ProbSet 3, January 30

Gideon Gordon gideongordon2029@u.northwestern.edu

Due: January 30, 2026

0.1 Problem 1

1a. In a multiple regression model, $Y = {}_0 + {}_1X_1 + {}_2X_2 + $, interpret ${}_1$ in your own words.

${}_1$ is the effect on Y of a one-unit change in X_1, all else held equal and conditional on X_2. So, ${}_1$ in this equation might be different from the value it would take if there were an additional variable involved or if ${}_1X_1$ were the only variable term in the model.

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

$$\text{Terrorism} = 10 \;\; 0.2 \text{ Œ TrumpShare} + 0.1 \text{ Œ D12Margin}$$

Interpret what happens to predicted terrorism incidents when: - TrumpShare increases by 5 percentage points, holding D12Margin constant. - D12Margin decreases by 3 percentage points, holding TrumpShare constant.

TrumpShare increasing by 5 percentage points: TrumpShare2 = TrumpShare1 + 5

$$Terrorism_1 = 10 \;\; 0.2 \text{ Œ TrumpShare}_1 + 0.1 \text{ Œ D12Margin}$$

$$Terrorism_2 = 10\ \ 0.2 \text{ Œ } (\text{TrumpShare}_1 + 5) + 0.1 \text{ Œ } \text{D12Margin}$$

$$Terrorism_2\ \ Terrorism_1 = \text{Change in terrorist incidents}$$

$$Terrorism_2 Terrorism_1 = (1010)0.2Œ(\text{TrumpShare}_1+5)+0.2Œ\text{TrumpShare}_1+0.1ŒD12Margin0.1ŒD12M$$

Everything cancels out except $0.2 \text{ Œ } (\text{TrumpShare}_1 + 5)\ \ 0.2 \text{ Œ } \text{TrumpShare}_1$, and we get $0.2 \text{ Œ } 5$, so the change is a decrease of 1 terrorist incident.

We do the same with D12Margin decreasing by three percentage points, and I won't replicate all the math here, but everything cancels out until we get $0.1 \text{ Œ } 3$. A decrease of three points in D12Margin leads to an (AVERAGE) decrease of 0.3 terrorist incidents.

1c. The slides show that in multivariate regression, $_1$ 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 $_1$ as a partial effect?

The all others held constant condition is crucial because otherwise you don't know which term produced the increase in Y. But also, it enables us to unpack the effects of changes in X_1 when a nonlinear term is involved.

0.2 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.*

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

$Y\ \ m(X) = Y\ \ E[Y|X] + E[Y|X]\ \ m(X)$

The square of $Y\ \ m(X)$ is $Y^2\ \ 2Ym(X) + m(X)^2$.

Expectations are linear, so by linearity of expectations, we can say: $MSE(m) = E[Y^2]$ $2E[Ym(X)] + E[m(X)^2]$.

And now I'm stuck.

Remember: If we average $E[Y|X]$ across all values of X, it's just Y. So $E[E[Y|X]] = E[Y]$. Also, $E[Y|Y] = Y$. We can replace E[E[Y|X]] with E[Y] and vice versa.

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

...

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

Let me try doing what the hint says.

$Y\ \ m(X) = (Y\ \ E[Y|X]) + (E[Y|X]\ \ m(X))$ We've added and subtracted the same amount, so this checks out. Now we can square it. $(Y\ \ m(X))^2 = Y^2\ \ YE[Y|X] + YE[Y|X]$ $Ym(X)...\ \ YE[Y|X] + E[Y|X]^2\ E[Y|X]^2 + E[Y|X]m(X)...\ \ + YE[Y|X]\ E[Y|X]^2 + E[Y|X]^2$ $E[Y|X]m(X)...\ \ Ym(X) + E[Y|X]m(X)\ \ E[Y|X]m(X) + m(X)^2$

God that's a mess. We can cancel a bunch of terms though. $(Y\ \ m(X))^2 = Y^2\ \ 2Ym(X) +$ $m(X)^2$ ...and when we do, we get back where we started.

