Problem Set 3

Applied Stats/Quant Methods 1

Due: November 11, 2024

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 be 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
setwd("C:/Users/julia/OneDrive/Desktop/R Data")
getwd()

incumbent_subset <- read.csv("incumbents_subset.csv")

install.packages("stargazer")
library(stargazer)

model <- lm(voteshare ~ difflog, data = incumbent_subset)
summary(model)

stargazer(model, type = "latex")
```

Table 1: Campaign Spending and Voteshare

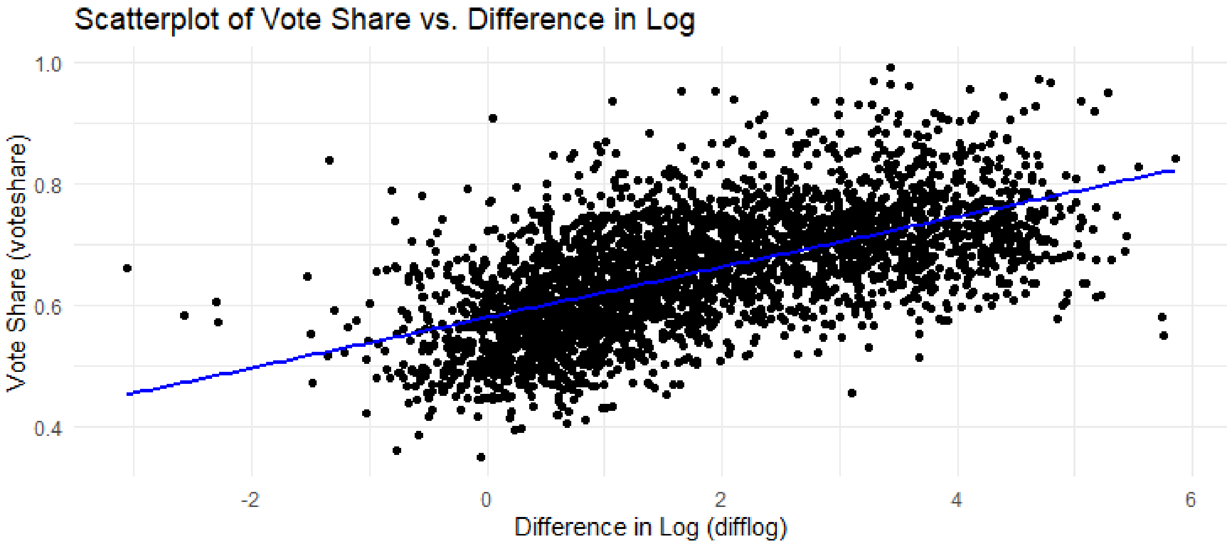
	Dependent variable:
	voteshare
difflog	0.042***
	(0.001)
Constant	0.579***
	(0.002)
Observations	3,193
R^2	0.367
Adjusted R^2	0.367
Residual Std. Error	0.079 (df = 3191)
F Statistic	1,852.791*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

The outcome of regressing difflog on voteshare cannot be interpreted easily, as difflog, meaning difference in campaign spending of incumbent and challenger party, is a logged varible. However, a positive relationship can be assumed, as the coefficient 0.042 is positive and significant at a p-value of 0.001. Consequently, one can assume, that a larger difference between campaign spending,on average has a positive effect on the voteshare of the incumbent party.

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

```


library(ggplot2)
ggplot(incumbent_subset, aes(x = difflog, y = voteshare)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatterplot of Vote Share vs. Difference in Log",
       x = "Difference in Log (difflog)",
        y = "Vote Share (voteshare)") +
  theme_minimal()
```



The scatterplot shows on the Y-axis the voteshare of the incumbent, and on the X-axis the difference in campaign spending between the incumbent and the challenger. Again, a positive correlation is evident, showing that a growing difference in campaign spending increases the voteshare for the incumbent.

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

```
residuals_model <- residuals(model)
head(residuals_model)
```

4. Write the prediction equation

$$\text{voteshare} = \beta_0 + \beta_1 \times \text{difflog} + \epsilon \tag{1}$$

$$\text{voteshare} = 0.579031 + 0.041666 \times \text{difflog} + \epsilon \tag{2}$$

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

model_presvote <- lm(presvote ~ difflog , data = incumbent_subset)

summary(model_presvote)
```

