Problem Set 3

Applied Stats/Quant Methods 1

Qin Guo 24338859

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should b(e submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday November 11, 2024. No late assignments will be accepted.

In this problem set, you will run several regressions and create an add variable plot (see the lecture slides) in `R` using the `incumbents_subset.csv` dataset. Include all of your code.

Question 1

We are interested in knowing how the difference in campaign spending between incumbent and challenger affects the incumbent's vote share.

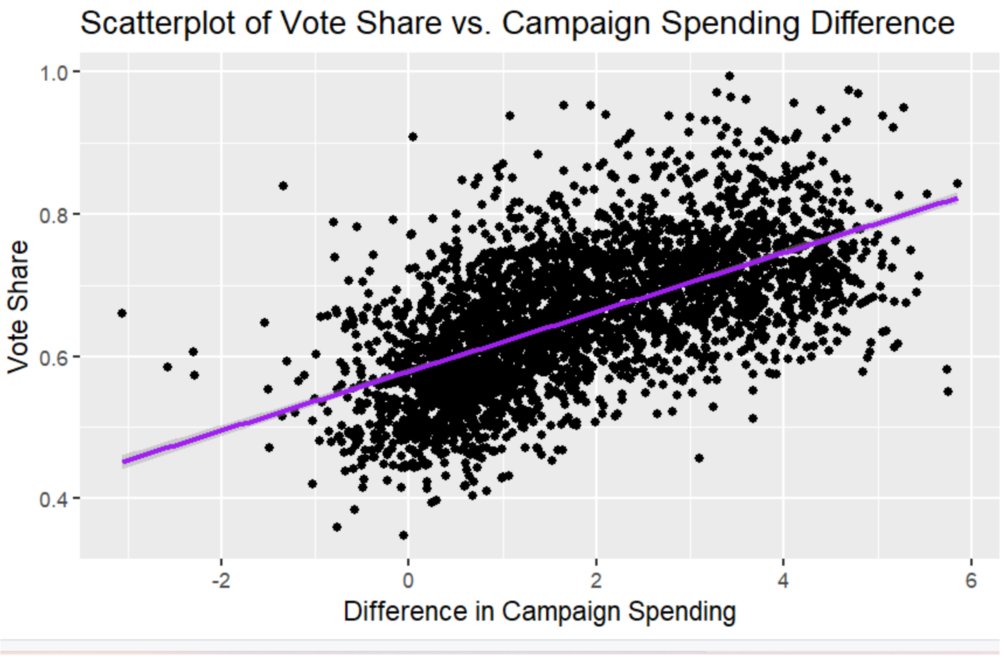
1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `difflog`.

```r
# read in data
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/incumbents_subset.csv")


#Question1:
# Run a regression where the outcome variable is voteshare and the explanatory variable is difflog
model1 <- lm(voteshare ~ difflog, data=inc.sub)
```

2. Make a scatterplot of the two variables and add the regression line.

```r
# Make a scatterplot of the two variables and add the regression line
library(ggplot2)
ggplot(inc.sub, aes(x=difflog, y=voteshare)) +
  geom_point() +
  geom_smooth(method="lm", col="purple") +
  labs(title="Scatterplot of Vote Share vs Campaign Spending Difference",
     x="Difference in Campaign Spending", y="Vote Share")
```



3. Save the residuals of the model in a separate object.

```r
# Save the residuals of the model in a separate object
residuals1 <- residuals(model1)
```

4. Write the prediction equation.

```r
# Write the prediction equation
coefficients1 <- coef(model1)
prediction_equation1 <- paste("voteshare = ",
                              round(coefficients1[1], 2), " + ",
                              round(coefficients1[2], 2), " * difflog")
print(prediction_equation1)
```

$$\textit{voteshare} = 0.58 + 0.04 \times \text{difflog}$$

Question 2

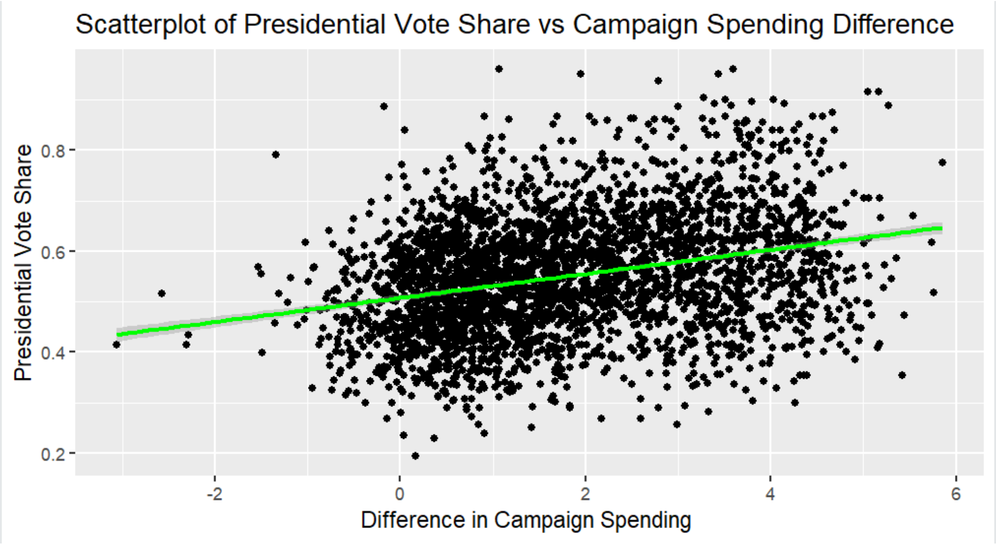
We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