$(Y\ m(X))^2 = Y(Y\ 2m(X)) + m(X)^2\ E[(Y\ m(X))^2] = E[Y^2]\ 2E[Ym(X)] + E[m(X)^2]$

My main obstacle here is this center term $2E[Ym(X)]$.

We know that $E[XY] = E[X] + E[Y]\ Cov(X, Y)$, and that $Cov(X, Y) = E[(X\ E[X])(Y\ E[Y])]$.

We can rewrite $2E[Ym(X)]$ as $2[E[Y] + E[m(X)]\ Cov(Y, m(X)]$.

$E[(Y\ m(X))^2] = E[Y^2] + E[m(X)^2]\ 2E[Y]\ 2E[m(X)] + 2Cov(Y, m(X))$. And then by linearity of expectations we can regroup all this: $ ditto = E[Y^2 - 2Y] + E[m(X)^2 - 2m(X)] + 2Cov(Y, m(X)). ditto = E[Y(Y - 2)] + E[m(X)(m(X) - 2)] + 2[E(m(X) - E[m(X)])(Y - E[Y])]$ The covariance term here equals: $E[Ym(X)\ m(X)E[Y]\ YE[m(X)] + E[Y]E[m(X)]]$

By linearity of expectations we can say that this is also: $2E[Ym(X)]\ 2E[m(X)E[Y]]\ 2E[YE[m(X)]] + 2E[E[Y]E[m(X)]]$

$E[(Y\ m(X))^2] = E[Y^2] + E[m(X)^2]\ 2E[Y]\ 2E[m(X)] + 2E[Ym(X)]\ 2E[m(X)E[Y]]\ 2E[YE[m(X)]] + 2E[E[Y]E[m(X)]]$.

Aeeeeeeghhh.

$E[Y^2] = E[Y\ Œ\ Y]$, which is $E[Y] + E[Y] + Cov(Y, Y)$, and something's covariance with itself is its variance. So that gets us $E[Y^2] = 2E[Y] + Var(Y)$. Same goes for $E[m(X)^2]$. Maybe this can get me somewhere?

Okay I'm just sort of stuck.

Let's try starting at the end, one more time.

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

We can unpack these squares here. $(Y\ E[Y|X])^2 = Y^2 + E[Y|X]^2\ 2YE[Y|X]\ (E[Y|X]\ m(X))^2 = E[Y|X]^2 + m(X)^2\ 2m(X)E[Y|X]$

$E[(Y\ m(X))^2] = E[Y^2 + E[Y|X]^2\ 2YE[Y|X]] + E[E[Y|X]^2 + m(X)^2\ 2m(X)E[Y|X]]$ $ ditto

= E[Yˆ2] + E[E[Y|X]ˆ2] - 2E[Y E[Y|X]] + E[E[Y|Xˆ2]] + E[m(X)ˆ2] - 2E[m(X)E[Y|X]]$

We know that $E[E[Y|X]] = E[Y]$. We know that $E[Y^2] = 2E[Y] + Var(Y)$. We know that $Var(Y) = E[(Y \ E[Y])^2]$. So also, $Var(Y) = E[(Y \ E[E[Y|X]])^2]$.

Okay, I'm stuck for good and spending far too long on one problem.

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

$MSE(m) = E[(Y \ E[Y|X])^2] + E[(E[Y|X] \ m(X))^2]$. If we set the predictor $m(X) = E[Y|X]$, then: $MSE(m) = E[(Y \ E[Y|X])^2] + E[(E[Y|X] \ E[Y|X])^2]$. We cancel out the second term, so: $MSE(m) = E[(Y \ E[Y|X])^2] + 0$.

No other value for m(X) will cancel out the second term.

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?

Basically, if the CEF error term is correlated with X, some part of the error will be caught in m(X), so the CEF would not in that case be the best predictor.

