211-Final Exam

Yu-Shiuan (Lily) Huang

Fall 2023

Problem 1

Consider a random snapshot of the ANES 2022 dataset (which is named as `df` in R):

`fttrump`	`ideology`	`female`	`education`	`partyID`	`partyID3`
10	1	1	6	3	
83	5	1	3	4	
32	3	1	5	3	
95	6	0	2	5	
25	2	0	4	2	

`fttrump` records the feeling thermometer scores a voter rated for Donald Trump on a scale of 0-100; `ideology` records a voter's self-reported ideology on a scale of 1 to 7, where 1 indicates extremely liberal and 7 indicates extremely conservative; `female` records a voter's gender, where 1 indicates female and 0 indicates male; `education` records a voter's education level, where 1 indicates no high school, 2 indicates high school graduate, 3 indicates some college, 4 indicates 2-year college, 5 indicates 4-year college, and 6 indicates post-graduate; `party ID` records a voter's level of party affiliation, where 1 indicates strong Democrat, 2 indicates not very strong Democrat, 3 indicates lean Democrat, 4 indicates Independent, 5 indicates lean Republican, 6 indicates not very strong Republican, and 7 indicates strong Republican.

a. Which of the variables in this dataset are cardinal, which are ordinal and which are categorical? (optional for students in the methods subfield)

b. What are the mean, median, and standard deviation for the `fttrump` and `ideology` variables? Round your answer to 2 decimal places. (optional for students in the methods subfield)

c. What does the below lines of R code mean? Briefly explain the code and fill the output into the blank `partyID3` column in the above table.

```
df <- df %>%
  mutate(partyID3 = case_when(partyID < 4 ~ "Democrat",
                              partyID == 4 ~ "Independent",
                              partyID > 4 ~ "Republican"))
```

d. Briefly explain the following lines of R code and manually write out the output the code produces. Round your answer to 2 decimal places. (optional for students in the methods subfield)

```
df2 <- df %>%
  group_by(partyID3) %>%
  summarise(mean = mean(fttrump))
```

e. Briefly explain the following lines of R code and manually write out the output the code produces. Round your answer to 2 decimal places. (required for students in the methods subfield)

```
df3 <- df %>%
  filter(partyID3 == "Democrat") %>%
  mutate(college = ifelse(education > 4, 1, 0)) %>%
  group_by(college) %>%
  summarise(mean = mean(fttrump)) %>%
  summarise(diff = mean[college == 1] - mean[college == 0])
```

f. Calculate the covariance and correlation between `fttrump` and `ideology`. Round your answer to 2 decimal places. (optional for students in the methods subfield)

g. Calculate the linear regression slope and intercept for a regression with `fttrump` as the dependent variable and `ideology` as the independent variable. Interpret the relationship captured by this regression slope and the intercept in intuitive terms. Round your answer to 2 decimal places.

h. What are the differences between the covariance, correlation, and regression slope for `fttrump` and `ideology` you calculated earlier? What information does each of these statistics provide? (optional for students in the methods subfield)

i. Assuming one aims to estimate the linear relationship between a voter's feelings toward Trump (`fttrump`) and their level of holding populism values (`populism`), considering that one's populism level is influenced by both ideology (`ideology`) and support for radical politicians (i.e., `populism` = `ideology` + `fttrump`), what are the regression intercept and slope in the relationship between `fttrump` and `populism`? Round your answer to 2 decimal places and briefly interpret the relationship captured by this regression slope and the intercept in intuitive terms. (required for students in the methods subfield)

j. The linear regression line you calculated earlier is seen as the "best" guess for the linear relationship between `fttrump` and `ideology` as the line minimizes the sum of the squared residuals (SSR), which is defined as below.

$$SSR = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

By employing the formula for SSR, demonstrate that minimizing SSR leads to the same formulas used to calculate the intercept and slope in **g**. (required for students in the methods subfield)

Answer:

a. Cardinal variables: `fttrump` and `ideology`; ordinal variables: `education` and `partyID`; categorical variables: `female`.

