Problem Set 3

Due Date: January 30, 2026
Submission: https://canvas.northwestern.edu/courses/245562/assignments/1676747

Problem 1

1a. *In a multiple regression model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, interpret β_1 in your own words.*

β_1 represents the relationship between X_1 and Y while subtracting off the linear relationship between X_2 and Y. In other words, it is the change in the mean in Y expected when there is a unit change in X_1 after removing the linear relationship between X_2 and Y.

1b. Consider the slides' example of terrorism incidents predicted by Trump vote share and 2012 margin. If the estimated equation is:

$$\text{Terrorism} = 10 - 0.2 \times \text{TrumpShare} + 0.1 \times \text{D12Margin}$$

Interpret what happens to predicted terrorism incidents when: - TrumpShare increases by 5 percentage points, holding D12Margin constant.

```
Terrorism would on average decrease by 1.
```

- *D12Margin decreases by 3 percentage points, holding TrumpShare constant.*

  Terrorism would on average decrease by 0.3.

1c. The slides show that in multivariate regression, β_1 represents the expected change in Y when X_1 increases by 1 unit, *with all other variables held constant.* Why is the "all other variables held constant" condition crucial for interpreting β_1 as a partial effect?

TrumpShare and D12Margin might be connected between themselves; a change in the Democrats' past margin might be expected to affect Trump's vote share, so these variables may have some statistical and even causal linkages between themselves. If we don't stipulate that they are held constant, adjusting D12Margin might be expected to change Terrorism directly but also indirectly by changing TrumpShare.

Problem 2

The Conditional Expectation Function (CEF) has a key property: it is the best predictor of Y given X in the mean-squared error sense.

Let $m(X)$ be any function of X used to predict Y. The mean-squared error (MSE) is defined as:

$$MSE(m) = E[(Y - m(X))^2]$$

2a. Show that for any predictor $m(X)$, the MSE can be decomposed as:

$$E[(Y - m(X))^2] = E[(Y - E[Y|X])^2] + E[(E[Y|X] - m(X))^2]$$

Hint: Start with $Y - m(X) = (Y - E[Y|X]) + (E[Y|X] - m(X))$, then expand the square.

$E[(Y-m(X))^2] = E[((Y-E[Y|X])+(E[Y|X]-m(X)))^2] = E[((Y-E[Y|X])^2+2(Y-E[Y|X])(E[Y|X]-m(X))+(E[Y|X]-m$

We can get rid of the middle term using the Law of Iterated Expectations. Specifically, we can take the expectation of that middle term conditional on X, and since that is already in an expectation, that is the same thing as the original term.

$$E(2(Y - E[Y|X])(E[Y|X] - m(X))|X)$$

Since $E[Y|X]$ and $m(X)$ are both functions of X, conditioning on X lets us treat both as constant and we can take them out of the inner expectation.

$$E[((Y - E[Y|X])^2 + E(2(Y - E[Y|X]|X))(E[Y|X] - m(X)) + (E[Y|X] - m(X))^2]$$

The first part of the product is:

$$E(2(Y - E[Y|X]|X)) = E(2(Y|X)) - E(E[Y|X]|X))$$

Double-conditioning on X is redundant, as is the double expectation, so this is the same as:

$$E(2(Y - E[Y|X]|X)) = E(2(Y|X)) - E([Y|X]) = 0$$

Since this part of the product collapses to zero, so does the whole product.

Then we get:

$$E[(Y - m(X))^2] = E[((Y - E[Y|X])^2] + (E[Y|X] - m(X))^2$$

2b. *Using the decomposition from 2a, explain why the CEF ($E[Y|X]$) minimizes MSE among all possible predictors $m(X)$.*

There is some inherent CEF error, which is the first part of the decomposition. We can't get away from that; there's just some scatter that can't be reduced without adding new variables to the input set.

