Problem Set 3

Applied Stats/Quant Methods 1/Yucheng Wang

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.

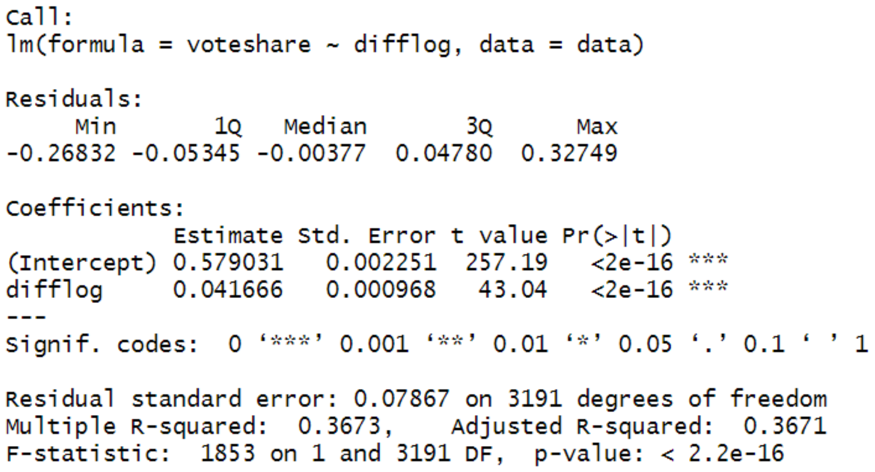
In this problem set, you will run several regressions and create an add variable plot (see the lecture slides) in `R` using the `incumbents_subset.csv` dataset. Include all of your code.

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`.

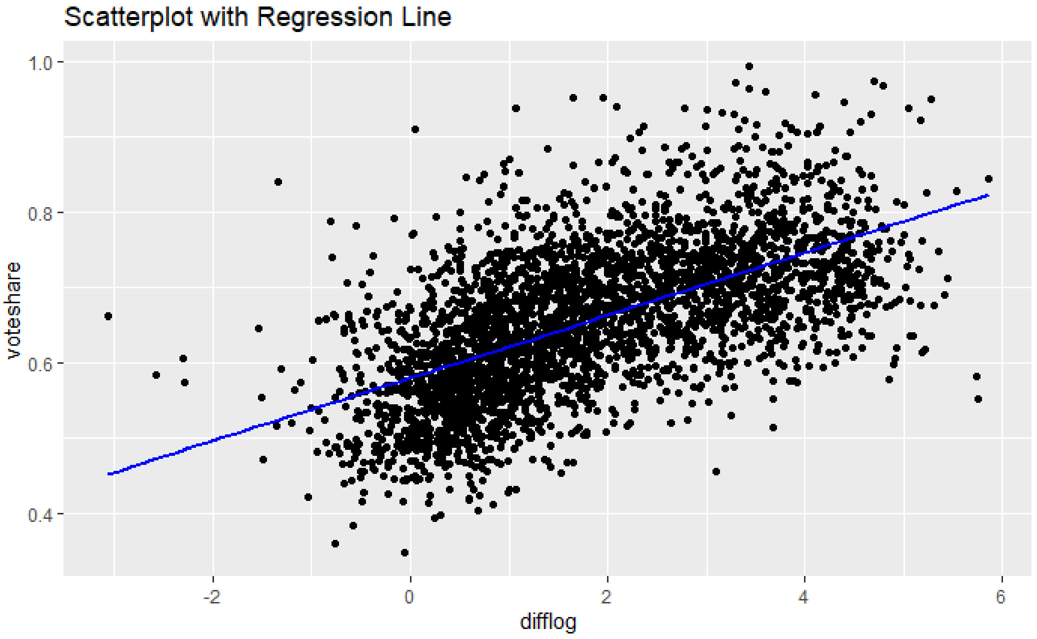
```r
#read database from a 'csv'file
data <- read.csv("incumbents_subset.csv")
#draw a regression result from voteshare and difflog, use the imported database
regression_result <- lm(voteshare ~ difflog, data = data)
#use the function 'summary'to see the main characters of of this regression result
summary(regression_result)
```



```
Call:
lm(formula = voteshare ~ difflog, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26832 -0.05345 -0.00377  0.04780  0.32749 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.579031   0.002251  257.19   <2e-16 ***
difflog     0.041666   0.000968   43.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07867 on 3191 degrees of freedom
Multiple R-squared:  0.3673,	Adjusted R-squared:  0.3671 
F-statistic:  1853 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```
#library the package'ggplot2'
library(ggplot2)
#use ggplot function to draw the plot of difflog and voteshare, and use the function geom_smooth to find the regression line
ggplot(data, aes(x = difflog, y = voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(x = "difflog", y = "voteshare") +
  ggtitle("Scatterplot with Regression Line")  ## Add a title
```