b.

mean of `fttrump` $= \frac{10+83+32+95+25}{5} = 49$

standard deviation of `fttrump` $= \sqrt{\frac{(10-49)^2+83-49)^2+(32-49)^2+(95-49)^2+(25-49)^2}{5-1}} \cong 37.61$

mean of `ideology` $= \frac{1+5+3+6+2}{5} = 3.4$

standard deviation of `ideology` $= \sqrt{\frac{(1-3.4)^2+(5-3.4)^2+(3-3.4)^2+(6-3.4)^2+(2-3.4)^2}{5-1}} \cong 2.07$

c.

The below R code aims to generate a new variable, named `partyID3`, based on the original 7-point party ID variable `partyID`. Respondents who self-reported as strong Democrat (1), not very strong Democrat (2), or lean Democrat (3) are reclassified as `Democrat` in the `partyID3` variable. Those who self-reported as strong Republican (5), not very strong Republican (6), or lean Republican (7) are reclassified as `Republican` in the `partyID3` variable. Those who self-reported as independent (4) are classified as `Independent` in the `partyID3` variable.

`fttrump`	`ideology`	`female`	`education`	`partyID`	`partyID3`
10	1	1	6	3	**Democrat**
83	5	1	3	4	**Independent**
32	3	1	5	3	**Democrat**
95	6	0	2	5	**Republican**
25	2	0	4	2	**Democrat**

d. The following R code aims to calculate the mean ratings of respondents who identify as Democrats, Independents, and Republicans in terms of their feelings toward Trump, which are 22.33, 83, and 95, respectively.

e. The following R code is intended to compute the difference in the mean ratings of Democrats' feelings toward Trump between those with college degrees and those without. The difference should be -4, indicating that Democrats with college degrees rated Trump 4 units lower than those without college degrees.

f.

$Cov(\texttt{ideology}, \texttt{fttrump}) = \frac{(1-3.4)(10-49)+(5-3.4)(83-49)+(3-3.4)(32-49)+(6-3.4)(95-49)+(2-3.4)(25-49)}{5-1} = 77$

$Cor(\texttt{ideology}, \texttt{fttrump}) = \frac{77}{2.07*37.61} = 0.99$

g.

slope $= \frac{77}{2.07^2} = 17.97$

intercept $= 49 - 3.4 * 17.97 = -12.10$

The intercept represents the average ratings of respondents' feelings toward Trump when the ideology is equal to 0. The slope indicates the change in respondents' average ratings of their feelings toward Trump associated with a one-unit increase in ideology. As ideology becomes more conservative by one unit, respondents' ratings of Trump increase by 17.97 units.

h. Covariance is a useful measure at describing the direction of the linear association between two quantitative variables. In this case, the positive covariance computed in **f** provides information that both `ideology` and `fttrump` are larger than the average values in the data set. However, a larger covariance does not always mean a stronger relationship, and we cannot compare the covariances across different sets of relationships. To account for the weakness, we normalize the covariance by the standard deviation of the `ideology` and `fttrump`, to get the correlation coefficient. The correlation coefficient is a value between -1 and 1, and measures both the direction and the strength of the linear association. In this case, the correlation computed in **f** suggests that `ideology` and `fttrump` are strongly positively correlated, the correlation is very close to 1. Both covariance and correlation do not account for the slope of the relationship. In other words, we do not know how a change in one variable could impact the other variable. Regression is the technique that fills this void — it allows us to make the best guess at how one variable affects the other variables.

i.