The second part measures the error added by approximating the CEF ($E[Y|X]$) by $m(X)$. If $m(X)$ equals the CEF, then that second term will always identically equal zero. It cannot get smaller! But otherwise, there will be at least some situation where $m(X) \neq E[Y|X]$ such that the second term has a positive squared expectation and the sum as a whole is larger than for the CEF.

2c. *Connect this proof to the lecture discussion about why we can't improve the CEF by adding something that depends on X. What does this imply about the relationship between the CEF error ($Y - E[Y|X]$) and X?*

The CEF already is the best possible predictor based on the information in the input variables. To do better at predicting Y, new information is needed. A variable that is just a transformation of X doesn't add new information, and therefore can't help with the CEF error, which is the first part of the decomposition. That's the only part that we can actually improve!

Problem 3

Install and load the required data:

```
#install.packages("poliscidata")
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)
```

3a.

Run a bivariate regression predicting voter turnout (`vep12_turnout`) based on income (`prcapinc` or `income_thousands`).

```
# Your code here
turnout_bivariate <- lm(vep12_turnout ~ income_thousands, data = states_data)
summary(turnout_bivariate)
```

```
##
## Call:
## lm(formula = vep12_turnout ~ income_thousands, data = states_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8558  -3.3015   0.0432   3.9256  13.5216
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       41.6116     6.2935   6.612  2.9e-08 ***
## income_thousands   0.5735     0.1951   2.939  0.00505 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.076 on 48 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.1349
## F-statistic: 8.639 on 1 and 48 DF,  p-value: 0.005047
```

Create a visualization that shows: 1. The raw data points 2. The BLP (linear regression line) 3. A LOESS curve to approximate the true CEF 4. Compare the two curves. Does the relationship appear linear?

```
# Your code for 3a here
library(ggplot2)

# Create the plot with both LOESS (CEF approximation) and linear (BLP) fits
turnout_wealth_plot <- ggplot(states_data, aes(x = income_thousands, y = vep12_turnout)) +
  geom_point(alpha = 0.3, color = "gray50") +  # Raw data
  geom_smooth(method = "loess", se = TRUE, color = "blue",
              aes(color = "LOESS (CEF approx)"), size = 1.2) +
  geom_smooth(method = "lm", se = TRUE, color = "red",
```

