   0.025965  56.341   <2e-16 ***
## C            0.955866   0.030749  31.086   <2e-16 ***
## U            0.516028   0.026120  19.756   <2e-16 ***
## I            0.075639   0.057169   1.323    0.186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.165 on 1995 degrees of freedom
## Multiple R-squared:  0.8232, Adjusted R-squared:  0.8229
## F-statistic:  2323 on 4 and 1995 DF,  p-value: < 2.2e-16
```

Controlling for the exogenous variable (U) did not meaningfully change the coefficient for X (1.477 vs. 1.499), confirming that omitting U does not introduce bias. However, including U reduced the standard error of the estimate (from 0.026 to 0.024). This indicates that while exogenous variables are not necessary for identification, they improve model efficiency by reducing unexplained variance.

4. To estimate the total effect of refugee numbers (X) on far-right party support (Y), we include confounders (C) and exogenous variables affecting Y (U) but exclude the mediator (M) and the collider (S). The regression results show that X has a coefficient of 1.477, indicating that it increases Y both directly and indirectly through M. Controlling for confounders and exogenous predictors improves estimate precision without blocking causal pathways. Including the collider reduces the X coefficient, introducing bias, while including the instrument does not meaningfully change it.