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.

```r
# read in data
incumbent <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/incumbents_subset.csv")
```

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
question1_regression <- lm(voteshare ~ difflog, data = incumbent)
print(question1_regression)
summary(question1_regression)
```

   Findings: Based on the model, for every one-unit increase in difflog there is an expected increase of about 0.04 in voteshare. This shows a positive relationship, as difflog increases, voteshare will generally increase as well.

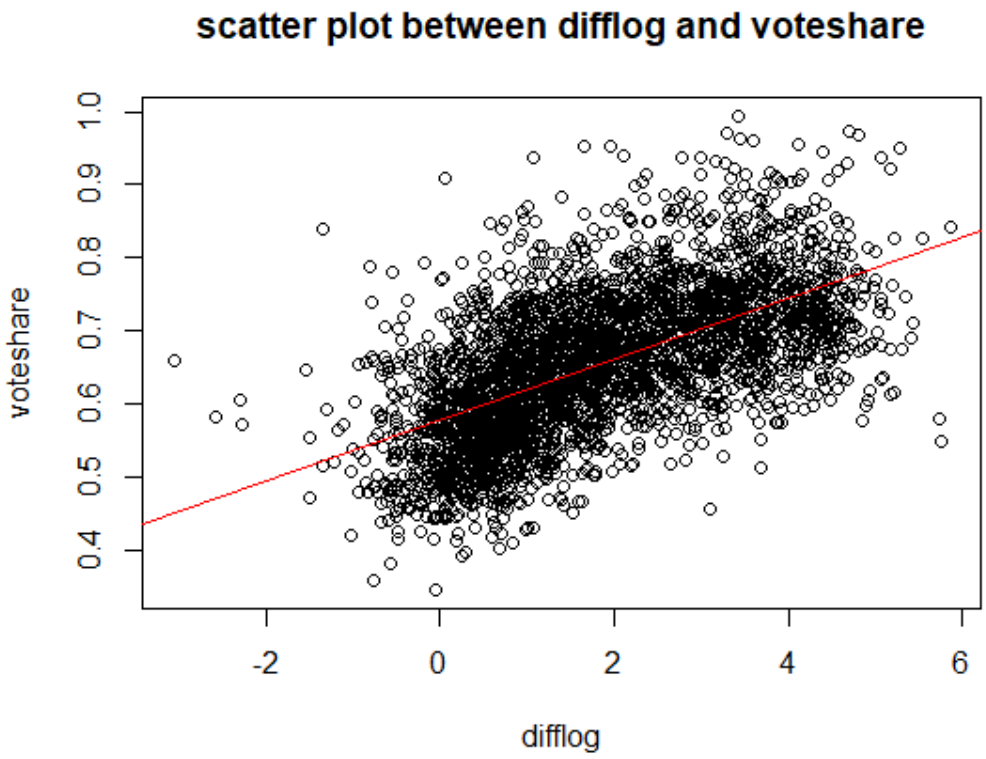
Figure 1: Regression between voteshare and difflog

```
> print(question1_regression)

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

Coefficients:
(Intercept)      difflog
    0.57903      0.04167

> summary(question1_regression)

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

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.

```
question1_scatter <- plot(incumbent$difflog, incumbent$voteshare,
                          xlab = "difflog",
                          ylab = "voteshare",
                          main = "scatter plot between difflog and
  voteshare") +
                    abline(question1_regression, col="red")
```

Figure 2: Question1 Scatter Plot



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

```
question1_residuals <- question1_regression$residuals
```

4. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{difflog}$$
$$voteshare = 0.579 + 0.042 \times \text{difflog}$$

```
intercept <- round(question1_regression$coefficients[1],3)
slope <- round(question1_regression$coefficients[2],3)
cat(intercept, "+", slope, "* 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
question2_regression <- lm(presvote ~ difflog, data = incumbent)
print(question2_regression)
summary(question2_regression)
```

