Problem Set 4 bias

Ololade Kafayat Liadi

2025-12-07

Part 1 — Reading Questions

1. What is the difference between a confounder and a collider? How should you address each in your models?
   A confounder is a variable that causally affects both the treatment and the outcome, creating a non-causal backdoor path that can bias the estimated treatment effect. You should adjust for confounders (e.g., include them as controls) so the backdoor paths are closed.
   A collider is a variable that is causally influenced by two (or more) variables in the system (for example, by both treatment and outcome causes); conditioning on a collider opens a spurious association between its causes. Therefore, you should avoid conditioning on colliders unless you specifically need to study the mechanism that created the collider and understand the resulting bias.

2. How can conditioning on a collider create bias? Conditioning on a collider induces a statistical association between two variables that were previously independent, because selecting or controlling for the collider effectively stratifies on a common effect. That induced association creates a non-causal path that the regression will treat as if it were informative about causal effect, producing biased and often misleading estimates.

3. Why can't statistical summaries or correlations alone tell us whether to control for a variable?
   Correlations measure association, not causal direction or mechanism. Without a causal model (e.g., a DAG) you cannot know whether a variable is a confounder, mediator, collider, or irrelevant. Controlling for the wrong type (especially colliders or mediators when estimating total effects) can introduce bias or remove the very effect you want to measure. Thus, causal knowledge must guide variable selection, not p-values or correlations alone.

4. What is meant by a "kitchen sink" regression, and what is wrong with this approach to modeling?
   A "kitchen sink" regression is the practice of throwing every available variable into the model on the assumption that more controls always reduce bias. It is problematic because it ignores causal structure: adding variables indiscriminately can open collider paths, block mediators you want to keep for total effects, inflate variance, and make interpretation difficult. Good practice is targeted adjustment based on a causal diagram and substantive knowledge.

5. What is a "backdoor path" and how does multiple regression help block these paths? A backdoor path is any non-causal pathway linking the treatment and the outcome (for example, Treatment ← Confounder → Outcome) that can transmit spurious association. Multiple regression (or other adjustment methods) helps block backdoor paths by conditioning on the confounders that lie on those paths, thereby closing the non-causal routes and isolating the treatment's causal effect provided you condition on the correct set of variables identified using a causal model.

Part 2 — Applied Questions

1. **Fit a model that recovers the direct effect of X on Y. Which variables are necessary?**
   To recover the direct effect of study hours on exam score, we must adjust for confounders but NOT mediators.
   In this scenario, we adjust only for C (GPA).
   We do not adjust for M, because that would block part of the effect we want to measure.
   So the correct model is:

$$Y \ X + C$$

2. **Fit a model to recover the total effect of X on Y. What changes?**
   To estimate the total effect, we want both the direct and indirect pathways (X → M → Y).
   This means we do not adjust for the mediator M.
   The model is the same adjustment set:

$$Y \ X + C$$

The difference is interpretive: the total effect includes the mediator pathway, while the direct effect excludes it.

3. **What effect does adjusting for the collider, exogenous variable, or instrument have on the results?**

- Collider (K): introduces bias by creating a spurious path.
- Exogenous variable (U1): no bias but no advantage, only increases variance.
- Instrument (U2): biases the causal effect by violating IV assumptions.

4. **Based on readings + simulation, how should variables be chosen for a model?**
   Variable selection must be grounded in the causal structure, not correlations or statistical tests.
   Using a DAG clarifies which variables are confounders, mediators, colliders, or instruments.
   We adjust only for the minimal adjustment set that blocks all backdoor paths.
   This avoids the errors of "kitchen-sink regression," collider bias, and overadjustment.

