Problem Set 9

Due date: 4 December

Table of contents

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 40

```
library(tidyverse)
library(broom)
library(modelsummary)
library(marginaleffects)
library(poliscidata)
library(ggdist)
```

Question 1

Points: 10

Table 1 below reports the results from two regression models. In Model 1, in Table 1, Y is regressed on X_1 and, in Model 2, Y is regressed on both X_1 and X_2. Why might X_1 be statistically significant at conventional levels in Model 1 but statistically insignificant in Model 2? Be as specific as possible.

There are several reasons why we might see this change in the relationship between X_1 and Y with the introduction of X_2 into our model:

1. **Multi-collinearity**: If X_1 and X_2 largely vary together (in other words, they are highly correlated), together they provide redundant information into the model. This inflates the variance of the estimated regression coefficients, leading to higher standard errors. Higher standard errors lead to lower statistical significance.
2. **Mediation**: X_2 could be driving the relationship between X_1 and Y. Therefore, when we exclude X_2 from our model, its relationship with Y is captured through variation in X_1. Once we include X_2 in the model, we account for its effect directly.
3. **Different sample size**: If we have a lot more data on X_2 than X_1, our ability to capture the relationship between X_1 and Y is diminished.

The model that includes both X_1 and X_2 has an R^2 value of 0.54. Therefore, we continue to have a weak model of the drivers of Y. Going forward, I would fit a model of Y that only includes X_2 and compare the predictive power of that model including only X_1 and the model including both X_1 and X_2.

Question 2

Points: 10

Using the `censusAggregate` dataset (posted on ELMs) — which is survey data aggregated to the state level (1972-2000) — estimate a regression with `vote` as the dependent variable and the following independent variables: `nonSouth`, `edr`, and `pcthsg`. Report the results in a professionally formatted table and interpret the regression results.

Also, create a figure to display the predicted values (with confidence intervals) for the effect of `pcthsg` on the turnout rate. Lastly, is it meaningful to interpret the constant term on its own? Why, or why not?

> **Note**
>
> `vote` is the turnout rate in a state in a given year (i.e., the number of people who voted divided by the number eligible to vote).
> `nonSouth` is a dummy variable equal to `0` for Southern states and a `1` for non-Southern states.
> `pcthsg` is the percentage of the population in a state that graduated high school.
> `edr` is a dummy variable equal to `1` for states that used election-day registration and a `0` for states without election-day registration.

	(1)
(Intercept)	54.002
	t = 20.643
	se = 2.616
	p = <0.001
Non-Southern state	5.546
	t = 6.958
	se = 0.797
	p = <0.001
Election-day registration	5.792
	t = 5.483
	se = 1.056
	p = <0.001
Population that graduated high school (%)	0.101
	t = 2.787
	se = 0.036
	p = 0.006
Num.Obs.	357
R2	0.281
R2 Adj.	0.275
AIC	2255.8
BIC	2275.2
Log.Lik.	−1122.884
F	46.049
RMSE	5.62

```
census_df <- read_csv(here::here("data", "censusAggregate.csv"))
```

```
m <- lm(vote ~ nonSouth + edr + pcthsg, data = census_df)
```

```
modelsummary(m,
             coef_rename = c("nonSouth" = "Non-Southern state",
                             "edr" = "Election-day registration",
                             "pcthsg" = "Population that graduated high school (%)"),
             statistic = c("t = {statistic}", "se = {std.error}", "p = {p.value}"))
```

This model indicates that the turnout rate in a state in a given year is associated with whether the state is located in the South, or not; the percentage of the state's population that graduated high school; and whether the state used election-day registration.