Figure 3: Regression between presvote and difflog

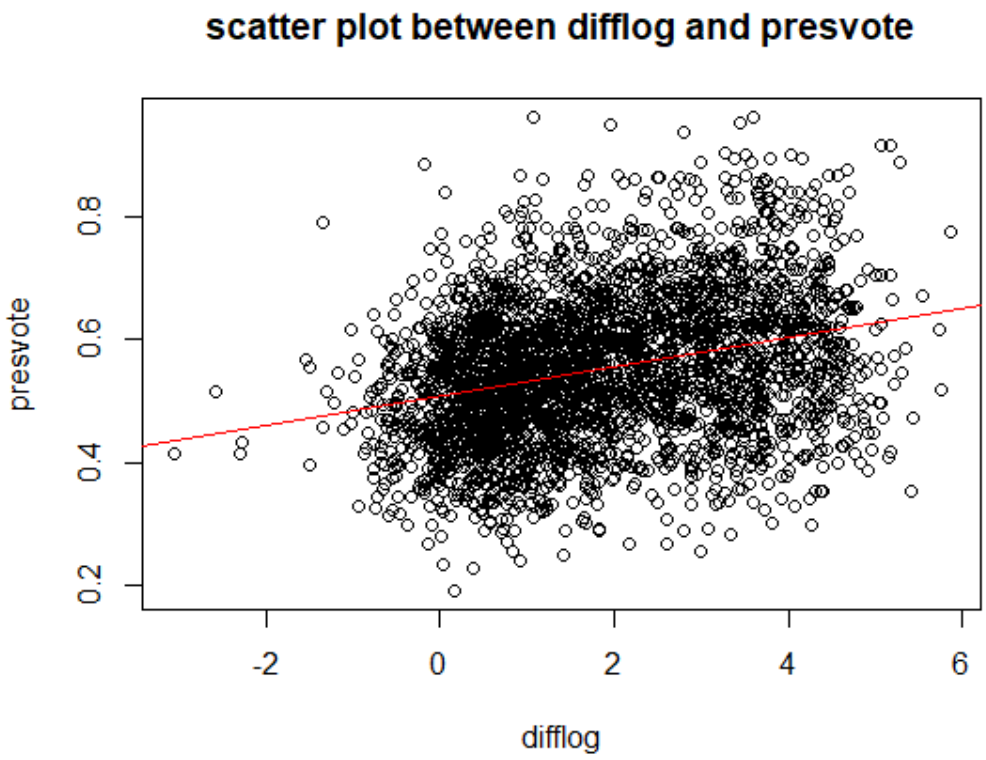
```
> print(question2_regression)

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

Coefficients:
(Intercept)      difflog
    0.50758      0.02384

> summary(question2_regression)

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

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
```

Findings: Based on the model, a one-unit increase in difflog will generally see a 0.02 increase in presvote. This is a positive relationship that shows when difflog rises presvote values tend to slightly increase.

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

```
question2_scatter <- plot(incumbent$difflog , incumbent$presvote ,
                          xlab = "difflog",
                          ylab = "presvote",
                          main = "scatter plot between difflog and
  presvote") +
 abline(question2_regression , col="red")
```

Figure 4: Question 2 Scatter Plot



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

```
question2_residuals <- question2_regression$residuals
```

4. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{difflog}$$
$$\textit{presvote} = 0.508 + 0.024 \times \text{difflog}$$

```
intercept2 <- round(question2_regression$coefficients[1],3)
slope2 <- round(question2_regression$coefficients[2],3)
cat(intercept2 , "+", slope2 , "* 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
question3_regression <- lm(voteshare ~ presvote, data = incumbent)
print(question3_regression)
summary(question3_regression)
```

