Problem Set 3

Tolga Bag - 23371290

Due: November 19, 2022

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 19, 2023. 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`. I am checking the table and running the regression

```
head(inc.sub)
summary(inc.sub)
str(inc.sub) #I inspected the dataset.
regmod_a <- lm(voteshare ~ difflog, data = inc.sub) #I run the regression
    model
```

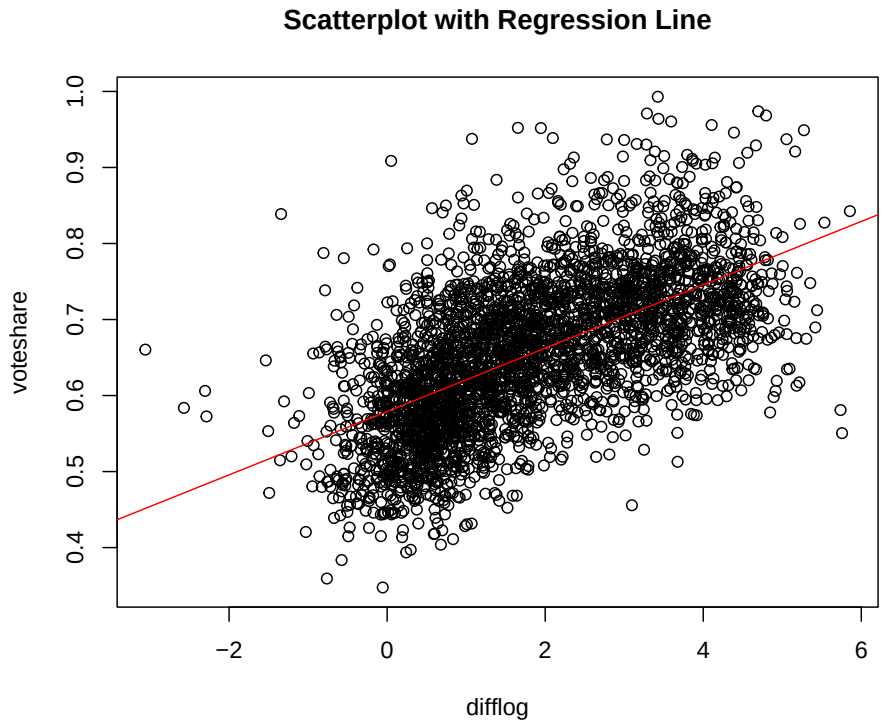
   voteshare is the dependent or outcome variable and difflog is the independent or explanatory variable. My null hypothesis is there is no relationship between voteshare and difflog.

```r
summary(regmod_a) #p value is 0.0000000000000002, so I can reject the
    null
```

p value is 0.0000000000000002, so I can reject the null hypothesis. Slope is 0.041666. In other words, for each unit change in difflog, voteshare is expected to increase by that amount. And it is a statistically significant result per the p value.

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

```r
#I prepare a scatter plot with voteshare as the dependent variable
    against
#difflog as the independent variable.
plot(inc.sub$difflog, inc.sub$voteshare,
     xlab = "difflog",
     ylab = "voteshare",
     main = "Scatterplot with Regression Line")
#I use the abline function to add the regression line per https://www.
    geeksforgeeks.org/adding-straight-lines-to-a-plot-in-r-programming-
    abline-function/
abline(regmod_a, col = "red") #I made it red to make it clear.
#as many observations are close to the regression line, it looks like a
    strong
#relationship and strong linearity.
```



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

```
residuals_a <- residuals(regmod_a)
summary(residuals_a) #my residuals are on this object. The symmetry
#between the minimum and maximum suggest a normally distributed model.
```

4. Write the prediction equation.

```
#I need to obtain the coefficients for the prediction equation.
coef_a <- coef(regmod_a)
#I need the intercept and slope of the coefficients to make the equation.
intercept_a <- coef_a[1]
slope_a <- coef_a[2]
#I make some research to find the code for the prediction. This provides
    detailed
#information: https://www.dataquest.io/blog/statistical-learning-for-
    predictive-modeling-r/
#I also used chatgpt to fox the errors I came across.
prediction_equation <- function(difflog) {
  voteshare_prediction = intercept_a + slope_a * difflog
  return(voteshare_prediction)
}
#I use the equation to check the voteshare value for a difflog value of
    1.42
prediction_equation(1.42)
#it predicts a voteshare value of 0.6381969. The prediction equation
    works.
```