Table 2: Campaign Spending and Presvote

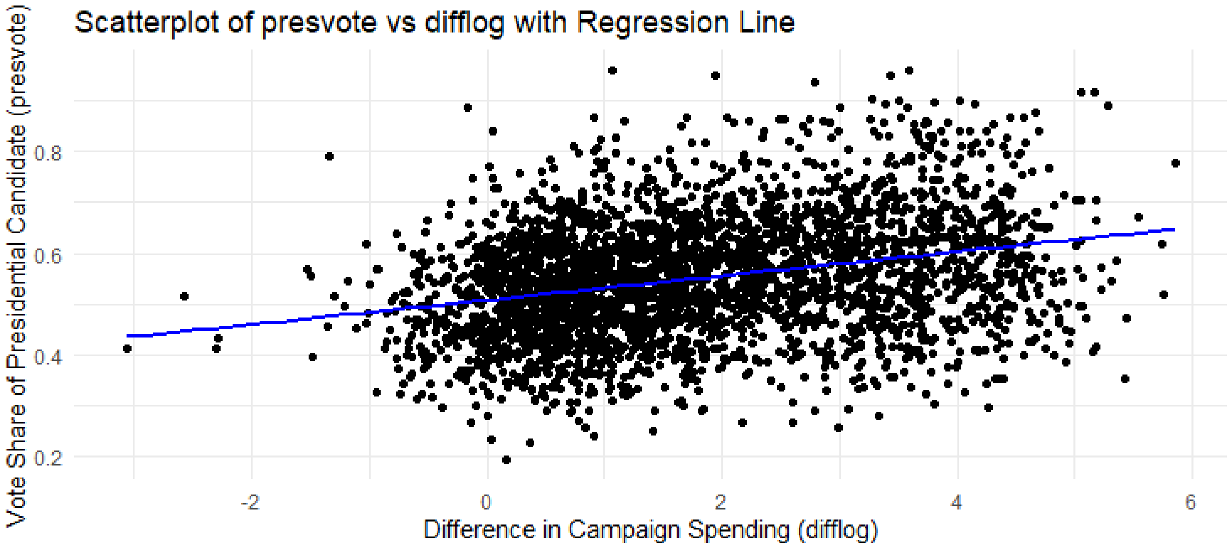
	Dependent variable:
	presvote
difflog	0.024***
	(0.001)
Constant	0.508***
	(0.003)
Observations	3,193
R^2	0.088
Adjusted R^2	0.088
Residual Std. Error	0.110 (df = 3191)
F Statistic	307.715*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

In table 2, the difference in campaign spending (difflog) is regressed on presidential voteshare (presvote). The coefficient is again not straightforward to interpret, as it remains a logged varibale, however, the positive coefficient of 0.024, significant at a p-value of 0.001 assumes a positive relationship.

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

```
library(ggplot2)

ggplot(incumbent_subset, aes(x = difflog, y = presvote)) +
  geom_point() +                                     # Scatterplot
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(x = "Difference in Campaign Spending (difflog)",
       y = "Vote Share of Presidential Candidate (presvote)",
       title = "Scatterplot of presvote vs difflog with Regression Line")
    +
  theme_minimal()
```



This scatterplot shows the Voteshare of the presidential candidate on the Y-axis, and the differences in campaign spending between the challanger and the incumbent on the X-axis. One can assume from this data, that a larger gap in campaign spending, increases the voteshare of the presidential candidate.

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

```
residuals_presvote <- residuals(model_presvote)

```

```r
head(residuals_presvote)
```

4. Write the prediction equation.

$$\text{presvote} = 0.507583 + 0.023837 \times \text{difflog} + \epsilon \tag{3}$$

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
model_voteshare <- lm(voteshare ~ presvote, data = incumbent_subset)

