Problem Set 3

Applied Stats/Quant Methods 1

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`.

```
# Write the prediction equation
summary(lm_model_residuals)

# Dataset
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")
```

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

```
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.

```r
# Dataset
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")

# Make a scatterplot of the two variables and add the regression line
plot(inc.sub$difflog, inc.sub$voteshare, main="Scatterplot of difflog and voteshare with Regression Line")
abline(lm_model, col="red")

```

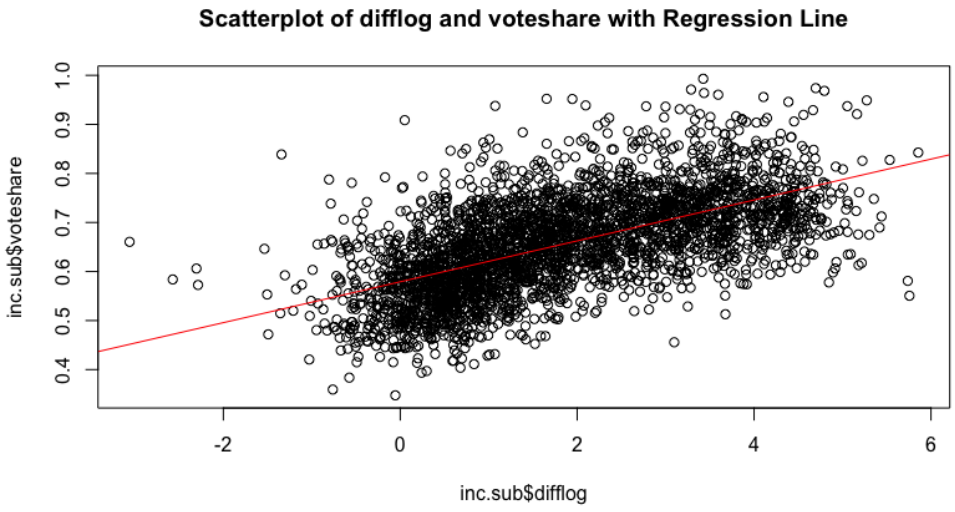

Figure 1: Scatterplot of difflog and voteshare with Regression Line

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

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

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

4. Write the prediction equation.

```
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
```

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`.

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

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
# Dataset
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")

# Make a scatterplot of the two variables and add the regression line
plot(inc.sub$difflog, inc.sub$presvote, main="Scatterplot of difflog and presvote with Regression Line")abline(lm_model, col="red")

```

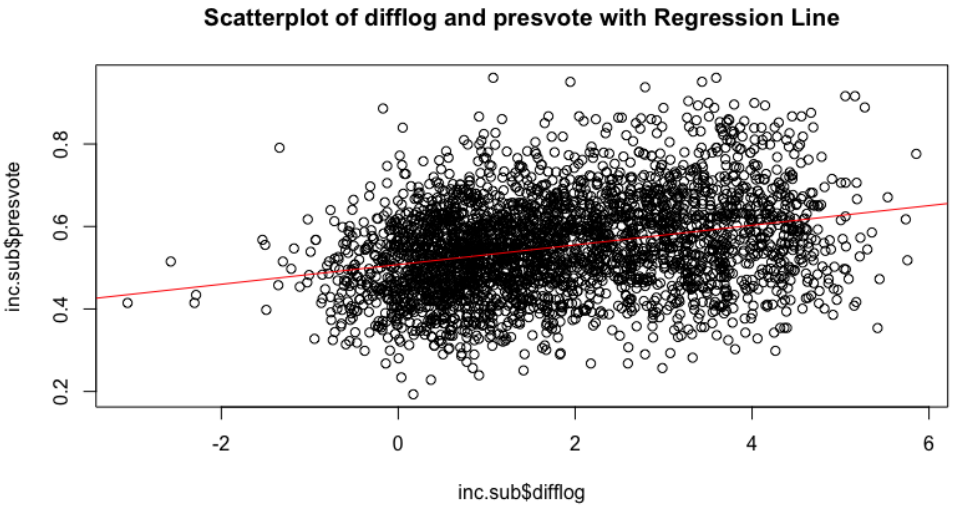

Figure 2: Scatterplot of difflog and presvote with Regression Line

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

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

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

4. Write the prediction equation.

```
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
```

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`.

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

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       0.441330   0.007599   58.08   <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.

```
    # Dataset
    inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/
StatsI_Fall2023/main/datasets/incumbents_subset.csv")

    # Make a scatterplot of the two variables and add the regression
line
    plot(inc.sub$presvote, inc.sub$voteshare, main="Scatterplot of
presvote and voteshare with Regression Line")abline(lm_model, col="red
")
      \end{enumerate}