Question 2

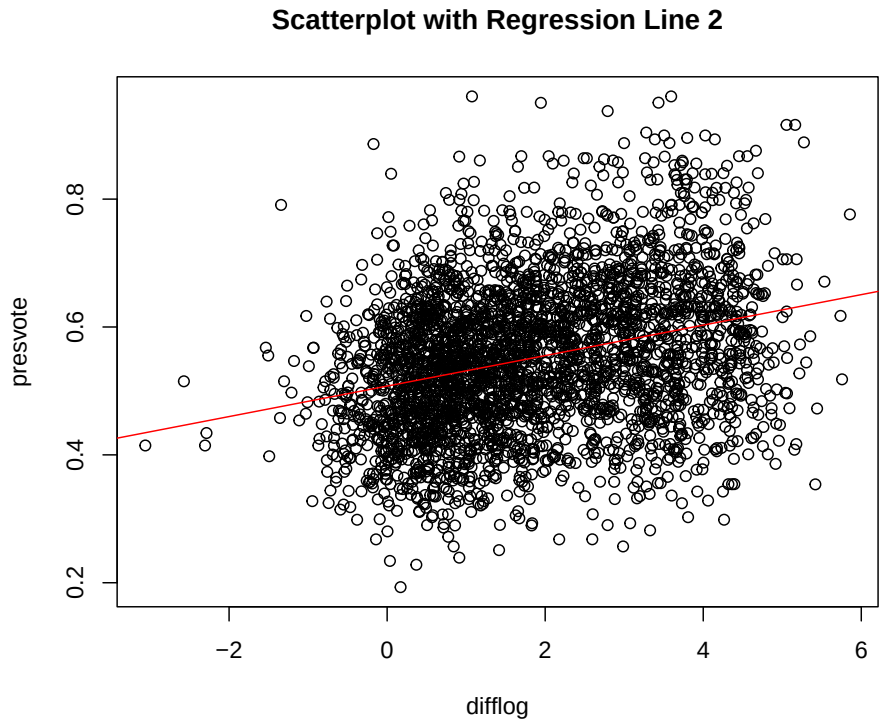
We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

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

```
regmod_b <- lm(presvote ~ difflog, data = inc.sub) #I run the regression
    model
#presvote is the dependent or outcome variable and difflog is the
    independent or
#explanatory variable.
summary(regmod_b) #p value is 0.0000000000000002, so I can reject the
    null
#hypothesis. Slope is 0.023837. In other words, for each unit change in
    difflog,
#voteshare is expected to increase by that amount. And it is a
    statistically
#significant result per the p value.
```

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

```
plot(inc.sub$difflog, inc.sub$presvote,
     xlab = "difflog",
     ylab = "presvote",
     main = "Scatterplot with Regression Line 2")
#I make it just as the first one
abline(regmod_b, col = "red") #Again, I made it red to make it clear.
```



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

```
residuals_b <- residuals(regmod_b)
summary(residuals_b) #my residuals are on this object.
```

4. Write the prediction equation.

```
#I need to obtain the coefficients for the prediction equation.
coef_b <- coef(regmod_b)
#I need the intercept and slope of the coefficients to make the equation.
intercept_b <- coef_b[1]
slope_b <- coef_b[2]
prediction_equation_b <- function(difflog) {
  voteshare_prediction_b = intercept_b + slope_b * difflog
  return(voteshare_prediction_b)
}
```

```r
#I use the equation to check the voteshare value for a difflog value of 1.42
prediction_equation_b(1.42)
#it predicts a voteshare value of 0.5414322. The prediction equation works.
```

