Problem_Set_4

Santiago Vidal Calvo

2025-12-07

```
# Readings:
# - "Building a better model: abandon kitchen sink regression"
# - "Causal Inference Is Not Just a Statistics Problem"
# Question 1
# What is the difference between a confounder and a collider? How should
# > you address each in your models?
#
# Answer 1:
# A confounder is a variable that affects both the treatment (or exposure)
# > and the outcome, and so it opens a "backdoor path" that creates a
# > spurious association between them.
# In contrast, a collider is a variable that is caused by both the
# > treatment and the outcome (or their causes), so the arrows point into
# > the collider rather than out of it.
# We usually want to adjust for confounders, because conditioning on them
# > blocks the backdoor paths and helps us get closer to the causal effect.
# Colliders are the opposite: we generally do not adjust for colliders,
# > because conditioning on them opens up a path that was previously
# > closed and can introduce bias rather than remove it.
# Question 2
# How can conditioning on a collider create bias?
```

```
#
# Answer 2:
# Conditioning on a collider can create bias because it makes the
# > treatment and outcome (or their causes) statistically dependent even
# > if they were independent before.
# When we restrict our sample to a specific value of the collider, or
# > include it as a control in a regression, we "open" a path through the
# > collider that connects the treatment and outcome.
# This can generate a spurious association that looks like a causal
# > effect but is really just an artifact of selection.
# The readings use examples where conditioning on a health-related
# > collider makes two risk factors appear negatively correlated, even
# > though they are not, which illustrates how this bias can be quite
# > misleading.
# Question 3
# Why can't statistical summaries or correlations alone tell us whether
# > to control for a variable?
#
# Answer 3:
# Statistical summaries like correlations, regression coefficients, or
# > balance tables only describe patterns in the data; they do not tell us
# > about the direction of causality.
# A variable could be a confounder, a mediator, a collider, or just an
# > unrelated factor, and it might still have a strong correlation with
# > both the treatment and the outcome.
# The papers emphasize that we need a story about the data-generating
# > process, often expressed as a DAG, to decide whether adjusting for a
```

```
# > variable will help or hurt.
# Two datasets can have identical summary statistics and even identical
# > regression results but require totally different adjustment sets
# > depending on the underlying causal structure.
# Question 4
# What is meant by a "kitchen sink" regression, and what is wrong with
# > this approach to modeling?
#
# Answer 4:
# A "kitchen sink" regression is when we throw every available variable
# > into the model (or use stepwise procedures) without a clear causal
# > plan, and then interpret the coefficients as if they had causal
# > meaning.
# The "Building a better model" paper argues that this approach ignores
# > directionality, encourages overfitting, inflates type I error, and
# > makes coefficients unstable and hard to interpret.
# It also risks adjusting for mediators or colliders, which can distort
# > causal estimates.
# Instead of kitchen sink models, the authors suggest starting with a
# > DAG based on subject-matter knowledge and using it to decide which
# > variables to adjust for.
# Question 5
# What is a "backdoor path" and how does multiple regression help block
# > these paths?
#
# Answer 5:
# A backdoor path is any non-causal path in a DAG that links the
```

```r
# > treatment to the outcome by going "into the back" of the treatment
# > through a common cause.
# For example, if a confounder affects both the treatment and the
# > outcome, there is a backdoor path through that confounder.
# Multiple regression helps by conditioning on confounders, which
# > effectively blocks those backdoor paths so that the remaining
# > association between the treatment and outcome is closer to the causal
# > effect.
# The key is to adjust only for variables that close backdoor paths while
# > avoiding colliders and post-treatment variables that could introduce
# > new bias.
```

```r
# Question: Think of some social causal relationship and simulate data
# > with:
# > - A treatment and outcome
# > - A confounder
# > - A mediator
# > - A collider
# > - An exogenous predictor of Y only
# > - An exogenous predictor of the treatment only (an instrument)
set.seed(123)
n <- 5000
# Confounder C: affects both treatment T and outcome Y
C <- rnorm(n, mean = 0, sd = 1)
# Exogenous predictor for treatment only: Zx (instrument)
Zx <- rnorm(n, mean = 0, sd = 1)
# Exogenous predictor for outcome only: Zy
Zy <- rnorm(n, mean = 0, sd = 1)
```

```r
# Treatment T: affected by confounder C and instrument Zx

Treat <- 0.5 * C + 0.8 * Zx + rnorm(n, 0, 1)

# Mediator M: affected by treatment and confounder

M <- 0.7 * Treat + 0.3 * C + rnorm(n, 0, 1)

# Outcome Y: affected by treatment (direct), mediator, confounder, Zy

# True direct effect of treatment is 0.6; indirect path via M is 0.5*0.7

Y <- 0.6 * Treat + 0.5 * M + 0.5 * C + 0.7 * Zy + rnorm(n, 0, 1)

# Collider K: affected by both treatment and outcome

K <- 0.6 * Treat + 0.6 * Y + rnorm(n, 0, 1)

sim_data <- data.frame(
Y = Y,
Treat = Treat,
C = C,
M = M,
K = K,
Zx = Zx,
Zy = Zy
)

head(sim_data)
```

```
##           Y      Treat           C           M          K          Zx          Zy
## 1 -1.846112 -2.0294258 -0.56047565 -2.42503749 -2.5189290 -0.4941739  2.3707252
## 2 -1.314553  0.2076088 -0.23017749 -0.14430010 -0.4060195  1.1275935 -0.1668120
## 3  0.241090 -0.9992497  1.55870831 -2.33537705 -0.9932084 -1.1469495  0.9269614
## 4  1.415373  2.1927474  0.07050839 -0.11173183  0.9858095  1.4810186 -0.5681517
## 5  1.246210  1.4167426  0.12928774 -0.06745671  2.4984187  0.9161912  0.2250901
## 6  2.976494  2.5110830  1.71506499  0.60665638  3.2762965  0.3351310  1.1319859
```

```r
# This head() output shows the first few rows of the simulated data and

# > lets me confirm that all variables were created as expected.
```

```r
# Question 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?

#

# To get the direct effect of Treat on Y, we need to:
```

```r
# - Adjust for the confounder C to block the backdoor path C -> Treat and

# > C -> Y.

# - Adjust for the mediator M so that we block the indirect path

# > Treat -> M -> Y and isolate the direct path Treat -> Y.

# We do not want to adjust for the collider K.

mod_direct <- lm(Y ~ Treat + C + M + Zy, data = sim_data)
summary(mod_direct)
```

```
##
## Call:
## lm(formula = Y ~ Treat + C + M + Zy, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4991 -0.6681 -0.0186  0.6875  3.5071
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004701   0.014034  -0.335    0.738
## Treat        0.593404   0.014658  40.484   <2e-16 ***
## C            0.502130   0.015717  31.947   <2e-16 ***
## M            0.503421   0.013906  36.202   <2e-16 ***
## Zy           0.703879   0.014021  50.202   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.992 on 4995 degrees of freedom
## Multiple R-squared:  0.7805, Adjusted R-squared:  0.7803
## F-statistic:  4441 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```r
# In this model, the coefficient on Treat should be close to the true

# > direct effect of 0.6, because we have adjusted for both the

# > confounder and the mediator.

# Zy is included to improve precision, since it affects Y but does not

# > confound Treat and Y.
```

```r
# Question 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
# For the total effect, we want both the direct and mediated paths:

# > Treat -> Y and Treat -> M -> Y.

# So we should adjust for the confounder C but NOT for the mediator M.

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