summary(model_voteshare)
```

Table 3: Presvote and Voteshare of the Incumbent

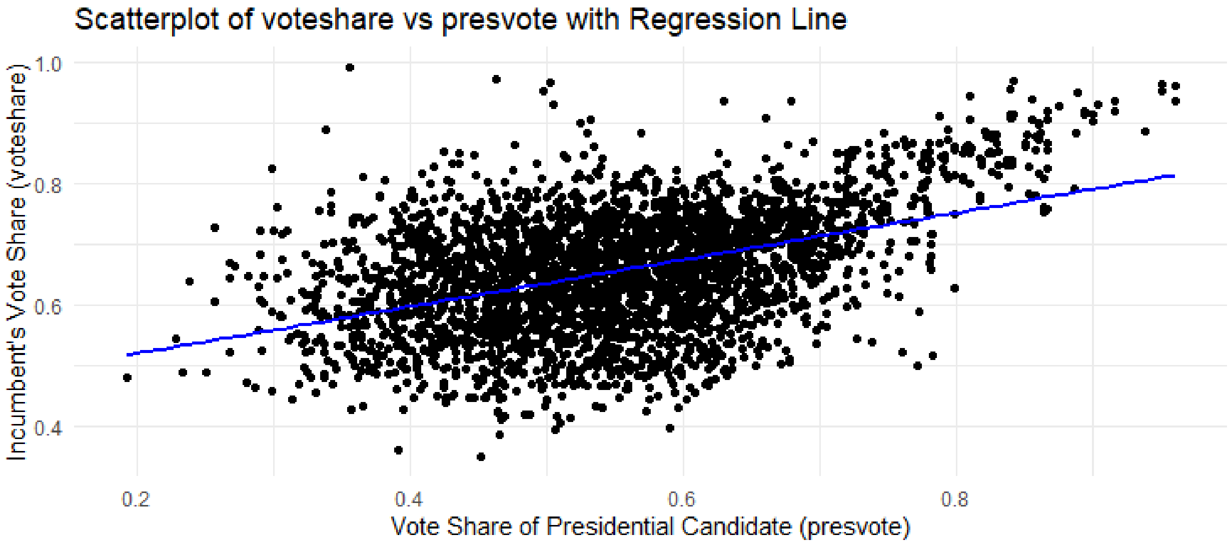
	Dependent variable:
	voteshare
presvote	0.388***
	(0.013)
Constant	0.441***
	(0.008)
Observations	3,193
R^2	0.206
Adjusted R^2	0.206
Residual Std. Error	0.088 (df = 3191)
F Statistic	826.950*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

In table 3, the voteshare of the presidential candidate of the incumbent (presvote) is regressed on the voteshare of the incumbent party (voteshare). Here, the coefficient is 0.388 and significant at a p-value of 0.013. Hence, one can conclude that a one-unit increase of the presvote causes a 0.388 percent increase in voteshare, holding all else constant. Thus, the increase of the voteshare of the incumbents presidential candidate has a significant and positive effect on the incumbents electoral success overall.

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

```r
library(ggplot2)

ggplot(incumbent_subset, aes(x = presvote, y = voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(x = "Vote Share of Presidential Candidate (presvote)",
       y = "Incumbent's Vote Share (voteshare)",
       title = "Scatterplot of voteshare vs presvote with Regression Line
  ") +
  theme_minimal()
```



This scatterplot shows the incumbents voteshare on the Y-axis, and the voteshare of the incumbents presidential candidate on the X-axis. Just as it was shown by the regression above, a positive relationship between the two varibales can be assumed. Although strong outliners exist, made visible by the scatterplot, an upard trend of the regression line is visible.

3. Write the prediction equation.

$$\text{voteshare} = 0.441330 + 0.388018 \times \text{presvote} + \epsilon \tag{4}$$

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.

```r
residuals_voteshare <- residuals(model_voteshare)

residuals_presvote <- residuals(model_presvote)

residual_model <- lm(residuals_voteshare ~ residuals_presvote)
summary(residual_model)