Question 3

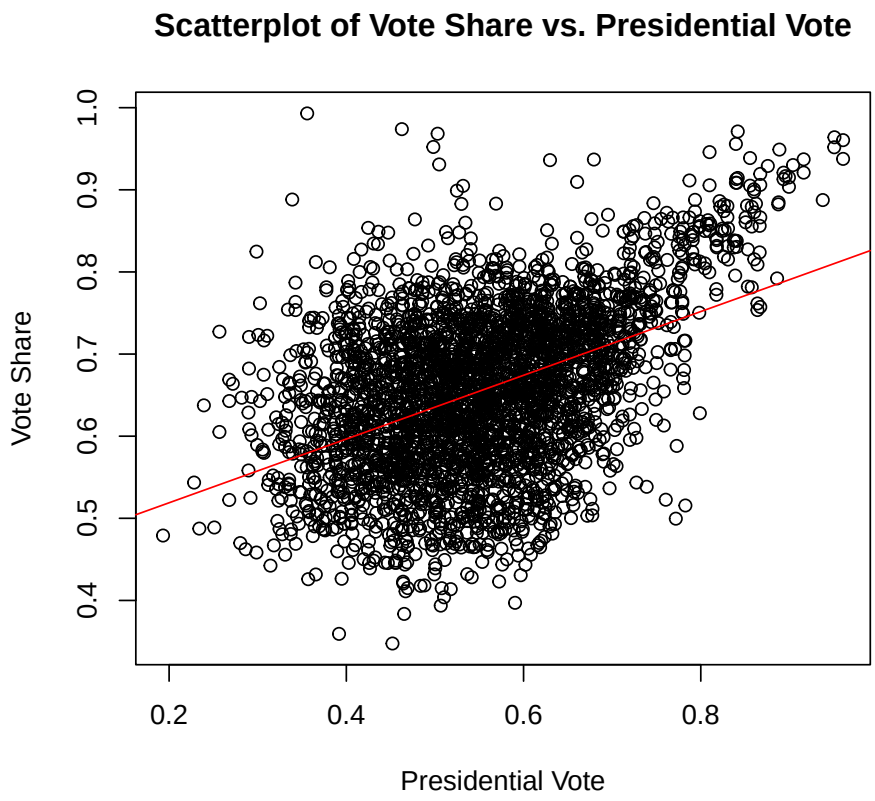
We are interested in knowing how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success.

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

```r
regmod_c <- lm(voteshare ~ presvote, data = inc.sub)
summary(regmod_c)
```

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

```r
plot(inc.sub$presvote, inc.sub$voteshare,
     xlab = "Presidential Vote",
     ylab = "Vote Share",
     main = "Scatterplot of Vote Share vs. Presidential Vote")
abline(regmod_c, col = "red")
```



3. Write the prediction equation.

```r
#Again, #I need to obtain the coefficients for the prediction equation.
coef_c <- coef(regmod_c)
#I need the intercept and slope of the coefficients to make the equation.
intercept_c <- coef_c[1]
slope_c <- coef_c[2]
prediction_equation_c <- function(difflog) {
  voteshare_prediction_c = intercept_c + slope_c * difflog
  return(voteshare_prediction_c)
}
#I use the equation to check the voteshare value for a difflog value of
    1.42
prediction_equation_c(1.42) #it returns 0.99 and it works.
```

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.

```
regmod_d <- lm(residuals_a ~ residuals_b, data = inc.sub)
summary(regmod_d)
```

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

```
plot(residuals_a, residuals_b,
     xlab = "Residuals_a from Q1",
     ylab = "Residuals_b from Q2",
     main = "Scatterplot of the two residuals")
abline(regmod_d, col = "red")
```

3. Write the prediction equation.

```
#Again, #I need to obtain the coefficients for the prediction equation.
```

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`.

```
regmod_e <- lm(voteshare ~ difflog + presvote, data = inc.sub) #this is a
    multi
#variate regression, so I edit the code accordingly.
summary(regmod_e) #per p values this is a good fit of a model.
```

2. Write the prediction equation.

```
#again, I start with the coefficients.
coef_e <- coef(regmod_e)
#I need to assign the intercept and slopes to variables:
intercept_e <- coef_e[1]
slopee_difflog <- coef_e[2]
slopee_presvote <- coef_e[3]

prediction_equation_e <- function(difflog, presvote) {
  voteshare_prediction_e = intercept_e + slopee_difflog * difflog +
    slopee_presvote * presvote
  return(voteshare_prediction_e)
}
#I test my model. The first input is difflog and the second is presvote
    to predict
#voteshare
prediction_equation_e(1.42, 0.44)
#it gives me 0.61. It works.
```

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 residuals are exactly the same. I think it is because both regressions are checking the effect of difflog. The one in Question 4 does it indirectly through the residuals of models that already accounted difflog, while the multi variate model does it directly.