```
##
## Call:
## lm(formula = Y ~ Treat + C + Zy, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8280 -0.7626  0.0034  0.7754  3.6711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009526   0.015765  -0.604    0.546
## Treat        0.948432   0.012239  77.495   <2e-16 ***
## C            0.662582   0.016941  39.111   <2e-16 ***
## Zy           0.711940   0.015750  45.204   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.114 on 4996 degrees of freedom
## Multiple R-squared:  0.7229, Adjusted R-squared:  0.7228
## F-statistic:  4345 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```r
# Now the coefficient on Treat should be close to the total effect,

# > which is the sum of the direct effect (0.6) and the mediated effect

# > (0.7 * 0.5 = 0.35), so the total is about 0.95.

# By leaving M out of the model, we allow that indirect path to stay

# > open, so the estimated effect of Treat includes both components.
```

```r
# Question 3

# How do your results change when you control for:

# > - the collider K,

# > - the exogenous predictor of Y (Zy),

# > - the instrument for Treat (Zx),

# > individually, not all at once?

#

# We start from the "total effect" model Y ~ Treat + C and then modify it.
```

```r
# Baseline total-effect model (no extra controls)

mod_base <- lm(Y ~ Treat + C, data = sim_data)
summary(mod_base)
```

```
##
## Call:
## lm(formula = Y ~ Treat + C, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1347 -0.8753 -0.0112  0.8962  4.2281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003934   0.018711   -0.21    0.833
## Treat        0.940797   0.014525   64.77   <2e-16 ***
## C            0.669749   0.020106   33.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.323 on 4997 degrees of freedom
## Multiple R-squared:  0.6096, Adjusted R-squared:  0.6094
## F-statistic:  3901 on 2 and 4997 DF,  p-value: < 2.2e-16
```