stargazer(residual_model, type = "latex")
```

Table 4: Regression of Residuals of Table 1 and Table 2

	Dependent variable:
	residuals_voteshare
residuals_presvote	0.257***
	(0.012)
Constant	−0.000
	(0.001)
Observations	3,193
R^2	0.130
Adjusted R^2	0.130
Residual Std. Error	0.073 (df = 3191)
F Statistic	476.975*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

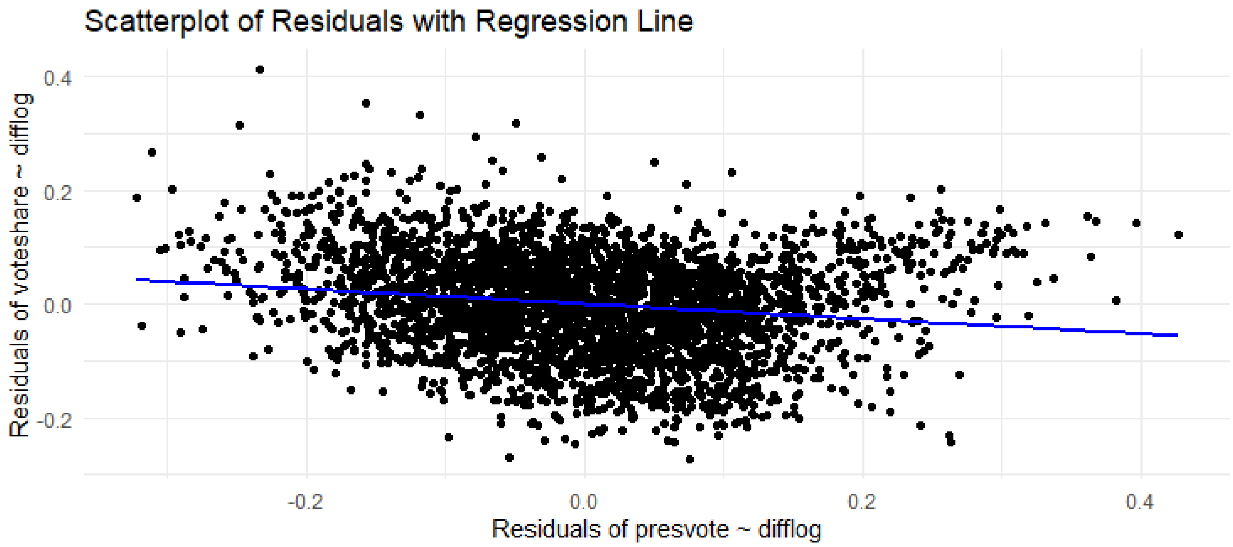
Table 4 shows the residuals of Table 1 (Difference in campaign spending and incumbent voteshare) regressed on the residuals of Table 2 (Differences in campaign spending and voteshare of presidential candidate). The positive coefficient of 0.257 is significant at a p-value of 0.0012, indicating that the unexplained variation in Table 1 is positively correlated with the unexplained correlatinon in Table 2. Meaning the variation within Table 1 and 2, which was not explained by the variation of their respective dependent varibales, is positively correlated with each other.

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

```r
library(ggplot2)

residuals_data <- data.frame(
  residuals_voteshare = residuals_voteshare,
  residuals_presvote = residuals_presvote)

ggplot(residuals_data, aes(x = residuals_presvote, y = residuals_
   voteshare)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(x = "Residuals of presvote ~ difflog",
       y = "Residuals of voteshare ~ difflog",
       title = "Scatterplot of Residuals with Regression Line") +
  theme_minimal()
```



This scatterplot shows the residuals of Table 1 on the Y-axis and the residuals of Table 2 on the X-axis. However, contrary to the output of the regression conducted above, a slight negative correlation is visible in this scatterplot, as the regression line seems to have a negative slope. This scatterplot is possibly caused by a computational mistake of mine, which I could not deteced, although I ran the tests multiple times and double checked using the correct data.

3. Write the prediction equation.

$$\text{residuals\_voteshare} = -5.934 \times 10^{-18} + 0.2569 \times \text{residuals\_presvote} + \epsilon \qquad (5)$$

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`.

```
model_voteshare_combined <- lm(voteshare ~ difflog + presvote, data = 
    incumbent_subset)

summary(model_voteshare_combined)

stargazer(model_voteshare_combined, type = "latex")
```

Table 5: Multivariate Model: Difflog and Presvote on Voteshare

	Dependent variable:
	voteshare
difflog	0.036***
	(0.001)
presvote	0.257***
	(0.012)
Constant	0.449***
	(0.006)
Observations	3,193
R^2	0.450
Adjusted R^2	0.449
Residual Std. Error	0.073 (df = 3190)
F Statistic	1,302.947*** (df = 2; 3190)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5 shows a multivariate model, where the two independent variables difflog (Differences in campaign spending) and presvote (voteshare of the incumbent presidential candidate) are regressed on voteshare (voteshare of the incumbent). As in previous models, the correlation coefficients of difflog and presvote are positive and significant with 0.036 at 0.001 a alpha level, and 0.257 at a 0.012 alpha level respectively. However, again the coefficient of difflog cannot be directly interpreted as this varibale is logged, and the coefficient of presvote can be interpreted following the standard interpretation of regression outputs.

2. Write the prediction equation.

$$\text{voteshare} = 0.4486442 + 0.0355431 \times \text{difflog} + 0.2568770 \times \text{presvote} + \epsilon \tag{6}$$

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

In the outputs of Question 4 and Question 5, the coefficient for residual presvote and presvote is identical (approximately 0.257). This similarity occurs because, in both cases, we are capturing the effect of presvote on voteshare while accounting for difflog. In Question 4, this effect is isolated by using residuals (the unexplained variation in voteshare and presvote after controlling for difflog), while in Question 5, it is calculated through a direct multiple regression. According to the Frisch-Waugh-Lovell theorem, these approaches yield the same coefficient for presvote since they both effectively control for difflog, allowing us to observe the unique contribution of presvote to voteshare.