From **g**, we know that:

$$\hat{\texttt{fttrump}}_i = -12.10 + 17.97\texttt{ideology}_i$$

We can rewrite the regression to:

$$\texttt{ideology}_i = \frac{12.10}{17.97} + \frac{\hat{\texttt{fttrump}}_i}{17.97}$$

According to the question, we also know that:

$$\texttt{populism}_i = \texttt{fttrump}_i + \texttt{ideology}_i$$

By inserting the previous equation, we obtain:

$$\hat{\texttt{populism}}_i = \texttt{fttrump}_i + \frac{12.10}{17.97} + \frac{\hat{\texttt{fttrump}}_i}{17.97} = 0.67 + 1.06\texttt{fttrump}_i$$

The intercept represents respondents' average levels of populism attitudes when the average of their feelings toward Trump is 0. The slope indicates the change in respondents' average levels of holding populism values associated with a one-unit positive increase in their feelings toward Trump. As they like Trump more by one unit, respondents' levels of holding populism values increase by 1.06 units.

j.

$$
\begin{aligned}
SSR &= \sum_{i=1}^{n}(y_i - (a + bx_i))^2 \\
&= \sum_{i=1}^{n}(y_i^2 - 2y_i(a + bx_i) + a^2 + 2abx_i + b^2x_i^2)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial SSR}{\partial a} &= \sum_{i=1}^{n}(-2y_i + 2a + 2bx_i) \\
0 &= \sum_{i=1}^{n}(-y_i + a + bx_i) \\
0 &= -n\bar{y} + na + bn\bar{x} \\
a &= \bar{y} - b\bar{x}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial SSR}{\partial b} &= \sum_{i=1}^{n}(-2x_iy_i + 2ax_i + 2bx_i^2) \\
0 &= -\sum_{i=1}^{n} x_iy_i + a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \\
0 &= -\sum_{i=1}^{n} x_iy_i + (\bar{y} - b\bar{x})\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \\
b &= \frac{n\sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\
b &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \\
b &= \frac{Cov(x, y)}{s_x^2}
\end{aligned}
$$

Problem 4

According to a study by Danny Hayes published in AJPS in 2005, the average American voter perceives Republican candidates as stronger leaders and more morally inclined. On the other hand, Democratic candidates are seen as more compassionate and empathetic. The table below presents voters' average ratings of Democratic and Republican candidates. Each voter provides ratings for both Republican and Democratic candidates on a scale from 1 to 4, with 1 being the lowest and 4 being the highest. The numbers in parentheses represent the standard errors.

	Republican Candidate	Democratic Candidate	Difference
Strong Leader ($n = 7,396$)	2.73 (0.01)	2.50 (0.01)	0.23
Moral ($n = 7,060$)	3.05 (0.01)	2.75 (0.01)	0.30
Compassionate ($n = 3,531$)	2.67 (0.01)	2.93 (0.01)	-0.26
Empathetic ($n = 7,152$)	2.37 (0.01)	2.69 (0.01)	-0.32

Regarding the trait of being a "strong leader," at a significance level of $\alpha = 0.05$, perform a two-tailed hypothesis test to demonstrate whether there is significant difference between Republican and Democratic candidates.

a. What are the null and alternative hypotheses? (optional for students in the methods subfield)

b. What is the distribution of the difference in average strong leader rating between Republican and Democratic candidates under the null hypothesis? (optional for students in the methods subfield)

c. What is the critical value of t that you would use to perform the hypothesis test? Round your answer to 2 decimal places.

d. Does the following R code calculate the p-value for observing the data you observed or data that is more extreme, assuming the null hypothesis is true? If so, why? If not, why not?

```
pnorm(t, 0, 1)
```

e. Briefly describe the process for making a decision regarding the null hypothesis. Under what conditions will you reject or fail to reject it? (optional for students in the methods subfield)

f. Briefly discuss when will the Type I and Type II errors occur. What is the probability of Type I error? How can we reduce the chance of Type I error?

g. For a two-tailed test with a significance level of 0.05, the critical value is approximately -1.96 and 1.96 on the left and right tails of a standard normal distribution, respectively. Does the following R code calculate the Type II error? If so, why? If not, why not? **Note**: The `SD` in the R code represents the standard error you computed in **b**. (required for students in the methods subfield)

```
pnorm(1.96*SD, 0, SD)
```

h. If the author, Danny Hayes, intends to conduct a replication study on the same topic, aiming for a power of 0.95 in a one-tailed test comparing whether there is a positive significant difference in the average strong leader ratings between Republican and Democratic candidates, what sample size would be required to achieve this level of power? Round your answer to the nearest integer. **Note**: Assuming a significance level of 0.01, and that the difference in means and the standard deviation of the average strong leader rating of Republican and Democratic candidates remain the same as in the previous questions when Prof. Hayes draws new samples. The critical values on the right tail of a standard normal distribution are approximately 1.645 for a 95% cumulative distribution function (CDF) and 2.33 for a 99% CDF. (required for students in the methods subfield)

Answer:

a.

