Problem Set 4

Due Date: February 6, 2026
Submission: https://canvas.northwestern.edu/courses/245562/assignments/1676748

Problem 1

1a. *Define omitted variable bias in your own words.*

Omitted variable bias is the expected difference in estimates for a given X variable's associated coefficient between a larger model that includes a set of additional control variables and a smaller model that does not include them. In order for there to be a nonzero difference, it is necessary both that the additional control variables have nonzero slopes in the larger model and that they have nonzero covariances with the X variable of interest.

1b. Consider the linear models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$y = \beta_0^* + \beta_1^* x_1 + u^*$$

Derive the formula for β_1^* in terms of β_1, β_2, and the relationship between x_1 and x_2. Show all steps.

We know that:

$$\beta_1^* = \frac{\text{cov}(y, x_1)}{\text{var}(x_1)}$$

Let's plug in the larger model's expression for y into this statement.

$$\beta_1^* = \frac{\text{cov}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, x_1)}{\text{var}(x_1)}$$

Covariance is linear, so we can rewrite this as:

$$\beta_1^* = \frac{\text{cov}(\beta_0, x_1)}{\text{var}(x_1)} + \frac{\text{cov}(\beta_1 x_1, x_1)}{\text{var}(x_1)} + \frac{\text{cov}(\beta_2 x_2, x_1)}{\text{var}(x_1)} + \frac{\text{cov}(u, x_1)}{\text{var}(x_1)}$$

We know that β_0 is a constant, so the first covariance is zero. We know that the error term is unrelated to the independent variables, so the fourth covariance is zero. The second covariance is $\beta_1 \text{var}(x_1)$, which is also the denominator, so that cancels out and turns into just β_1. So we're left with:

$$\beta_1^* = \beta_1 + \beta_2 \frac{\text{cov}(x_2, x_1)}{\text{var}(x_1)}$$

1c. *Interpret each term in your formula from 1b. Under what conditions does omitted variable bias occur? When is it zero?*

The left-hand side is the slope for variable 1 when variable 2 is omitted. β_1 is the slope for variable 1 when variable 2 is included. So the difference between those two is the omitted variable bias; that is the remaining portion of the formula.

β_2 is the slope for variable 2. $\text{cov}(x_2, x_1)$ is a measure of the linear relationship between variables 1 and 2. var(x_{1}) is a measure of the variance of variable 1. Omitted variable bias occurs when variable 2 has a nonzero slope in the larger model, when the linear relationship between variables 1 and 2 is nonzero, and when the variance of variable 1 is finite.

Problem 2

The lecture slides show a nonlinear CEF for GDP and turnout, with different linear approximations (BLPs) for different ranges of GDP. In your own words, explain:

1. What does it mean for the BLP to be the "best linear approximation" of a nonlinear CEF?

Of all possible linear equations, it is the one that comes closest to fitting the form of the nonlinear CEF; it leaves the smallest overall remaining sum of squared errors.

2. How can the BLP coefficient change sign depending on which range of the data we focus on?

Because the CEF is nonlinear and we are fitting a line to it, emphasizing different parts of the range of the data will draw attention to different portions of the curve and will therefore produce a BLP with a different slope.

3. What are the implications for interpreting regression coefficients when the true CEF is nonlinear?

When the true CEF is nonlinear, regression coefficients must always be interpreted as at best local approximations that provide information about relationships near the data, and should never be extrapolated to areas of sparse data or beyond the range of the observed information.

Problem 3

Return to your voter turnout analysis from Problem Set 3.

3a. Re-examine the difference between your bivariate model (turnout ~ income) and multivariate model (turnout ~ income + religiosity + age). Set up a version of the multivariate model that only adds religiosity and does not use age. Calculate the omitted variable bias using the formula from Problem 1.

```
# Calculate the components needed for OVB formula

library(poliscidata)
```

```
## Registered S3 method overwritten by 'gdata':
##   method         from
##   reorder.factor gplots
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.2.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Clean and prepare the data
states_data <- states %>%
  select(state, vep12_turnout, prcapinc, religiosity, over64) %>%
  filter(!is.na(vep12_turnout), !is.na(prcapinc)) %>%
  mutate(income_thousands = prcapinc / 1000)

turnout_bivariate <- lm(vep12_turnout ~ income_thousands, data = states_data)
```

```r
turnout_multivariate <- lm(vep12_turnout ~ income_thousands + religiosity,
                           data = states_data)

# You'll need:
# 1. beta2 from multivariate model (coefficient for religiosity)
beta2 <- turnout_multivariate$coefficients[3]
beta2
```

```
## religiosity
## -0.02997725
```

```r
# 2. Covariance between income and religiosity
inc_rel_covariance <- cov(states_data$income_thousands,states_data$religiosity)
inc_rel_covariance
```

```
## [1] -150.9603
```

```r
# 3. Variance of income
inc_variance <- var(states_data$income_thousands)
inc_variance
```

```
## [1] 19.79052
```

```r
# Then compute: OVB = beta2 * Cov(x1, x2) / Var(x1)
ovb_computation <- beta2*inc_rel_covariance/inc_variance
ovb_computation
```

```
## religiosity
##   0.2286637
```