```r
# Add collider K

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

```
##
## Call:
## lm(formula = Y ~ Treat + C + K, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5540 -0.6962 -0.0126  0.6694  3.6596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.006077   0.014554  -0.418    0.676
## Treat        0.189443   0.017338  10.926   <2e-16 ***
## C            0.416542   0.016255  25.625   <2e-16 ***
## K            0.648819   0.011357  57.127   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.029 on 4996 degrees of freedom
## Multiple R-squared:  0.7639, Adjusted R-squared:  0.7637
## F-statistic:  5387 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```r
# Add exogenous predictor of Y (Zy)

mod_exogY <- lm(Y ~ Treat + C + Zy, data = sim_data)
```

```r
summary(mod_exogY)
```

```
##
## Call:
## lm(formula = Y ~ Treat + C + Zy, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8280 -0.7626  0.0034  0.7754  3.6711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009526   0.015765  -0.604    0.546
## Treat        0.948432   0.012239  77.495   <2e-16 ***
## C            0.662582   0.016941  39.111   <2e-16 ***
## Zy           0.711940   0.015750  45.204   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.114 on 4996 degrees of freedom
## Multiple R-squared:  0.7229, Adjusted R-squared:  0.7228
## F-statistic:  4345 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```r
# Add instrument Zx

mod_instrument <- lm(Y ~ Treat + C + Zx, data = sim_data)
summary(mod_instrument)
```

```
##
## Call:
## lm(formula = Y ~ Treat + C + Zx, data = sim_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1584 -0.8813 -0.0103  0.8997  4.2265
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004347   0.018710  -0.232   0.8163
## Treat        0.921290   0.018668  49.352   <2e-16 ***
## C            0.679528   0.020945  32.443   <2e-16 ***
## Zx           0.039876   0.023979   1.663   0.0964 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.322 on 4996 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.6096
## F-statistic:  2603 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```r
# Interpretation:

# - In mod_base, the Treat coefficient is an estimate of the total effect

# > (should be around 0.95) when we adjust only for the confounder C.
```

```
# - In mod_collider, including K (which is caused by Treat and Y) tends
# > to distort the estimated effect of Treat, because conditioning on a
# > collider opens a spurious path between Treat and Y.
# We usually see the Treat coefficient move away from the true total
# > effect, illustrating collider bias.
# - In mod_exogY, including Zy, which affects Y but not Treat, should not
# > bias the coefficient on Treat. It can reduce residual variance and
# > slightly shrink the standard error, but the point estimate should
# > stay close to the total effect.
# - In mod_instrument, including Zx, which affects Treat but not Y
# > directly, does not help estimate the causal effect and can add some
# > noise. In expectation it should not bias the Treat coefficient, but
# > in a finite sample it can wiggle the estimate around a bit.
# These comparisons illustrate that adding a collider is harmful, adding
# > a pure outcome predictor is mostly about precision, and adding an
# > instrument as a covariate is not especially useful in a simple OLS
# > setting.
```

```
# Question 4
# Given the reading and simulation results, how should you choose which
# > variables to include in a model?
#
# Answer 4:
# This simulation reinforces the message from the readings: we should not
# > decide what to control for based only on statistical significance or
# > how many variables we can throw into a "kitchen sink" regression.
# Instead, we should start with a causal diagram that reflects our
# > substantive understanding of the problem and then:
```

```
# - Adjust for true confounders that open backdoor paths between the
# > treatment and the outcome.
# - Avoid conditioning on colliders or post-treatment variables, because
# > they can introduce new bias rather than remove it.
# - Treat pure outcome predictors and instruments carefully: they may
# > help with precision or identification in more advanced settings, but
# > they do not automatically belong in every model.
# The big lesson is that good modeling is about understanding the data
# > generating process first and then using regression as a tool, not the
# > other way around.
```