Figure 5: Regression between presvote and voteshare

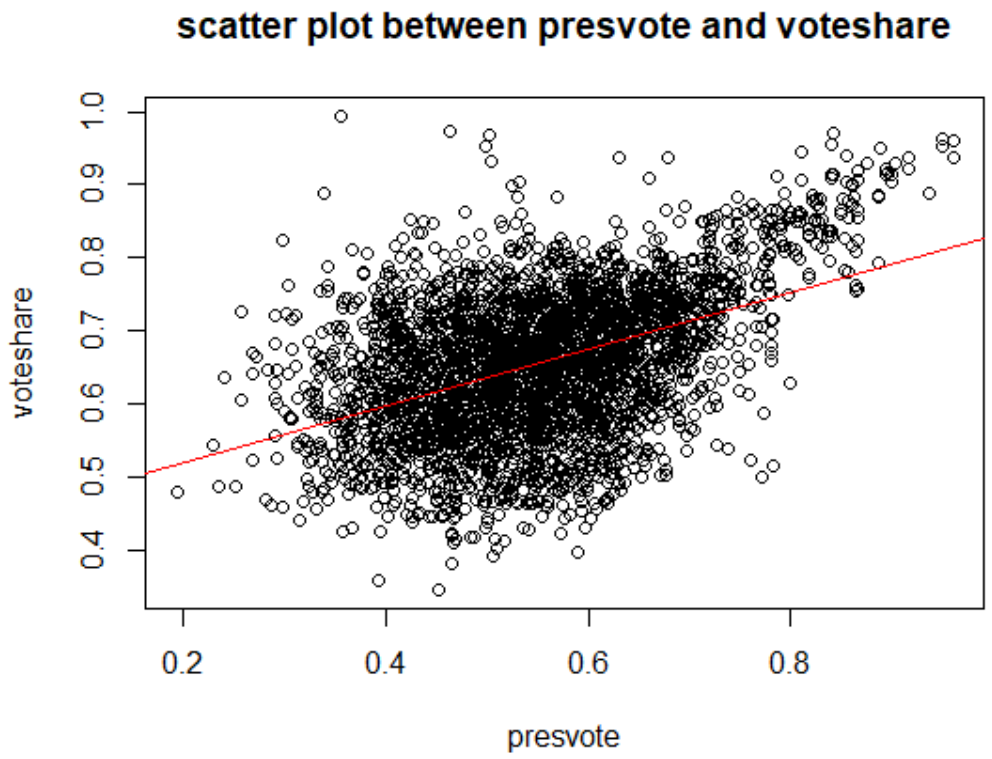
```
> print(question3_regression)

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

Coefficients:
(Intercept)     presvote
     0.4413       0.3880

> summary(question3_regression)

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

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
```

Findings: A one-unit increase in presvote is associated with a 0.38 increase in voteshare. This is a positive relation showing as presvote increases so does voteshare.

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

```
question3_scatter <- plot(incumbent$presvote, incumbent$voteshare,
                          xlab = "presvote",
                          ylab = "voteshare",
                          main = "scatter plot between presvote and
  voteshare") +
  abline(question3_regression, col="red")
```

Figure 6: Question 3 Scatter Plot



3. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{presvote}$$
$$\textit{voteshare} = 0.441 + 0.388 \times \text{presvote}$$

```
intercept3 <- round(question3_regression$coefficients[1],3)
slope3 <- round(question3_regression$coefficients[2],3)
cat(intercept3, "+", slope3, "* voteshare")
```

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.

```
question4_regression <- lm(question1_residuals ~ question2_residuals,
    data = incumbent)
print(question4_regression)
summary(question4_regression)
```