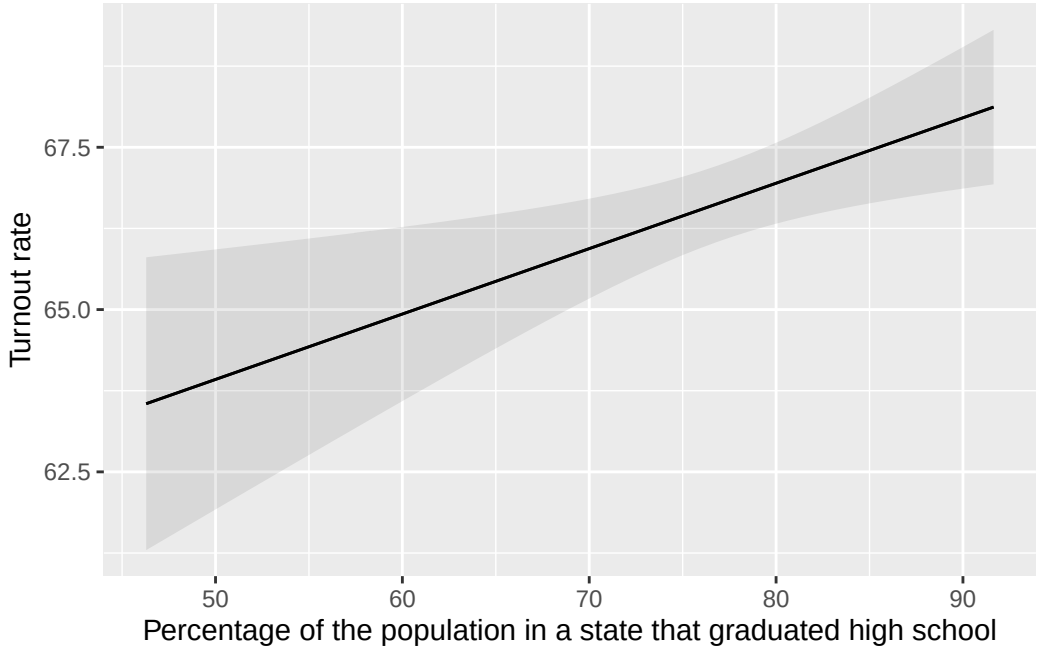
A non-Southern state's turnout rate is 5.55 percentage points higher than that of a Southern state, on average and holding all else constant.

A state that used election-day registration had a turnout rate that was 5.79 percentage points higher than that of a state without election-day registration, on average and holding all else constant.

A one percentage point increase in a state's high school graduation rate is associated with a 0.1 percentage point increase in its turnout rate, on average and holding all else constant.

We can use this model to predict the marginal effect of an increase in a state's high school graduation rate on its turnout rate, as visualized below.

```
plot_predictions(m, condition = "pcthsg") +
  labs(x = "Percentage of the population in a state that graduated high school",
       y = "Turnout rate")
```



A Southern state that did not use election-day registration and that has a high school graduation rate of zero percent is predicted to have a turnout rate of 54 percent. Because there are no states with a high school graduation rate of zero percent and there are unlikely to ever be states with no high school graduates, this intercept coefficient is not meaningful on its own.

	Full model	Null model
(Intercept)	54.002	65.936
	(2.616)	(0.348)
nonSouth	5.546	
	(0.797)	
edr	5.792	7.691
	(1.056)	(1.144)
pcthsg	0.101	
	(0.036)	
Num.Obs.	357	357
R2	0.281	0.113
R2 Adj.	0.275	0.110
AIC	2255.8	2326.9
BIC	2275.2	2338.5
Log.Lik.	−1122.884	−1160.454
F	46.049	45.183
RMSE	5.62	6.24

Question 3

Points: 5

Using the regression results from the previous question, evaluate the null hypothesis that the effects (i.e., regression coefficients) of `nonSouth` and `pcthsg` are jointly equal to zero. Can you reject the null hypothesis? Be sure to demonstrate how you reached a definitive conclusion.

My null model:

$$vote = \beta_0 + 0 * nonSouth + \beta_{EDR} EDR + 0 * pcthsg + \epsilon$$

```
m_null <- lm(vote ~ edr, data = census_df)
```

```
modelsummary(list("Full model" = m, "Null model" = m_null))
```

We can now compare the predictive power of our full model to that of the null model. The F-statistic associated with each model tells us whether the variables included provide more predictive power than if we were to simply guess the average value of our outcome of interest. We can compare the predictive power of our full model to that of the null model to determine whether the full model provides us with significantly more predictive power than the null model.