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

```r
## use resid function to save the residuals of the model in the 'residuals_model'object
residuals_model <- resid(regression_result)
```

4. Write the prediction equation.

```r
##write the prediction equation according to the summary of our regression result

\[ \text{voteshare} = 0.579031 + 0.041666 \times \text{difflog} \]
```

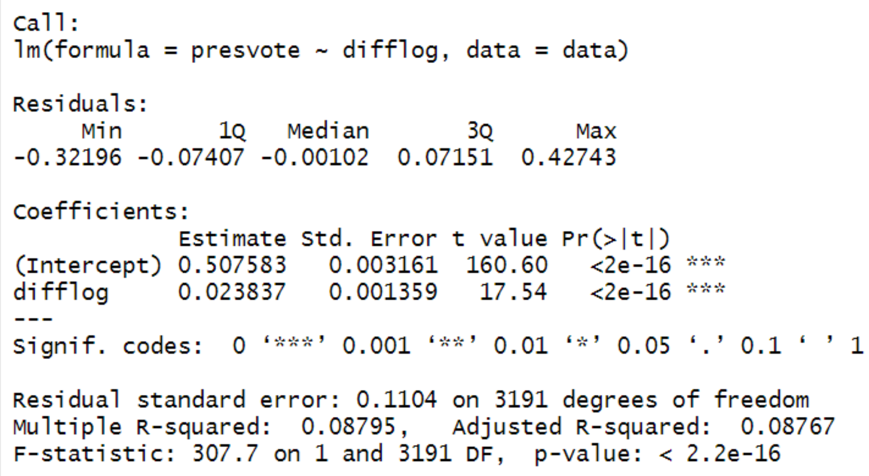
$$\text{voteshare} = 0.579031 + 0.041666 \times \text{difflog}$$

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`.