0.3 Problem 3

Install and load the required data:

```
library(poliscidata)
```

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

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.5.2

Warning: package 'tibble' was built under R version 4.5.2

Warning: package 'tidyr' was built under R version 4.5.2

Warning: package 'readr' was built under R version 4.5.2

Warning: package 'purrr' was built under R version 4.5.2

-- 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    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.1
-- 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
```

```
library(ggplot2)


# 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.

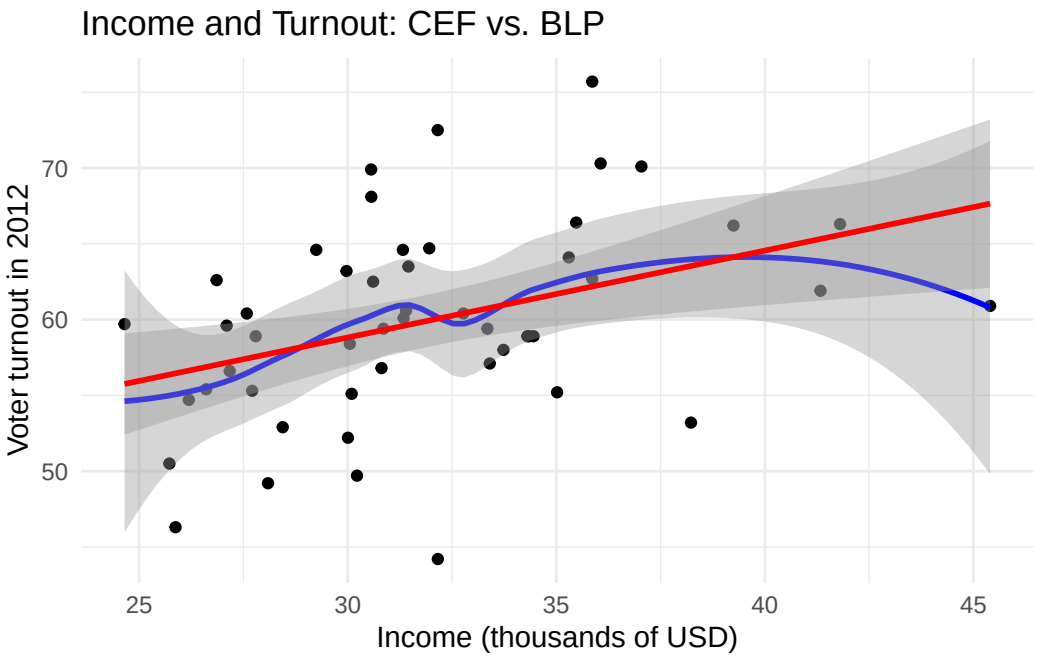
Run a bivariate regression predicting voter turnout (`vep12_turnout`) based on income (`prcapinc` or `income_thousands`).

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

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?

```
ggplot(data = states_data, mapping = aes(x = income_thousands, y = vep12_turnout)) +
  geom_point() +
  geom_smooth(method = "loess", se = TRUE, color = "blue",
              aes(color = "LOESS (CEF approx)")) +
  geom_smooth(method = "lm", se = TRUE, color = "red",
              aes(color = "Linear (BLP)")) +
  labs(title = "Income and Turnout: CEF vs. BLP",
       x = "Income (thousands of USD)",
       y = "Voter turnout in 2012",
       color = "Fit Type") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

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



The CEF does appear to be decently close to linear. Obviously hard to tell, but I think the linear relationship gives us a good approximation.

Questions for 3a: 1. Interpret the slope coefficient from the bivariate regression. The coefficient for per capita income is 0.5735, which tells me that, in general, if we have one group of states with a GDP per capita of N, and another group of states with a GDP per capita of N+1000, the latter group's average turnout in 2012 would probably be about half a percentage point higher. If we had *infinite* states for the 2012 election the difference between the two groups would be precisely 0.5735.

This is my very cautious phrasing.

I can't say that increasing GDP per capita increases turnout, because we only have data for one year and we haven't controlled for any confounding factors. I can't say that turnout will always be correlated with GDP per capita; it could be that in 2012, income had some special salience for turnout that it doesn't for other years.

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

There's a downturn at the very end in the LOESS curve because the curve fits the outlier at a bit over $45,000 per capita. But towards the center and on the left, the linear approximation works very well and for the whole length, the two are statistically indistinguishable.

3. Calculate and interpret R-squared. R-squared here is 0.1525, so it looks like about 15% of the variation in turnout can be explained by the bivariate regression we have.

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)
```

Questions for 3b: 1. Interpret each coefficient in the multivariate model. The coefficient for each variable is the average difference of observations one unit apart on that variable, conditional on all other variables, with all other variables held constant.