3b. Does the OVB formula correctly predict the difference between the bivariate and multivariate income coefficients? Show your calculations.

```r
turnout_bivariate$coefficients[2] - turnout_multivariate$coefficients[2]
```

```
## income_thousands
##        0.2286637
```

It is exactly the same.

3c. Based on the lecture slides' discussion of model specification: *1. Could adding more variables ever increase bias? Under what conditions?*

Adding more variables cannot increase value of the omitted variable bias formula, which measures the bias due to omission. However adding variables can increase bias in ways that formula does not measure in causal inference frameworks. For example, adding a control variable that is causally post-treatment relative to the main independent variable of interest will not increase the omitted variable bias according to this formula, and may well reduce it. However, it can add post-treatment bias to a causal inference.

2. When might it be better to use a bivariate model even if you suspect omitted variables?

If we are trying to carry out causal inference and we suspect that omitted variables belong to suspect categories (post-treatment, collider), or if adding the omitted variables distorts measurement quality or data availability severely.

Problem 4

Simulation Study of OVB

We'll study two scenarios of omitted variable bias through simulation.

Scenario A: Confounding (Both X1 and X2 cause Y)

```
set.seed(789)
n <- 1000
x1 <- rnorm(n)
x2 <- 0.7*x1 + rnorm(n)  # x2 correlated with x1
y <- 2 + 1.5*x1 + 2*x2 + rnorm(n, sd = 0.5)

# Run regressions
model_bivariate <- lm(y ~ x1)
model_multivariate <- lm(y ~ x1 + x2)

summary(model_bivariate)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1525 -1.3460 -0.0749  1.3630  5.9756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.96193    0.06260   31.34   <2e-16 ***
## x1           2.92022    0.06245   46.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.98 on 998 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6863
## F-statistic:  2187 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
summary(model_multivariate)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97057 -0.31425  0.01079  0.31582  1.52002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.99251    0.01541   129.3   <2e-16 ***
## x1           1.51282    0.01908    79.3   <2e-16 ***
## x2           1.98904    0.01598   124.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4871 on 997 degrees of freedom
## Multiple R-squared:  0.981,  Adjusted R-squared:  0.981
## F-statistic: 2.58e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

Questions for Scenario A: *1. What is the true value of β_1?*

1.5, as we can see from the formula that constructs Y.

2. What is the estimated beta$_1$ in the bivariate model? How biased is it?

The bivariate model gives an estimate of about 2.9. This is a substantial positive bias of about 1.4.

3. Use the OVB formula to calculate the expected bias. Does it match the actual bias?

```
model_multivariate$coefficients[3]*cov(x1,x2)/var(x1)
```

```
##       x2
## 1.407403
```

This is almost exactly the same as the actual bias.

Scenario B: Collider Bias (X2 is a common effect)

```
set.seed(789)
n <- 1000

# Correct setup for collider scenario
x1 <- rnorm(n)
y <- 2 + 1.5*x1 + rnorm(n, sd = 0.5)  # y depends only on x1
x2 <- 0.7*x1 - 1.5*y + rnorm(n, sd = 0.5)  # x2 is a collider

# Run regressions
model_correct <- lm(y ~ x1)  # Correct specification
model_collider <- lm(y ~ x1 + x2)  # Including collider

summary(model_correct)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.53024 -0.32887 -0.01085  0.32344  1.51190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.99231    0.01525  130.61   <2e-16 ***
## x1           1.50379    0.01522   98.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4824 on 998 degrees of freedom
## Multiple R-squared:  0.9073, Adjusted R-squared:  0.9072
## F-statistic:  9768 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
summary(model_collider)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.77772 -0.17814 -0.00328  0.17006  0.93515
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.625167   0.029928   20.89   <2e-16 ***
## x1           0.796156   0.017079   46.62   <2e-16 ***
## x2          -0.456355   0.009585  -47.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2667 on 997 degrees of freedom
## Multiple R-squared:  0.9717, Adjusted R-squared:  0.9716
## F-statistic: 1.711e+04 on 2 and 997 DF,  p-value: < 2.2e-16
```

Questions for Scenario B: *1. What is the true value of β_1?*

1.5, again.

2. What happens when we include x2 in the regression? Why?

The bivariate estimate, which is very close to correct, is distorted into producing a new estimate that is far from correct and shows a downward bias to about 0.8. This is because we have included an inappropriate control variable that distorts the results.

3. This demonstrates "bad control" or collider bias. Explain in your own words why including x2 creates bias even though x2 is correlated with both x1 and y.

Even though all the variables are correlated, the directions of causal flow are not what we expect. As a result, the statistical relations mean something different here. Unconditional on x_2, there is no confounding, but conditional on X_2, we induce a new spurious negative correlation between x_1 and y. This is the bias.

4c. Simulation Synthesis: *1. Create a table comparing both scenarios.*