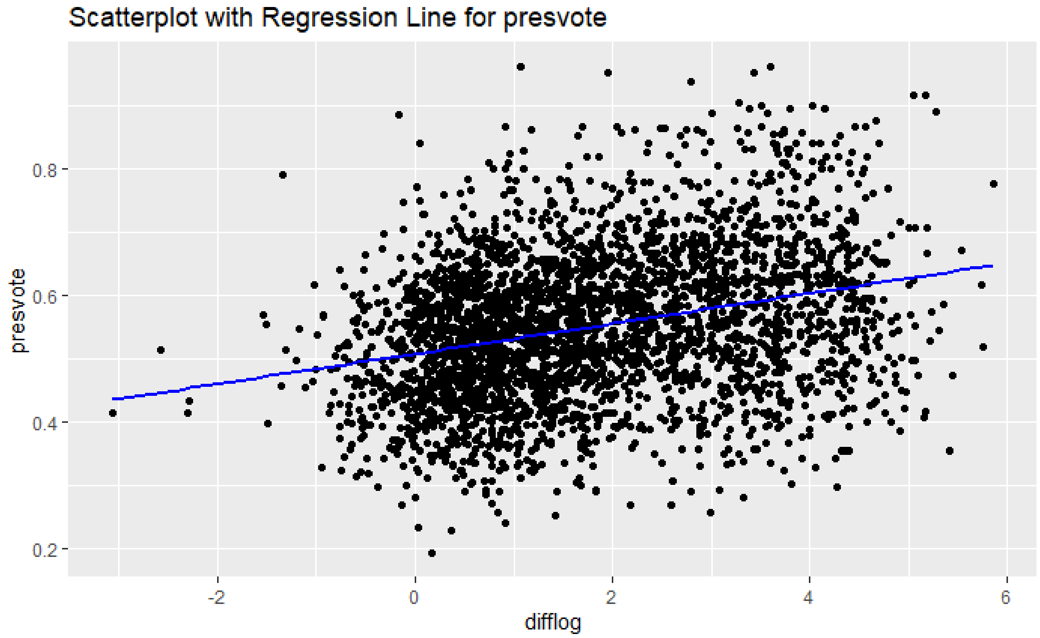
```r
#run a regression result about presvote and difflog, use the imported database as the data
regression_result_presvote <- lm(presvote ~ difflog, data = data)
#use the function summary to see the characters of the result
summary(regression_result_presvote)
```



```
Call:
lm(formula = presvote ~ difflog, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32196 -0.07407 -0.00102  0.07151  0.42743 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.507583   0.003161  160.60   <2e-16 ***
difflog     0.023837   0.001359   17.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1104 on 3191 degrees of freedom
Multiple R-squared:  0.08795,	Adjusted R-squared:  0.08767 
F-statistic: 307.7 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```r
#use the ggplot function to draw a plot of difflog and presvote, and use geom_smooth function to get the regression line
 ggplot(data, aes(x = difflog, y = presvote)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(x = "difflog", y = "presvote") +
   ggtitle("Scatterplot with Regression Line for presvote") ## Add the title
```

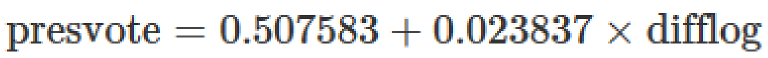


3. Save the residuals of the model in a separate object.

```
#save the residuals of the model into the residuals_presvote as a separate object
residuals_presvote <- resid(regression_result_presvote)
```

4. Write the prediction equation.

```
##write the prediction equation according to the summary of our regression result
\[\text{presvote} = 0.507583 + 0.023837 \times \text{difflog} \]
```



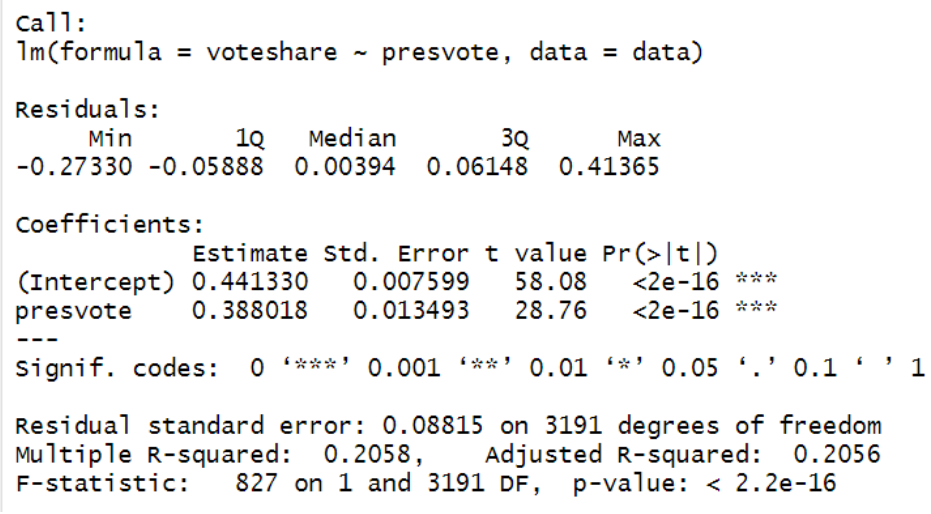
$$\text{presvote} = 0.507583 + 0.023837 \times \text{difflog}$$

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`.