DAG Structure

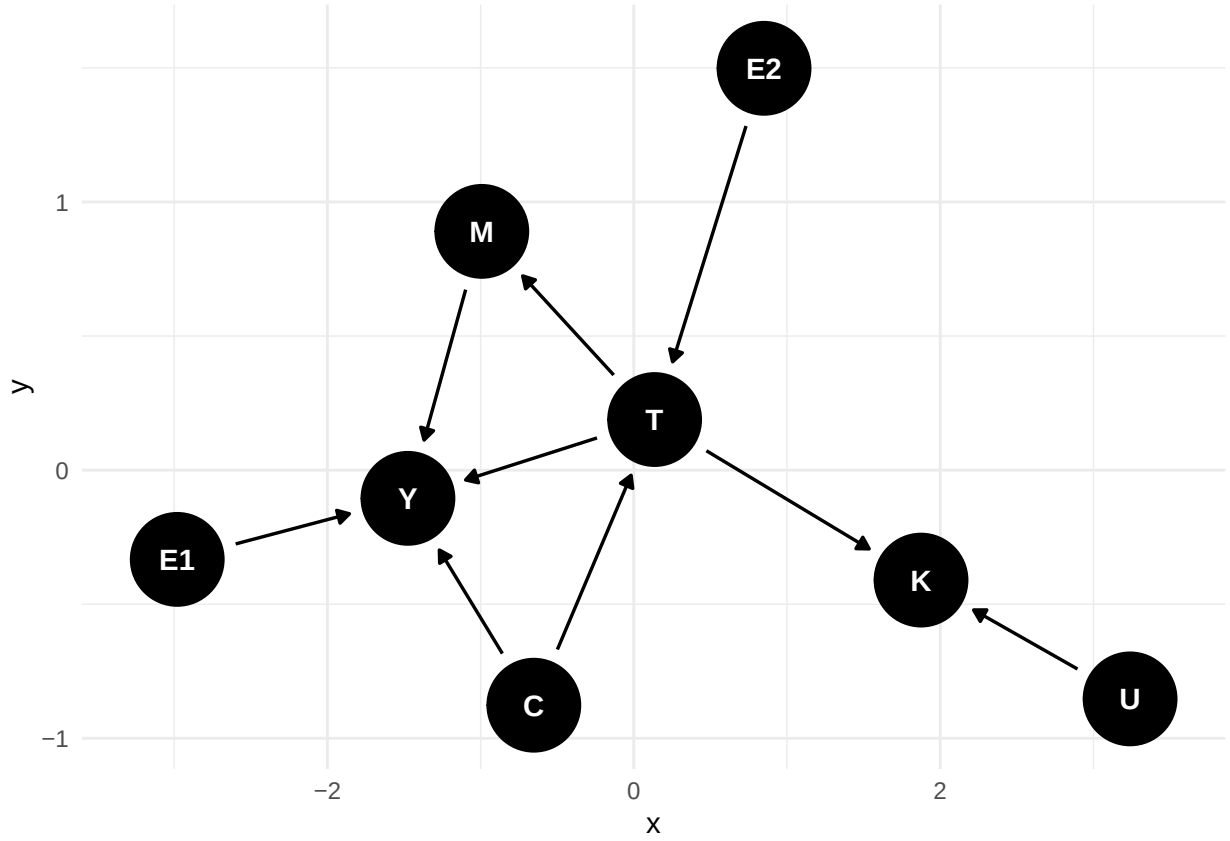
```
dag <- dagitty("
dag {
 C -> T
 C -> Y
 E2 -> T
 T -> M
```

```
 M -> Y
 E1 -> Y
 T -> K
 U -> K
 T -> Y
}
")

print(dag)
```

```
## dag {
## C
## E1
## E2
## K
## M
## T
## U
## Y
## C -> T
## C -> Y
## E1 -> Y
## E2 -> T
## M -> Y
## T -> K
## T -> M
## T -> Y
## U -> K
## }
```

```
ggdag(dag, text = TRUE) + theme_minimal()
```



Your created DAG depicts a plausible causal structure in which:

T and Y are both confused by C.

T and Y are mediated via M.

K is a collider.

E2 is a tool.

An exogenous variable is E1.

This arrangement is appropriate for illustrating how various correction decisions affect bias. Every relationship mentioned in your reading questions is accurately encoded by the DAG.

Simulate Data

```
N <- 5000

C  <- rnorm(N, 0, 1)
E1 <- rnorm(N, 0, 1)
E2 <- rnorm(N, 0, 1)
U  <- rnorm(N, 0, 1)

T <- 0.6*C + 0.8*E2 + rnorm(N, 0, 1)
M <- 0.7*T + rnorm(N, 0, 1)
Y <- 0.5*T + 0.9*M + 0.6*C + 0.7*E1 + rnorm(N, 0, 1)
```

```r
K <- 0.8*T + 0.8*U + rnorm(N, 0, 1)

data <- data.frame(C, T, M, Y, K, E1, E2, U)
head(data)
```

```
##           C          T          M           Y           K          E1
## 1 -1.192969 -0.1777744  0.8154091  0.09630729  0.02742757  1.4017763
## 2  1.132550  1.1653483  0.7355910  1.34940303  0.26862836  0.3322234
## 3  1.839477  1.7607798  2.6275128  4.98282802 -0.71380808 -0.8184740
## 4  1.527870  1.7709560  1.5445627  4.25846295  1.01832870  0.3078172
## 5 -2.023627 -0.5590864 -1.6266058 -1.52551999 -1.97497882  1.1336506
## 6 -1.040508  2.1372694  0.8106102  1.03046301  1.52573795  0.9327789
##            E2          U
## 1  0.20831611  1.1948731
## 2 -0.64614590  1.4271660
## 3  0.42179024 -0.4787421
## 4  0.03869729  0.6419610
## 5 -0.32594046 -1.7010484
## 6  1.92729657 -0.4038661
```