```r
# Simulation comparison
library(modelsummary)
simulations <- list("Bivariate Confounded" = model_bivariate,
                    "Multivariate Unconfounded" = model_multivariate,
                    "Collider Excluded" = model_correct,
                    "Collider Included" = model_collider)
modelsummary(simulations, stars = TRUE, output = "markdown")
```

	Bivariate Confounded	Multivariate Unconfounded	Collider Excluded	Collider Included
(Intercept)	1.962***	1.993***	1.992***	0.625***
	(0.063)	(0.015)	(0.015)	(0.030)
x1	2.920***	1.513***	1.504***	0.796***
	(0.062)	(0.019)	(0.015)	(0.017)
x2		1.989***		-0.456***
		(0.016)		(0.010)
Num.Obs.	1000	1000	1000	1000
R2	0.687	0.981	0.907	0.972
R2 Adj.	0.686	0.981	0.907	0.972
AIC	4207.7	1404.3	1383.7	199.8
BIC	4222.4	1423.9	1398.4	219.4
Log.Lik.	-2100.855	-698.131	-688.863	-95.903
F	2186.796	25804.617	9767.693	17105.635

	Bivariate Confounded	Multivariate Unconfounded	Collider Excluded	Collider Included
RMSE	1.98	0.49	0.48	0.27

- p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

2. Under what circumstances does adding a control variable reduce bias? When might it increase bias?

Adding control variables will generally not hurt us when the control variables in question are not caused by the independent variable of interest. They can be damaging, though, when they are caused by the independent variable and the dependent variable (collider), or when they are caused by the independent variable and cause the dependent variable (post-treatment). They reduce bias most significantly when they cause both the independent variable and the dependent variable (confounder).

3. How can researchers decide which variables to include in a regression model?

We need background knowledge of the research design and/or causal situation that comes from somewhere other than the data and the model. We can't build a successful regression-type model based solely on statistical considerations.

Problem 5

The lecture slides derive OLS using the plug-in principle and matrix algebra.

5a. Plug-in Principle: *1. Define the plug-in principle in your own words.*

We can estimate a feature of a population distribution by using the equivalent feature of an appropriate finite sample's empirical distribution.

5b. Matrix Derivation: The OLS estimator in matrix form is:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Using the turnout data with GDP and Temperature:

```
# Load and clean data
turnout <- read_csv("https://raw.githubusercontent.com/jnseawright/ps405/refs/heads/main/Data/turnout.cs
```

```
## Rows: 50 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): Year, Turnout, Temperature, GDP
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
turnout_clean <- turnout[13:nrow(turnout), ]  # Remove NA rows as in slides

# Create X matrix with intercept, GDP, Temperature
X <- as.matrix(cbind(1, turnout_clean$GDP, turnout_clean$Temperature))
colnames(X) <- c("Intercept", "GDP", "Temperature")

# Create y vector
y <- turnout_clean$Turnout

# Calculate OLS coefficients using matrix algebra
```

```r
betahat <- solve(t(X)%*%X, tol = 1e-19)%*%t(X)%*%y
betahat
```

```
##                      [,1]
## Intercept    5.750768e-01
## GDP         -3.524858e-09
## Temperature  9.408781e-04
```

```r
# Compare with lm() output
summary(lm(Turnout ~ GDP + Temperature, data=turnout_clean))
```

```
##
## Call:
## lm(formula = Turnout ~ GDP + Temperature, data = turnout_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13323 -0.05916 -0.01721  0.02748  0.19650
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.751e-01  5.458e-01   1.054    0.299
## GDP         -3.525e-09  3.006e-09  -1.173    0.249
## Temperature  9.409e-04  1.066e-02   0.088    0.930
##
## Residual standard error: 0.09201 on 35 degrees of freedom
## Multiple R-squared:  0.06614,    Adjusted R-squared:  0.01278
## F-statistic: 1.239 on 2 and 35 DF,  p-value: 0.3019
```

1. Verify that your matrix calculation matches the `lm()` output.

Great, they do!