To do this, we need to calculate the F-ratio of our two models:

$$F = \frac{\frac{RSS_R - RSS_{UR}}{m}}{\frac{RSS_{UR}}{n-k-1}}$$

Where:

```
rss_r <- glance(m_null) |>
  pull(deviance)
rss_r
```

```
[1] 13918.85
```

```
rss_ur <- glance(m) |>
  pull(deviance)
rss_ur
```

```
[1] 11277.05
```

```
k <- 3

n <- nrow(census_df)
```

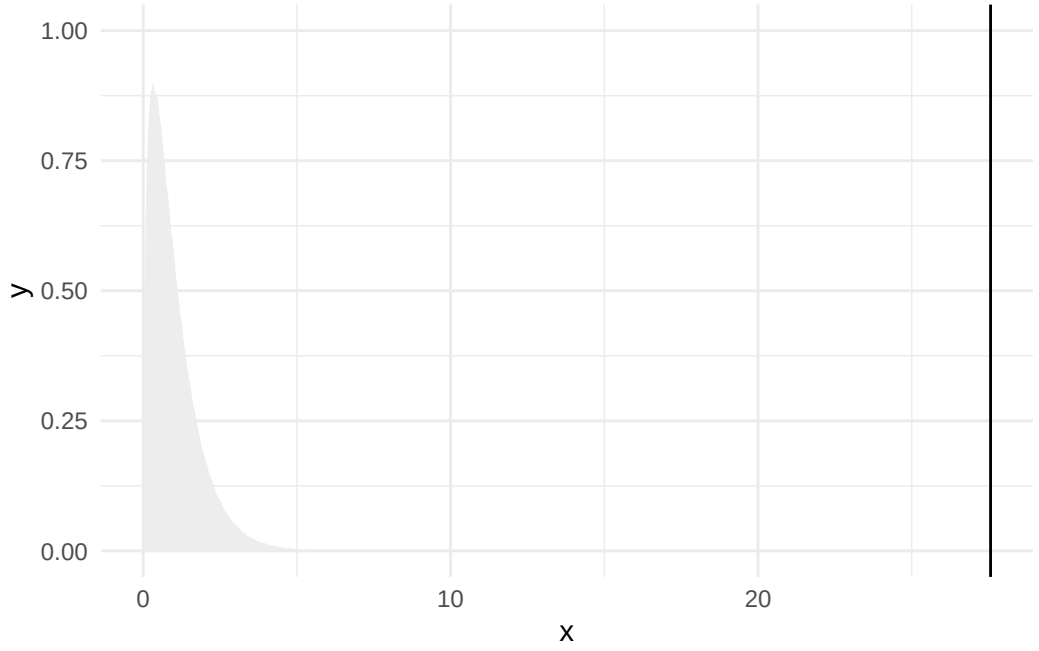
Therefore, the F-ratio is equal to:

```
f_ratio <- ((rss_r - rss_ur) / k) / (rss_ur / (n - k - 1))
f_ratio
```

```
[1] 27.565
```

Let's place this within the null world:

```
ggplot(tibble(x = rf(1e6, k, n-k-1)), aes(x = x)) +
  stat_slab(aes(fill_ramp = after_stat(x >= f_ratio))) +
  theme_minimal() +
  geom_vline(xintercept = f_ratio) +
  theme(legend.position = "none")
```



Another approach, which gives us the same result:

```
anova(m, m_null)
```

```
Analysis of Variance Table

Model 1: vote ~ nonSouth + edr + pcthsg
Model 2: vote ~ edr
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    353 11277
2    355 13919 -2   -2641.8 41.347 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 4

Points: 15

Using one of the other datasets available in the `poliscidata` package pick one dependent variable and two or more independent variables. Run a regression of the dependent variable on the independent variables. In your answer, describe why you picked the variables you did, produce a professionally formatted results table, and describe your statistical and substantive findings.