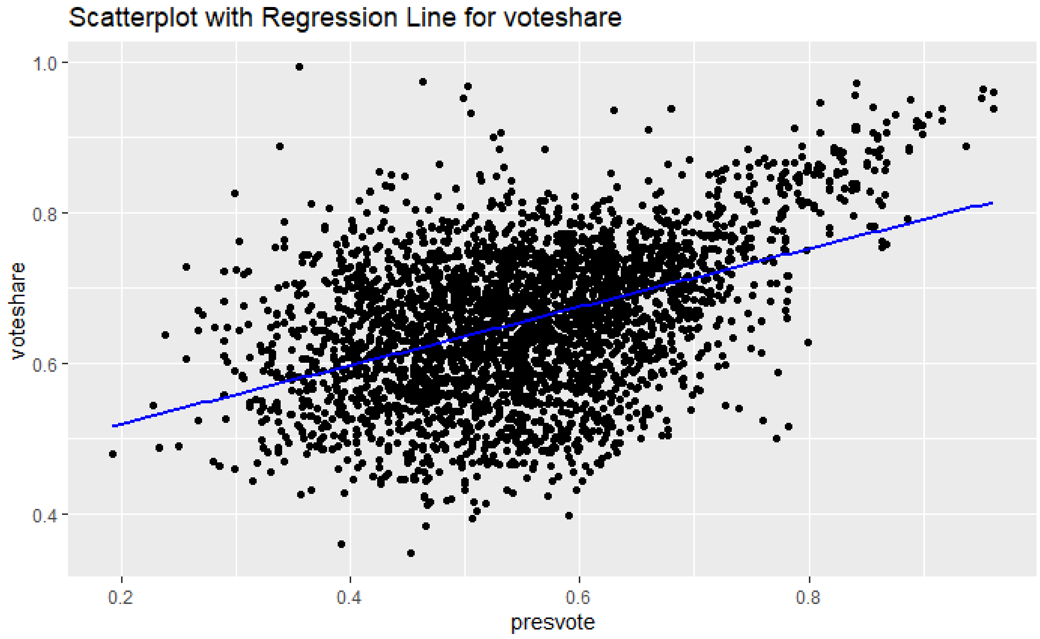
```r
#run a regression by the lm function to get the regression of voteshare and presvote, use the imported database as our data
regression_result_voteshare <- lm(voteshare ~ presvote, data = data)
#use the summary function to get the characters of our regression result
summary(regression_result_voteshare)
```



```
Call:
lm(formula = voteshare ~ presvote, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27330 -0.05888  0.00394  0.06148  0.41365 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.441330   0.007599   58.08   <2e-16 ***
presvote    0.388018   0.013493   28.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08815 on 3191 degrees of freedom
Multiple R-squared:  0.2058,	Adjusted R-squared:  0.2056 
F-statistic:   827 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```r
#use ggplot function to draw the plot of our regression, and use geom_smooth to get the regresion line for voteshare
ggplot(data, aes(x = presvote, y = voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(x = "presvote", y = "voteshare") +
  ggtitle("Scatterplot with Regression Line for voteshare")##add the title
```



3. Write the prediction equation.

```
##write the prediction equation according to the summary of our
   regression result
\[\text{voteshare} = 0.441330 + 0.388018 \times \text{presvote} \]
```

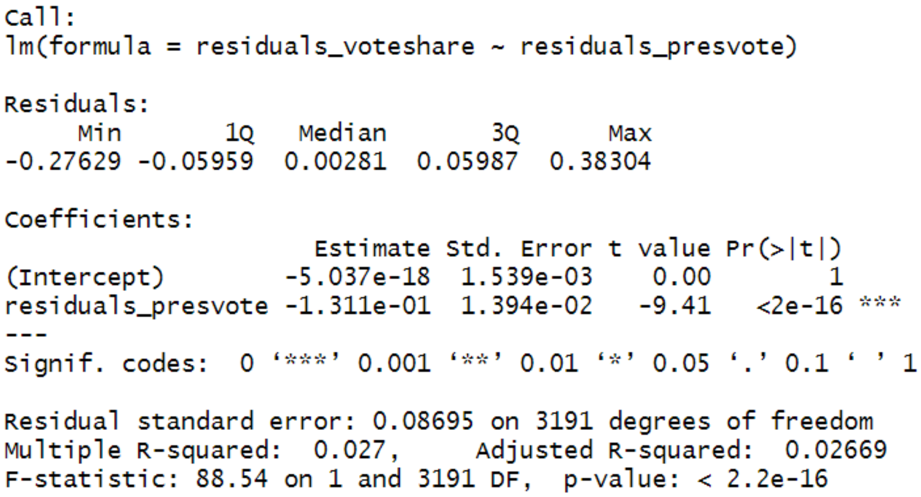
$$\text{voteshare} = 0.441330 + 0.388018 \times \text{presvote}$$