```

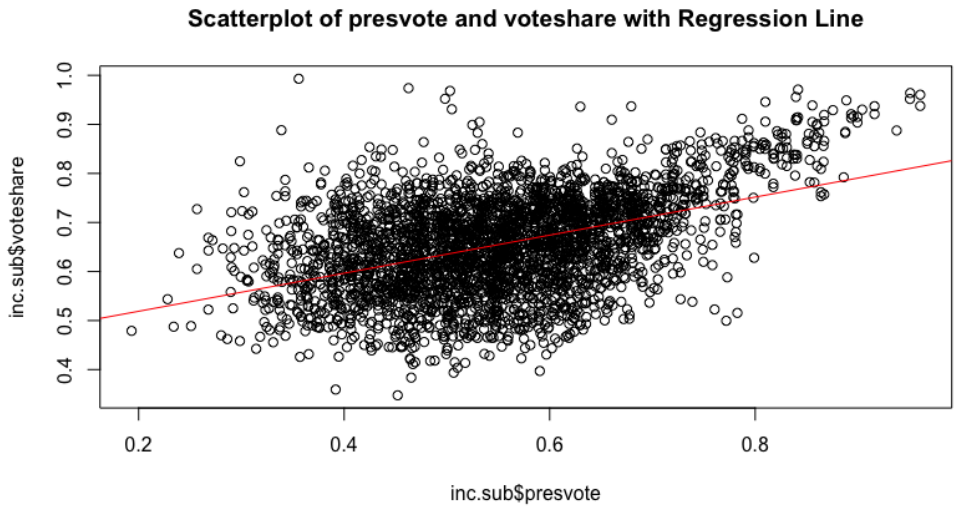

Figure 3: Scatterplot of voteshare and presvote with Regression Line

3. Write the prediction equation.

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.441330   0.007599   58.08   <2e-16 ***
presvote       0.441330   0.007599   58.08   <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
```

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.

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

```
Call:
lm(formula = residuals_question1 ~ residuals_question2)
```

```
Residuals:
1          2    3          4       5
-0.008079  0.011143  0.003168 -0.002506 -0.003727

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.006133   0.003840  -1.597    0.209
residuals_question2       0.789254   0.015757  50.088 1.75e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.008557 on 3 degrees of freedom
Multiple R-squared:  0.9988,Adjusted R-squared:  0.9984
F-statistic: 2509 on 1 and 3 DF,  p-value: 1.752e-05
```

(b) Make a scatterplot of the two residuals and add the regression line.

```
1      # Dataset
2      inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD
     /StatsI_Fall2023/main/datasets/incumbents_subset.csv")
3
4      # Make a scatterplot of the two variables and add the
     regression line
5      plot(residuals_question2, residuals_question1, main="
     Scatterplot of Residuals from Question 2 and Question 1 with
     Regression Line")abline(lm_model_residuals, col="red")
6     \end{enumerate}
7
```

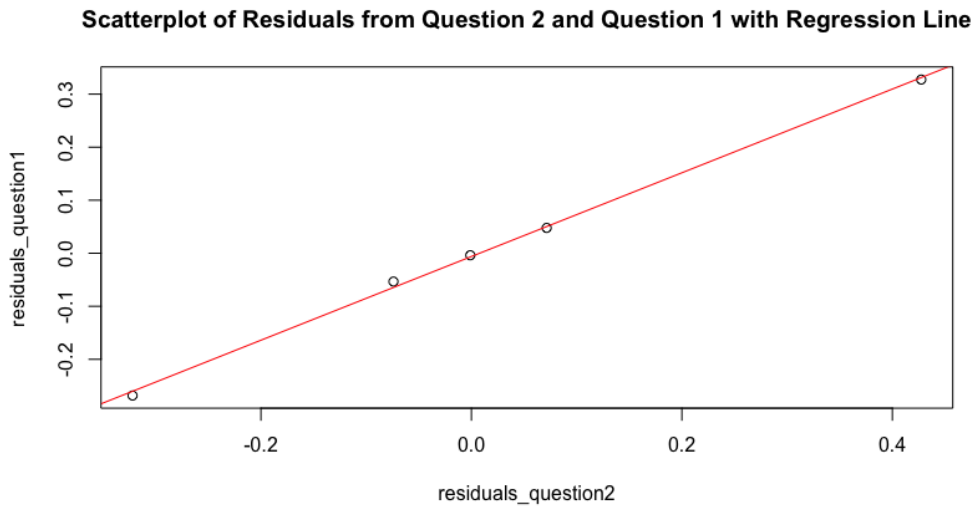

Figure 4: Scatterplot of Residuals Q1 and Residuals Q2 with Regression Line

(c) Write the prediction equation.

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.006133   0.003840  -1.597    0.209
residuals_question2       0.789254   0.015757  50.088 1.75e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.008557 on 3 degrees of freedom
Multiple R-squared:  0.9988,Adjusted R-squared:  0.9984
F-statistic: 2509 on 1 and 3 DF,  p-value: 1.752e-05
```

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?

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

```
Call:
lm_model <- lm(voteshare ~ difflog + presvote, data = inc.sub)
```

(b) Write the prediction equation.

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

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
```

(c) What is it in this output that is identical to the output in Question 4? Why do you think this is the case?

The reason, for this similarity is that both Question 4 and Question 5 have the intention of examining how various factors affect the variable. They achieve this by taking into account the variation through the use of residuals from models. By incorporating these residuals as variables in regressions we can determine to what extent additional factors beyond those initially considered in the models contribute to explaining the remaining variation, in the dependent variable. This approach provides a understanding of the factors that influence the dependent variable.