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

The coefficient is smaller; it looks like some large part of the effect of income was controlled for by the effects of religiosity and age.

I think the key to explaining the outcome here is age. It's hard to say, but if retirement pensions and income from investments count as income in this model, older people are more likely to have high incomes. Also, people who are low-income are less likely to live to old age. So people who are older are likely to be both richer and more likely to turn out (because it's well-known that the elderly turn out to vote at above-average rates). Controlling for age

captures some of the variation which would otherwise be explained by income.

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

```
(0.34790*50) - (-0.02987*50) + (0.09013*15) + 44.77
```

```
[1] 65.01045
```

In this moderately wealthy, pious, and elderly state, the predicted turnout is 65.01%.

3c.

Compare the two models:

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


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

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

The second model has a higher R2, so it seems at first glance that the new variables do improve the model. But also, the adjusted R2 is the same, meaning that the new variables did not improve the model by any more than we would expect from random change. So, it's not clear to me that the new model is much better, and the new variables are not significant.

0.4 Problem 4

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

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


# Property 1: Mean of residuals
mean_residual <- mean(residuals)
cat("Mean of residuals:", mean_residual, "\n")


# 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")
cat("Correlation with religiosity:", cor_res_relig, "\n")
cat("Correlation with age:", cor_res_age, "\n")
```

Questions for 4: 1. Do the residuals from your model satisfy the BLP properties? What might it mean if they don't? The residuals here are *VERY* slightly correlated with income, religiosity, and age. But it's such a tiny correlation that I think the BLP properties are mostly satisfied and the correlation may be a matter of random chance.

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.

```r
turnout_income_age <- lm(vep12_turnout ~ income_thousands + over64,
                         data = states_data)
summary(turnout_income_age)
```

```
Call:
lm(formula = vep12_turnout ~ income_thousands + over64, data = states_data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9718  -3.3880  -0.0718   3.8528  13.5598

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       40.1514     9.2553   4.338 7.56e-05 ***
income_thousands   0.5762     0.1975   2.918  0.00539 **
over64             0.1095     0.5044   0.217  0.82909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.138 on 47 degrees of freedom
Multiple R-squared:  0.1534,    Adjusted R-squared:  0.1173
F-statistic: 4.257 on 2 and 47 DF,  p-value: 0.01999
```

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

```
Call:
lm(formula = vep12_turnout ~ income_thousands + religiosity,
    data = states_data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9100  -4.1523   0.7776   4.0935  13.7442

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      45.99189    6.95785   6.610 3.19e-08 ***
income_thousands  0.34485    0.25197   1.369    0.178
religiosity      -0.02998    0.02121  -1.413    0.164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.014 on 47 degrees of freedom
Multiple R-squared:  0.1871,    Adjusted R-squared:  0.1525
F-statistic: 5.407 on 2 and 47 DF,  p-value: 0.007698
```

Looking at these alternative models (just including religiosity and income, just including age and income), we can see that the relationship between income and turnout remains significant when the percentage of people in a state over the age of 64 is included as a control variable, but controlling for religiosity reduces the estimated coefficient from about .55 to .34 and removes its significance. Note that most states have a negative value on the religiosity score, and outside the American South, the values are *very* negative. Also, the American South has a lower GDP per capita than the rest of the United States.

So controlling for religiosity, we separate out a source of inter-regional variation. The less religious a state is, the more wealthy it is, and also the more it votes. I can't say anything about causation, though. The chance that these findings are a result of random error is fairly high.