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.

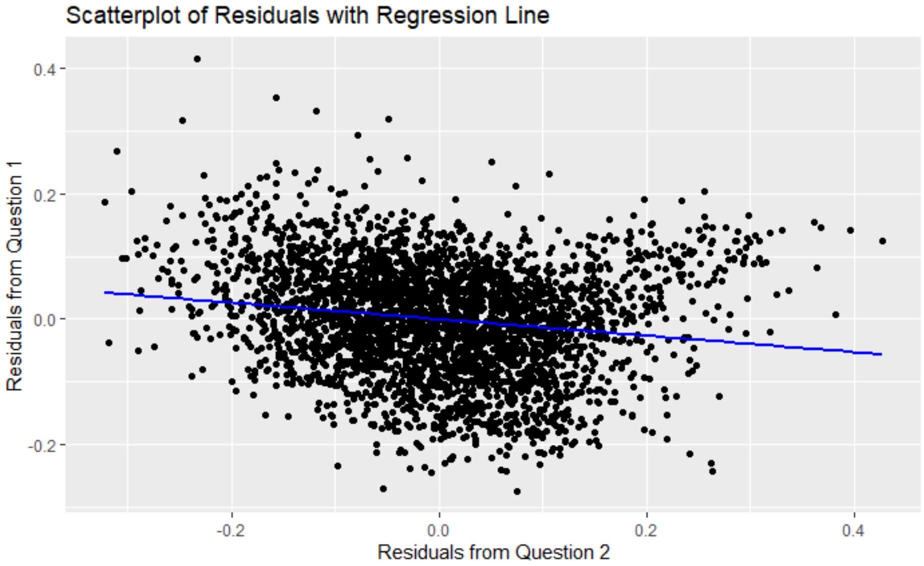
```
#use resid function to save the residuals of regression_result_voteshare
residuals_voteshare <- resid(regression_result_voteshare)
#run a regression of residuals_voteshare and resuals_presvote
residuals_regression <- lm(residuals_voteshare ~ residuals_presvote)
#use summary function to get the characters of this regression result
summary(residuals_regression)
```



```
Call:
lm(formula = residuals_voteshare ~ residuals_presvote)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27629 -0.05959  0.00281  0.05987  0.38304 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -5.037e-18  1.539e-03    0.00        1    
residuals_presvote -1.311e-01  1.394e-02   -9.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08695 on 3191 degrees of freedom
Multiple R-squared:  0.027,	Adjusted R-squared:  0.02669 
F-statistic: 88.54 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```r
#use gglot to draw the scatterplot, we use data.frame with residuals_voteshare and residuals_presvote,and use geom_smooth to get our regression line
ggplot(data.frame(Residuals_voteshare = residuals_voteshare, Residuals_presvote = residuals_presvote),
         aes(x = Residuals_presvote, y = Residuals_voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(x = "Residuals from Question 2", y = "Residuals from Question 1") +
  ggtitle("Scatterplot of Residuals with Regression Line")##add a title
```



3. Write the prediction equation.

```
##write the prediction equation according to the summary of our regression result
\[residuals\_voteshare = -0.1311 \times residuals\_presvote \]
```