H_0: There is no difference between Republican and Democratic candidates in terms of the trait of being a strong leader; $\mu_R - \mu_D = 0$

H_a: There is difference between Republican and Democratic candidates in terms of the trait of being a strong leader; $\mu_R - \mu_D \neq 0$

b. $\bar{X}_R - \bar{X}_Y \sim N(0, \sqrt{0.0002})$

- standard deviation of Republican candidate: $SE_R = \frac{s_R}{\sqrt{7396}} = 0.01$, thus, $s_R = 0.01 * \sqrt{7396} = 0.86$
- standard deviation of Democratic candidate: $SE_D = \frac{s_D}{\sqrt{7396}} = 0.01$, thus, $s_D = 0.01 * \sqrt{7396} = 0.86$
- standard error:$\sqrt{\frac{0.86^2+0.86^2}{7396}} = \sqrt{0.0002}$

c. $t = \frac{0.23-0}{\sqrt{0.0002}} \cong 16.26$

d. No, as this is a two-tailed test, the correct way to calculate the p-value in R should be:

```r
2*pnorm(t, 0, 1)
```

e. When the p-value we computed in d is larger than the threshold of 0.05, we would fail to reject the null hypothesis, indicating that there is not enough evidence to suggest a difference between Republican and Democratic candidates in terms of the trait of being a strong leader. On the other hand, if the p-value is smaller than the threshold of 0.05, we can reject the null hypothesis and conclude that there is a significant difference between Republican and Democratic candidates in terms of the trait of being a strong leader.

f. Type I error would occur when there is no difference between Republican and Democratic candidates in terms of the trait of being a strong leader (H_0), yet we reject it. Type II error would occur when there is difference, yet we fail to reject the null hypothesis. In this case, the probability of having Type I error is the significance level that is set in the question, 0.05. One way to reduce the probability of Type I error is by lowering the significance level (α) used in hypothesis testing.

g. When calculating the probability of Type II error, we assume the null hypothesis is false. Therefore, we should not calculate the probability of Type II error under the assumption that the null hypothesis is true, but rather under the assumption that the observed sample statistic is more likely to be true. In this case, when using the `pnorm` function in R, the `mean` argument should be set to 0.23 instead of 0. Additionally, as this is a two-tailed test, the probability of Type II error should consider both the left and right tail sides. To summarize, the correct way to calculate the probability of Type II error in R is as follows:

```r
pnorm(1.96*sqrt(0.0002), 0.23, sqrt(0.0002)) -
  pnorm(-1.96*sqrt(0.0002), 0.23, sqrt(0.0002))
```

h. $\bar{X}_R - \bar{X}_Y \sim N(0, \frac{1.4792}{\sqrt{n}})$

$$
\begin{aligned}
P(\frac{\Delta X - 0}{\frac{1.4792}{\sqrt{n}}} \geq 2.33) &= 0.95 \\
P(\frac{\Delta X - 0.23}{\frac{1.4792}{\sqrt{n}}} \geq 2.33 + \frac{0 - 0.23}{\frac{1.4792}{\sqrt{n}}}) &= 0.95 \\
P(Z < 2.33 + \frac{-0.23}{\frac{1.4792}{\sqrt{n}}}) &= 0.05 \\
2.33 - \frac{0.23}{\frac{1.4792}{\sqrt{n}}} &= -1.645 \\
\frac{0.23}{\frac{1.4792}{\sqrt{n}}} &= 3.975 \\
n &\cong 654
\end{aligned}
$$