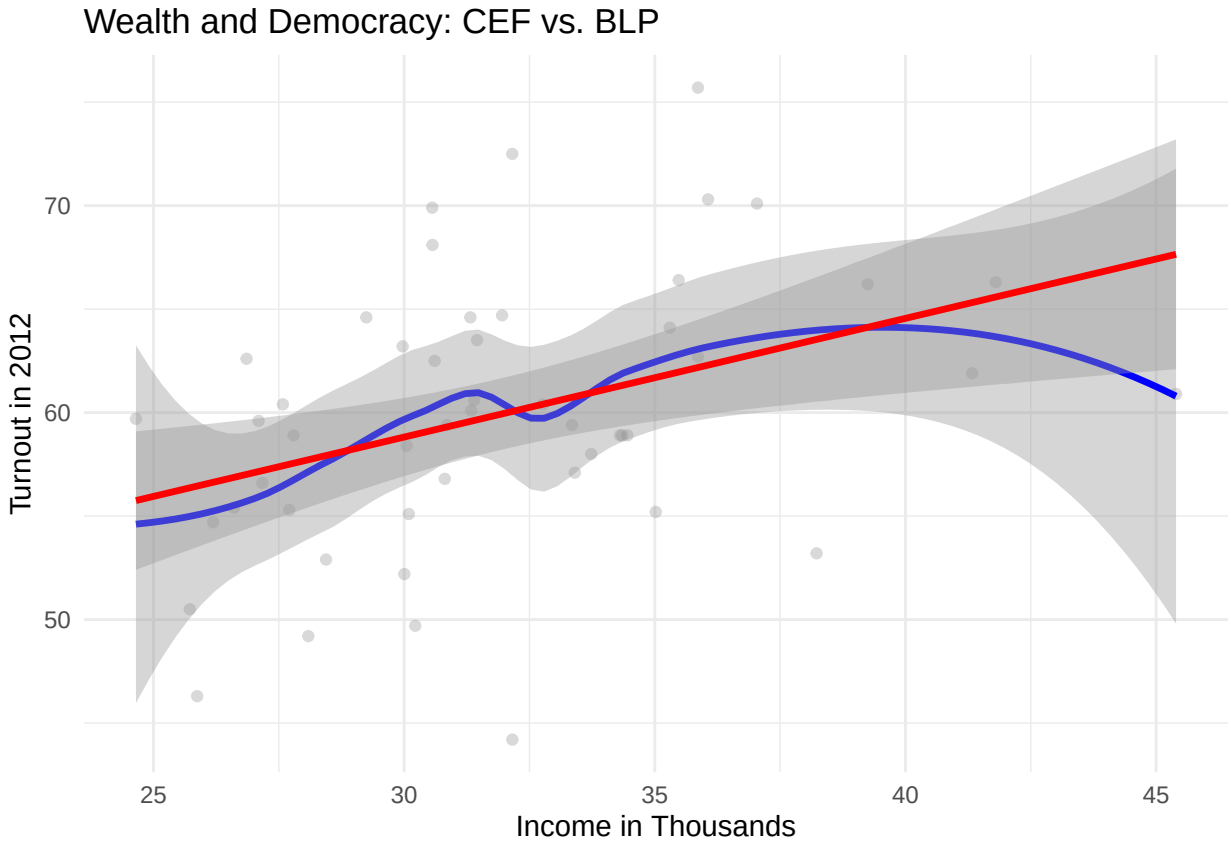
```
              aes(color = "Linear (BLP)"), size = 1.2) +
  scale_color_manual(values = c("LOESS (CEF approx)" = "blue",
                                "Linear (BLP)" = "red")) +
  labs(title = "Wealth and Democracy: CEF vs. BLP",
       x = "Income in Thousands",
       y = "Turnout in 2012",
       color = "Fit Type") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Display the plot
turnout_wealth_plot
```

```
## Ignoring unknown labels:
## * colour : "Fit Type"
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
```



Questions for 3a: *1. Interpret the slope coefficient from the bivariate regression.*

Every extra thousand dollars in average annual income at the state level is associated with an average 2012 turnout that is about half a point (0.57) higher.

2. Based on the visualization, does the BLP appear to be a good approximation of the CEF? Explain.

The LOESS curve is never far from the linear approximation and shows no systematic departures, suggesting that the BLP may indeed be a reasonable approximation of the CEF in this relationship.

3. Calculate and interpret R-squared.

R-squared is reported as 0.15, which tells us that the BLP corresponds to about fifteen percent of the variation in the underlying output variable.

3b.

Now run a multivariate regression predicting voter turnout based on income (`prcapinc`), religiosity (`religiosity`), and age distribution (`over64`).

```
# Your code here
turnout_multivariate <- lm(vep12_turnout ~ income_thousands + religiosity + over64,
                           data = states_data)
summary(turnout_multivariate)
```

```
##
## Call:
## lm(formula = vep12_turnout ~ income_thousands + religiosity +
##     over64, data = states_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0052  -3.7213   0.7419   3.9666  13.7749
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      44.77407    9.74654   4.594 3.39e-05 ***
## income_thousands  0.34790    0.25516   1.363    0.179
## religiosity      -0.02987    0.02144  -1.393    0.170
## over64            0.09013    0.49958   0.180    0.858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.077 on 46 degrees of freedom
## Multiple R-squared:  0.1876, Adjusted R-squared:  0.1347
## F-statistic: 3.542 on 3 and 46 DF,  p-value: 0.02169
```

Questions for 3b: *1. Interpret each coefficient in the multivariate model.*

Here, a shift of a thousand dollars in average income at the state level is only connected with a turnout increase of about a third of a percent. An increase of one point on the religiosity scale (which has a weird measurement scale) is connected with a minuscule decline in turnout of about 3 one hundredths of a percent. Finally, an increase of one percent in the population of a state that is over 64 is associated with about one tenth of a percent increase in turnout.