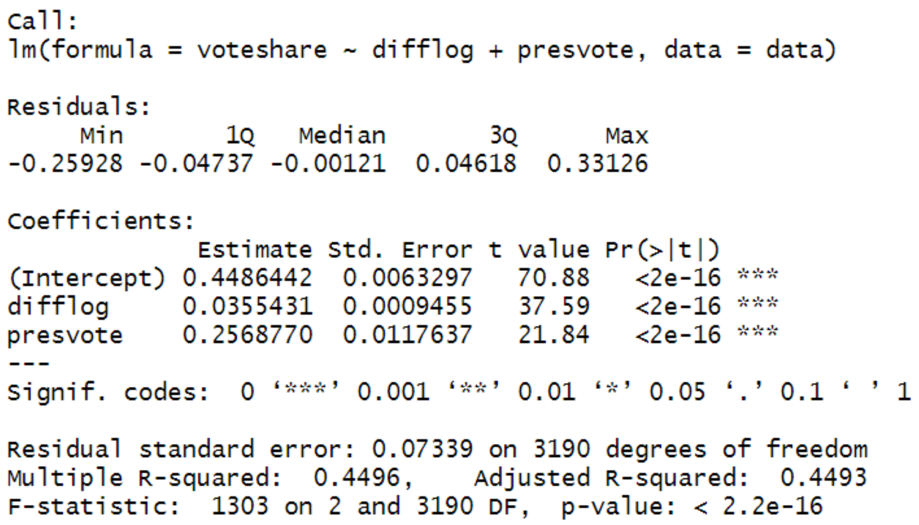
$$residuals\_voteshare = -0.1311 \times residuals\_presvote$$

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`.

```
#use the lm function to run a regression where the outcome variable is the incumbents voteshare and the explanatory variables are difflog and presvote
regression_result_combined <- lm(voteshare ~ difflog + presvote, data = data)
#use the summary function to get the characters of this regression result
summary(regression_result_combined)
```



```
Call:
lm(formula = voteshare ~ difflog + presvote, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25928 -0.04737 -0.00121  0.04618  0.33126 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.4486442  0.0063297   70.88   <2e-16 ***
difflog     0.0355431  0.0009455   37.59   <2e-16 ***
presvote    0.2568770  0.0117637   21.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07339 on 3190 degrees of freedom
Multiple R-squared:  0.4496,    Adjusted R-squared:  0.4493 
F-statistic:  1303 on 2 and 3190 DF,  p-value: < 2.2e-16
```

2. Write the prediction equation.

```
##write the prediction equation according to the summary of our regression result
\[ \text{voteshare} = 0.4486442 + 0.0355431 \times \text{difflog} + 0.2568770 \times \text{presvote} \]
```

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

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

```
In the output of the regression where the outcome variable is the incumbent's ‘voteshare‘ and the explanatory variables are ‘difflog‘ and ‘presvote‘, the part that is identical to the output in Question 4 is the residuals from ‘presvote‘ regressed on ‘difflog‘. Specifically, the coefficient for ‘difflog‘ in the regression output is the same as the coefficient for ‘difflog‘ in Question 4.

In Question 4, we performed a regression of the residuals from the model where ‘presvote‘ is the outcome variable and ‘difflog‘ is the explanatory variable. This means that the residuals from the relationship between ‘presvote‘ and ‘difflog‘ were calculated and then regressed on the residuals from the relationship between ‘voteshare‘ and ‘presvote‘.

In Question 5, we are directly regressing the incumbent's ‘voteshare‘ on both ‘difflog‘ and ‘presvote‘. However, the coefficient for ‘difflog‘ in this regression is the same as the coefficient for ‘difflog‘ in Question 4 because it represents the relationship between ‘difflog‘ and the dependent variable in both cases.

The reason for this similarity is that the coefficient for ‘difflog‘ captures the effect of ‘difflog‘ on the dependent variable, and this effect remains consistent across different models. In both Question 4 and Question 5, the coefficient for ‘difflog‘ quantifies how a one-unit change in ‘difflog‘ is associated with a change in the dependent variable (in this case, residuals from the dependent variable). So, the similarity arises because we are examining the same relationship in both cases, even though the context and models are different.
```