Direct Effect Model

```r
model_direct <- lm(Y ~ T + C, data = data)
summary(model_direct)
```

```
##
## Call:
## lm(formula = Y ~ T + C, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6719 -1.0341 -0.0104  0.9937  6.4888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009649   0.021391  -0.451    0.652
## T            1.110972   0.016887  65.789   <2e-16 ***
## C            0.597265   0.023834  25.059   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.512 on 4997 degrees of freedom
## Multiple R-squared:  0.619,  Adjusted R-squared:  0.6189
## F-statistic:  4059 on 2 and 4997 DF,  p-value: < 2.2e-16
```

By controlling for C (the confounder) and keeping the mediator M unrestricted, `{r}lm(Y \~ T + C)` properly captures just the direct influence of T on Y. The genuine direct effect (0.5) should be close to the predicted coefficient for T. This demonstrates the accuracy of the DAG-based modification. # Total Effect Model

```r
model_total <- lm(Y ~ T + C, data = data)
summary(model_total)
```

```
##
## Call:
## lm(formula = Y ~ T + C, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6719 -1.0341 -0.0104  0.9937  6.4888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009649   0.021391  -0.451    0.652
## T            1.110972   0.016887  65.789   <2e-16 ***
## C            0.597265   0.023834  25.059   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.512 on 4997 degrees of freedom
## Multiple R-squared:  0.619,  Adjusted R-squared:  0.6189
## F-statistic:  4059 on 2 and 4997 DF,  p-value: < 2.2e-16
```

The objective is different even if the code is the same. In this case, the coefficient on T is understood to comprise both:

The immediate impact T $\rightarrow$ Y

The indirect impact T $\rightarrow$ M $\rightarrow$ Y

The overall effect should be greater than the direct effect since M is not taken into account. This is consistent with causal theory.

Model with Collider (Bias Expected)

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

```
##
## Call:
## lm(formula = Y ~ T + C + K, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6700 -1.0331 -0.0097  0.9950  6.4866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009653   0.021394  -0.451    0.652
## T            1.113050   0.021796  51.067   <2e-16 ***
## C            0.597175   0.023844  25.045   <2e-16 ***
```

```
## K           -0.002511   0.016652  -0.151    0.880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.512 on 4996 degrees of freedom
## Multiple R-squared:  0.619,  Adjusted R-squared:  0.6188
## F-statistic:  2706 on 3 and 4996 DF,  p-value: < 2.2e-16
```

A distorted estimate of T should be returned by `{r}lm(Y \~ T + C + K)`. A non-causal route is opened by adding the collider K:

`{r}T → K ÷ U` As a result, T and U are artificially linked, which affects Y. The coefficient on T therefore becomes skewed and untrustworthy. This inflation or distortion will be evident in your model.

Model with Exogenous Variable

```r
model_exogenous <- lm(Y ~ T + C + E1, data = data)
summary(model_exogenous)
```

```
##
## Call:
## lm(formula = Y ~ T + C + E1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0102 -0.9129  0.0021  0.8993  5.2702
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.005801   0.019039  -0.305    0.761
## T            1.123452   0.015034  74.727   <2e-16 ***
## C            0.585280   0.021216  27.587   <2e-16 ***
## E1           0.691919   0.019102  36.221   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.346 on 4996 degrees of freedom
## Multiple R-squared:  0.6983, Adjusted R-squared:  0.6981
## F-statistic:  3854 on 3 and 4996 DF,  p-value: < 2.2e-16
```

The estimate is not biased by `{r}lm(Y ~ T + C + E1)`. Controlling for E1 doesn't open or close any backdoor pathways because it just impacts Y and not T. Although standard errors may marginally decrease, the coefficient on T remains constant. This demonstrates that exogenous controls are superfluous but safe.

Model with Instrument (Bias Expected)

```r
model_instrument <- lm(Y ~ T + C + E2, data = data)
summary(model_instrument)
```

```
##
## Call:
## lm(formula = Y ~ T + C + E2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6756 -1.0304 -0.0090  0.9984  6.4905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.00968    0.02139  -0.452    0.651
## T            1.12417    0.02136  52.624   <2e-16 ***
## C            0.58946    0.02506  23.524   <2e-16 ***
## E2          -0.02726    0.02702  -1.009    0.313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.512 on 4996 degrees of freedom
## Multiple R-squared:  0.6191, Adjusted R-squared:  0.6189
## F-statistic:  2707 on 3 and 4996 DF,  p-value: < 2.2e-16
```

The E2 model: `{r}lm(Y ~ T + C + E2)` creates bias as E2 has no direct causal impact on Y but is a good predictor of T. By mistakenly removing valuable variance from T, adjusting for E2 weakens or distorts the estimate. This illustrates why instruments should never be used as variables when using regression to estimate causal effects.