1. Run a regression where the outcome variable is `presvote` and the explanatory variable is `difflog`.

```r
# Run the regression
model2 <- lm(presvote ~ difflog, data = inc.sub)
```

2. Make a scatterplot of the two variables and add the regression line.

```r
# Create a scatterplot and add the regression line
ggplot(inc.sub, aes(x = difflog, y = presvote)) +
  geom_point() +  # Add scatterplot
  geom_smooth(method = "lm", color = "green") +  # Add regression line
  labs(title = "Scatterplot of Presidential Vote Share vs Campaign Spending Difference",
       x = "Difference in Campaign Spending ",
       y = "Presidential Vote Share")
```



3. Save the residuals of the model in a separate object.

```r
# Save the residuals of the model in a separate object
residuals2 <- residuals(model2)
```

4. Write the prediction equation.

```r
# Write the prediction equation
coefficients2 <- coef(model2)
prediction_equation2 <- paste("presvote = ",
                              round(coefficients[1], 2), " + ",
                              round(coefficients[2], 2), " * difflog")
print(prediction_equation2)
```

$$\textit{voteshare} = 0.51 + 0.02 \times \text{difflog}$$

Question 3

We are interested in knowing how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success.

1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`.

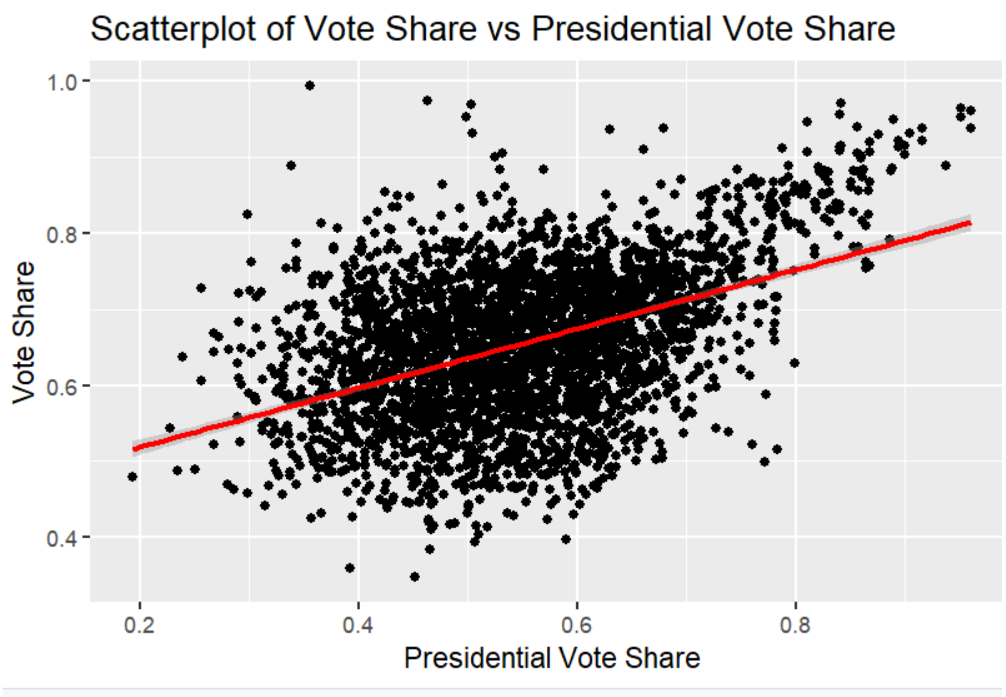
```r
# Run the regression
model3 <- lm(voteshare ~ presvote, data = inc.sub)
```

2. Make a scatterplot of the two variables and add the regression line.

```r
# Create a scatterplot and add the regression line
ggplot(inc.sub, aes(x = presvote, y = voteshare)) +
  geom_point() +  # Add scatterplot
  geom_smooth(method = "lm", color = "red") +  # Add regression line
  labs(title = "Scatterplot of Vote Share vs Presidential Vote Share",
       x = "Presidential Vote Share",
       y = "Vote Share")
```