2. How does the coefficient for income change from the bivariate to multivariate model? What might explain this change?

The coefficient declines from 0.57 to 0.35. This shift is omitted variable bias connected with adding the other two variables. The researcher would have to decide whether those two variables are desirable controls or not.

3. Calculate the predicted voter turnout for a state with: income = \$50,000, religiosity = 50, over64 = 15%.

$44.77407 + 0.34790 * 50 + -0.02987 * 50 + 0.09013 * 15 = 62.02752$

3c.

Compare the two models:

```
# Model comparison
library(modelsummary)
models <- list("Bivariate" = turnout_bivariate,
               "Multivariate" = turnout_multivariate)
modelsummary(models, stars = TRUE, output = "markdown")
```

	Bivariate	Multivariate
(Intercept)	41.612***	44.774***
	(6.293)	(9.747)
income_thousands	0.574**	0.348
	(0.195)	(0.255)
religiosity		-0.030
		(0.021)
over64		0.090
		(0.500)
Num.Obs.	50	50
R2	0.153	0.188
R2 Adj.	0.135	0.135
AIC	326.3	328.2
BIC	332.0	337.7
Log.Lik.	-160.147	-159.089
F	8.639	3.542
RMSE	5.95	5.83

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

```
# Calculate and compare R-squared
cat("Bivariate R-squared:", summary(turnout_bivariate)$r.squared, "\n")
```

```
## Bivariate R-squared: 0.1525238
```

```
cat("Multivariate R-squared:", summary(turnout_multivariate)$r.squared, "\n")
```

```
## Multivariate R-squared: 0.1876339
```

Questions for 3c: *1. Which model has better fit? Does adding variables substantially improve the model?*

R-squared is basically pretty similar, but the more sensitive measures of AIC and BIC favor the bivariate model. This suggests that adding the variables may not substantially improve the model, at least as a tool for forecasting. If the goal is causal inference, the problem is much more complicated.

Problem 4

4a. Recall from the lectures that the BLP has two key properties: (1) $E[e] = 0$ and (2) $E[e \times X] = 0$. Verify these properties for your multivariate model from Problem 3:

```
# Calculate residuals
residuals <- residuals(turnout_multivariate)

# Property 1: Mean of residuals
```

```r
mean_residual <- mean(residuals)
cat("Mean of residuals:", mean_residual, "\n")
```

```
## Mean of residuals: -3.527907e-16
```

```r
# Property 2: Correlation of residuals with each predictor
cor_res_income <- cor(residuals, states_data$income_thousands, use = "complete.obs")
cor_res_relig <- cor(residuals, states_data$religiosity, use = "complete.obs")
cor_res_age <- cor(residuals, states_data$over64, use = "complete.obs")

cat("Correlation with income:", cor_res_income, "\n")
```

```
## Correlation with income: 2.316249e-16
```

```r
cat("Correlation with religiosity:", cor_res_relig, "\n")
```

```
## Correlation with religiosity: 1.779433e-17
```

```r
cat("Correlation with age:", cor_res_age, "\n")
```

```
## Correlation with age: -7.878278e-19
```

Questions for 4: *1. Do the residuals from your model satisfy the BLP properties? What might it mean if they don't?*

Yes, the residuals have mean almost exactly zero and are almost exactly uncorrelated with the included X variables. If they didn't meet these criteria, then some mathematical error would likely have happened in the estimation process.

2. Based on all your analyses, write a brief conclusion (3-4 sentences) about what affects voter turnout in U.S. states and how well linear regression captures these relationships.

From what we have seen, there is a positive relationship at the state level between income and turnout. There is limited evidence for expanding the analysis to include other demographic factors (religiosity and age do not add much at the state level). Linear regression captures the economic relationship reasonably well.