Problem Set 3

Applied Stats/Quant Methods 1

Due: November 20, 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 20, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

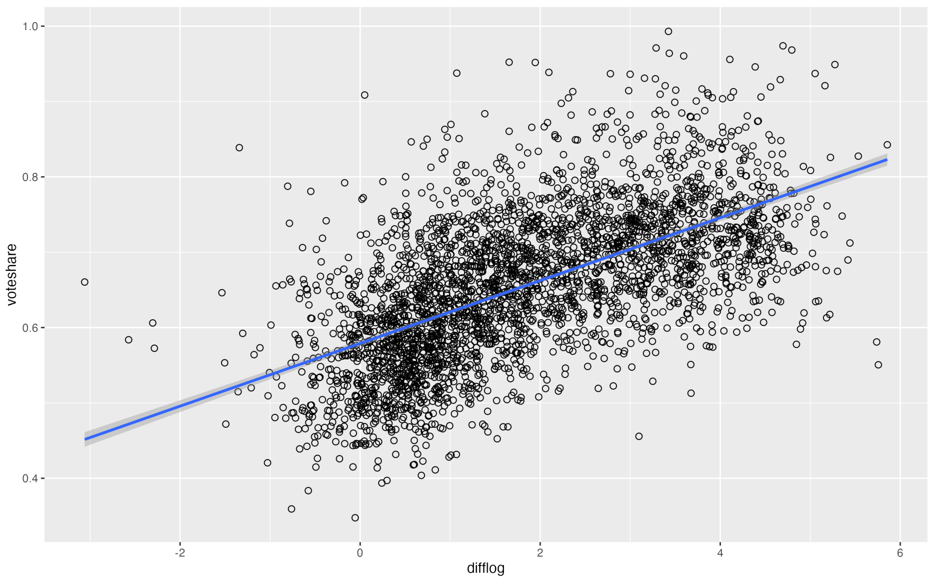
In this problem set, you will run several regressions and create an add variable plot (see the lecture slides) in R using the incumbents_subset.csv dataset. Include all of your code.

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.

```
# run regression
spend_lm <- lm(voteshare~difflog , data = data1)
summary(spend_lm) # small p-value , low se , 1 unit increase in
# difflog = 0.041666 increase in incumbent's voteshare
```



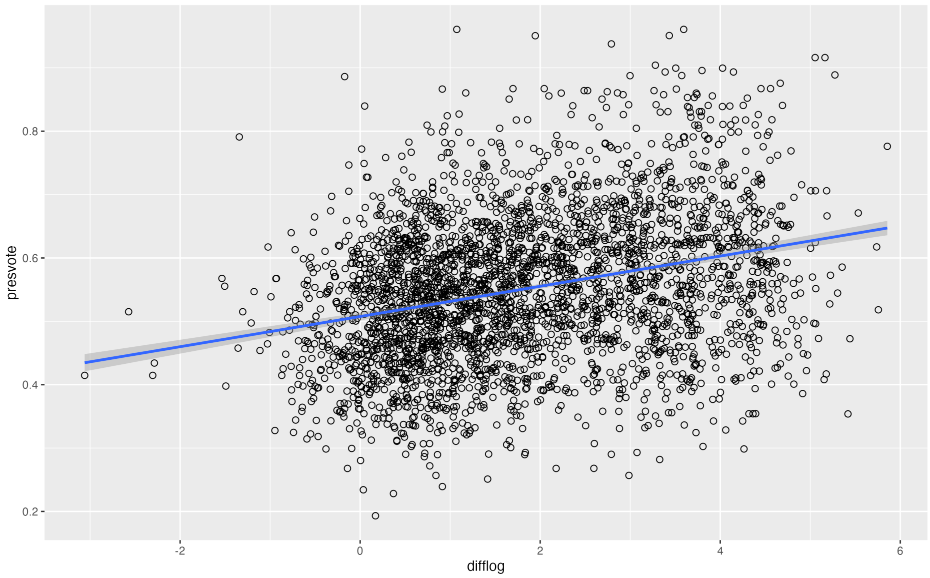
2. Make a scatterplot of the two variables and add the regression line.

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

```r
resids <- residuals(spend_lm)
```

4. Write the prediction equation.

   Y = 0.579031 + 0.041666X



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`.

```
plot(voteshare~difflog, data=data1) # very roughly linear

# run regression
spend_lm <- lm(voteshare~difflog, data = data1)
```