3. Write the prediction equation.

```r
# Write the prediction equation
coefficients3 <- coef(model3)
prediction_equation3 <- paste("voteshare = ",
                              round(coefficients3[1], 2), " + ",
                              round(coefficients3[2], 2), " * presvote")
print(prediction_equation3)
```



$$voteshare = 0.44 + 0.39 \times \text{presvote}$$

Question 4

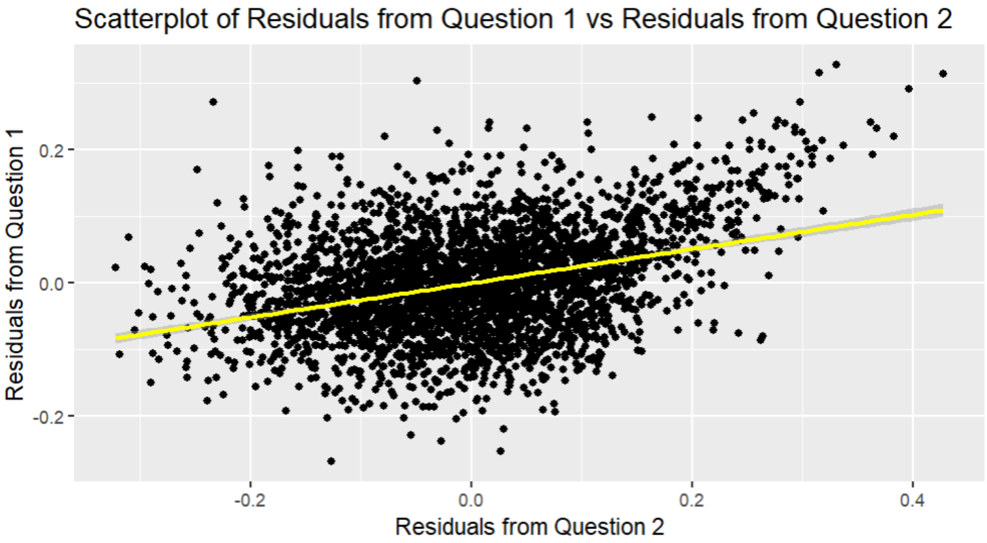
The residuals from part (a) tell us how much of the variation in `voteshare` is *not* explained by the difference in spending between incumbent and challenger. The residuals in part (b) tell us how much of the variation in `presvote` is *not* explained by the difference in spending between incumbent and challenger in the district.

1. Run a regression where the outcome variable is the residuals from Question 1 and the explanatory variable is the residuals from Question 2.

```
# Run the regression
model4 <- lm(residuals1 ~ residuals2)
```

2. Make a scatterplot of the two residuals and add the regression line.

```
# Create a scatterplot and add the regression line
ggplot(data.frame(residuals1, residuals2), aes(x = residuals2, y =
    residuals1)) +
  geom_point() + # Add scatterplot
  geom_smooth(method = "lm", color = "yellow") + # Add regression line
  labs(title = "Scatterplot of Residuals from Question 1 vs Residuals
    from Question 2",
       x = "Residuals from Question 2",
       y = "Residuals from Question 1")
```



3. Write the prediction equation.

```
# Write the prediction equation
coefficients4 <- coef(model4)
prediction_equation4 <- paste("residuals1 = ",
                              round(coefficients4[1], 2), " + ",
                              round(coefficients4[2], 2), " * residuals2"
  )
print(prediction_equation4)
```

$$residuals1 = 0 + 0.26 \times \text{residuals2}$$

Question 5

What if the incumbent's vote share is affected by both the president's popularity and the difference in spending between incumbent and challenger?

1. Run a regression where the outcome variable is the incumbent's `voteshare` and the explanatory variables are `difflog` and `presvote`.

```
  # Run the regression
  model5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
```

2. Write the prediction equation.

```
# Write the prediction equation
coefficients5 <- coef(model5)
prediction_equation5 <- paste("voteshare = ",
                              round(coefficients5[1], 2), " + ",
                              round(coefficients5[2], 2), " * difflog + "
  ,
                              round(coefficients5[3], 2), " * presvote")
print(prediction_equation5)
```

$$voteshare = 0.45 + 0.04 \times \text{difflog} + 0.26 \times \text{presvote}$$

3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case?

   The identical element in the output of this question compared to Question 4 is likely the coefficient for presvote.
   Because in this question, we conducted a multiple regression analysis in which we considered both the effects of campaign funding differences (difflog) and the presidential candidate's vote share (presvote) on the incumbent's vote share (voteshare). In this model, we are trying to understand how the presidential candidate's vote share affects the incumbent's vote share after accounting for funding differences. In Question 4, we are actually analyzing the relationship between two residuals: one is the residual of the incumbent's vote share (the part not explained by funding differences in Question 1), and the other is the residual of the presidential candidate's vote share (the part not explained by funding differences in Question 2). This analysis helps us understand the relationship between the presidential candidate's vote share and the incumbent's vote share after excluding the impact of funding differences.
   The key point is that in both cases, we are trying to solve the same mystery: how much impact does the presidential candidate's vote share have on the incumbent's vote share after excluding the impact of funding differences. In question 5, we did this by directly using funding differences and presidential candidate vote share as explanatory variables, whereas in question 4, we did this indirectly by looking at the residuals of the two variables.
   So, despite the different approaches, both analyses are trying to answer the same question, which is why the coefficient of presvote in question 5 is the same as the coefficient of presvote in the residual regression in question 4. This coefficient reflects the net effect of the presidential candidate's vote share on the incumbent's vote share after controlling for funding differences.