Figure 7: Regression between Q1 residuals and Q2 residuals

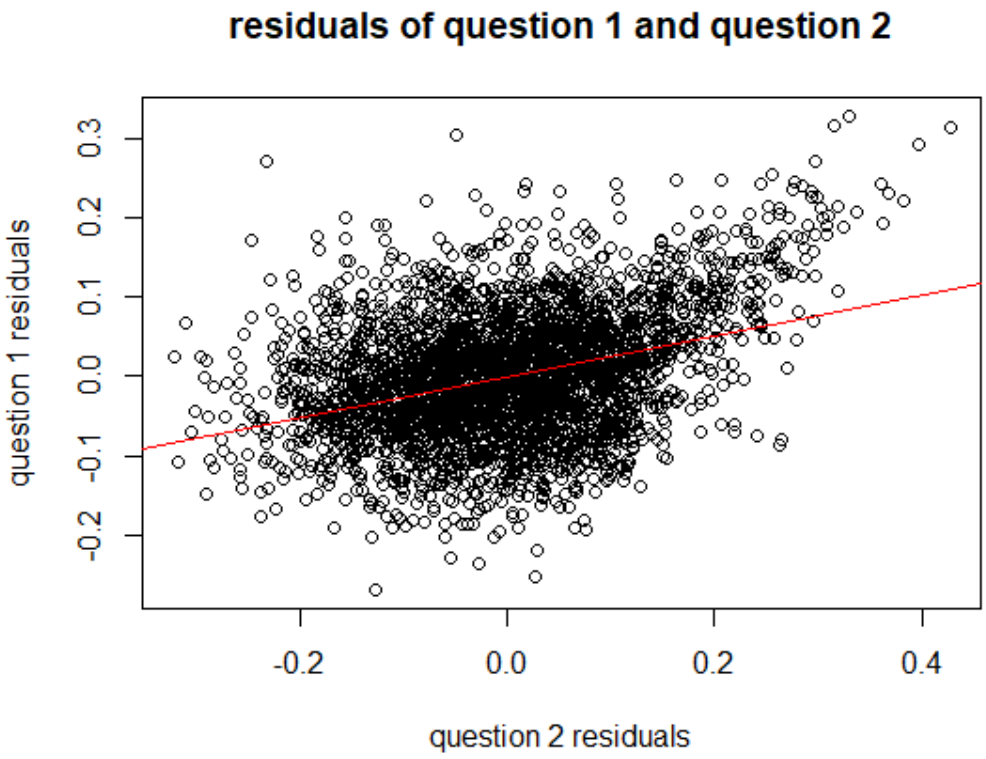
```
> print(question4_regression)

Call:
lm(formula = question1_residuals ~ question2_residuals, data = incumbent)

Coefficients:
        (Intercept)  question2_residuals
         -5.934e-18            2.569e-01

> summary(question4_regression)

Call:
lm(formula = question1_residuals ~ question2_residuals, data = incumbent)

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)         -5.934e-18  1.299e-03    0.00        1
question2_residuals  2.569e-01  1.176e-02   21.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07338 on 3191 degrees of freedom
Multiple R-squared:   0.13,     Adjusted R-squared:  0.1298
F-statistic:   477 on 1 and 3191 DF,  p-value: < 2.2e-16
```

Findings: Running this regression shows if the unexplained parts of presvote (residuals from question 2) help in explaining the remaining variation in voteshare. The positive coefficient of 0.2569 shows that some of the residual variation in presvote can help in explaining the residual variation in voteshare.

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

```r
question4_scatter <- plot(question2_residuals, question1_residuals,
                          xlab = "question 2 residuals",
                          ylab = "question 1 residuals",
                          main = "residuals of question 1 and question 2"
    ) +
  abline(question4_regression, col="red")
```

Figure 8: Question 4 Scatter Plot



3. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{question2 residuals}$$
$$question1residuals = 0 + 0.257 \times \text{question2 residuals}$$

```r
intercept4 <- round(question4_regression$coefficients[1],3)
slope4 <- round(question4_regression$coefficients[2],3)
cat(intercept4, "+", slope4, "* question2_residuals")
```

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`.

```
question5_regression <- lm(incumbent$voteshare ~ incumbent$difflog +
    incumbent$presvote)
print(question5_regression)
summary(question5_regression)
```

Figure 9: Regression between presvote and difflog

```
> print(question5_regression)

Call:
lm(formula = incumbent$voteshare ~ incumbent$difflog + incumbent$presvote)

Coefficients:
       (Intercept)   incumbent$difflog  incumbent$presvote
           0.44864             0.03554             0.25688

> summary(question5_regression)

Call:
lm(formula = incumbent$voteshare ~ incumbent$difflog + incumbent$presvote)

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 ***
incumbent$difflog  0.0355431  0.0009455   37.59   <2e-16 ***
incumbent$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
```

Findings: When presvote is held constant, a one-unit increase in difflog results in a 0.04 increase in voteshare. When difflog is held constant, a one-unit increase in presvote results in a 0.26 increase in voteshare. Both presvote and difflog have positive relationships with voteshare, however, presvote has a stronger impact.

2. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{difflog} + \hat{\beta}_2 \times \text{presvote}$$

$$\textit{voteshare} = 0.449 + 0.036 \times \text{difflog} + 0.257 \times \text{presvote}$$

```
intercept2 <- round(question2_regression$coefficients[1],3)
slope2 <- round(question2_regression$coefficients[2],3)
cat(intercept2, "+", slope2, "* difflog")
```

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 coefficient for presvote in Question 5 is identical to Question 4's coefficient for the residual of Question2 (0.257). This shows that after accounting for difflog, in question 5, there is still a partial, additional effect of presvote on voteshare. The 0.257 is a representation of the independent relationship between presvote and voteshare after removing the influence of difflog.