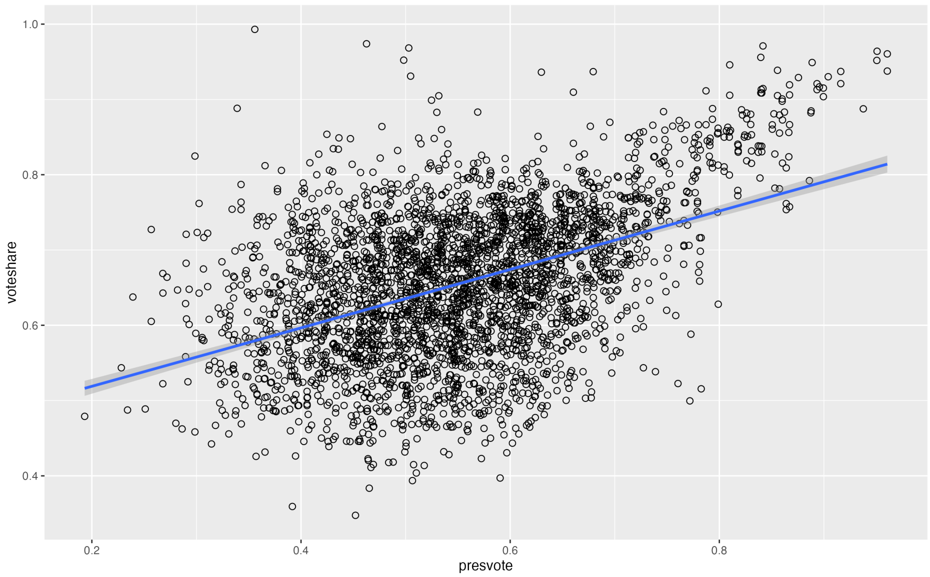
2. Make a scatterplot of the two variables and add the regression line.

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

```
resids2 <- residuals(spend_pres_lm)
```

4. Write the prediction equation.

   Y = 0.507583 + 0.023837X



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.

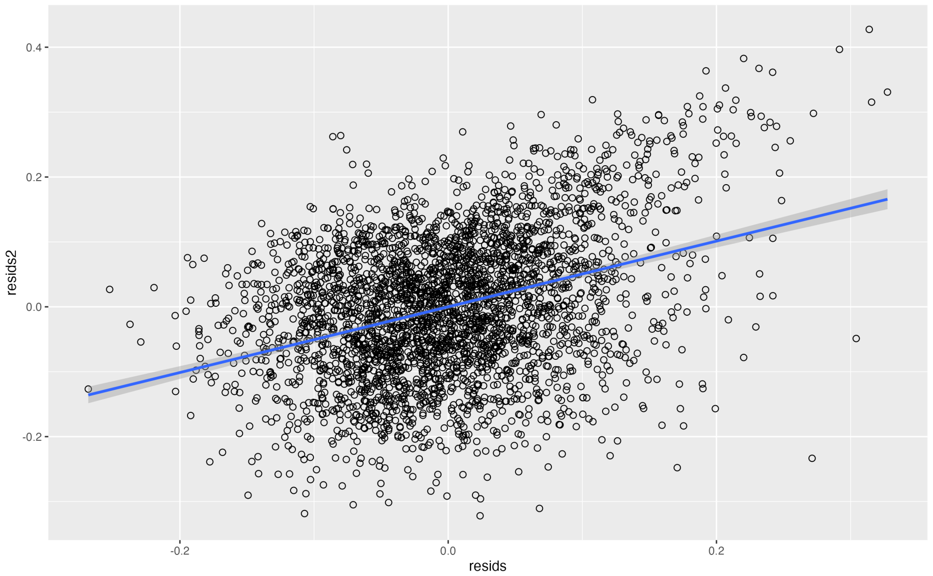
1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`.

```
# run regression
votesh_presv_lm <- lm(voteshare~presvote, data = data1)
summary(votesh_presv_lm) # small p-value, low se, 1 unit increase in
# presvote = 0.388018 increase in incumbent's voteshare
```

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

3. Write the prediction equation.

   Y = 0.441330 + 0.388018X



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.

```
residd_lm <- lm(resids~resids2)
summary(residd_lm) # small increase in resids 2 for 1 in resids
```

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

3. Write the prediction equation.

   Y = -4.860e-18 + 2.569e-01X

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`.

```
mrmodel <- lm(voteshare ~ difflog + presvote, data = data1)
summary(mrmodel)
```

2. Write the prediction equation.

   Y = 0.4486442 + 0.0355431X1 + 0.0117637X2

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 outputs include the statistics on the values of the residuals, including the interquartile ranges and their centring roughly around 0, and the residual standard errors (almost identical). The multiple regression model is regressed on the same variables which produced the linear models with the residuals used in Q4. The results of Q4 describe how much of the variation in voteshare unexplained by the linear model in Q1 can be explained by the variation in presvote which is also unexplained by the model in Q2. In other words, the residuals plotted against each other are telling us, after removing the effect of difflog (i.e. the linear models) from presvote and voteshare, if there's a relationship between voteshare and presvote. Hence, the multiple regression model, which aims to indicate how much variation in the outcome variable (voteshare) can be explained by variation in the input variables, will show some of the same results, indicating the interaction effects of the input variables.