Problem Set 3

Applied Stats/Quant Methods 1
Minh Trinh (Student ID:24350478)

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.

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`.

   We need to declare our assumptions:

   - Each observation is independent
   - Observations are randomly generated
   - There is a linear relationship between mean of the dependent variable and the value of independent variable
   - The error term is normally distributed with mean 0 and has constant variable for all value of independent variable

We first load the dataset into our environment and run regression with `voteshare` as outcome variable and `difflog` as the explanatory variable in R. We can then examine the regression coefficients using `summary()`:

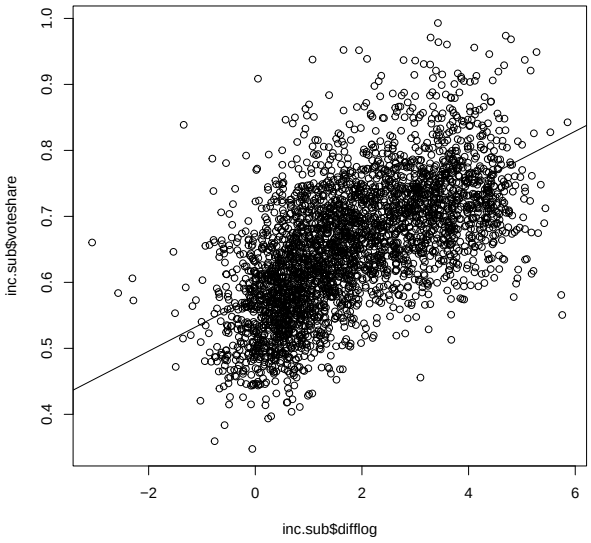
```r
# read in data
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/incumbents_subset.csv")

#Q1
#regression between voteshare and difflog
model_q1 <- lm(voteshare ~ difflog, data=inc.sub)
summary(model_q1)
```

```
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.
We draw the scatterplot and regression line in R:

```r
plot(inc.sub$difflog,inc.sub$voteshare)
abline(model_q1)
```

Figure 1: Scatterplot between difference in campaign spending and incumbent's vote share with regression line.



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

   The residual of the model can be obtained in the model object in R.

```
head(model_q1$residuals)
```

```
            1             2             3             4             5
-0.0004227622 -0.0316840149 -0.0045514943  0.0386688767  0.0355287965
```

   We will assign it to an object

```
model_resid_votes_difflog <- model_q1$residuals
```

4. Write the prediction equation.
   General form for prediction equation is:

$$\hat{y} = \beta_0 + \beta_1 x$$

   Based on the coefficients of the regression on section 1 we have the following prediction equation:

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

   With $\widehat{voteshare}$ as the predicted value for outcome variable and difflog as the explanatory variable

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`.

   We need to declare our assumptions:

   - Each observation is independent
   - Observations are randomly generated
   - There is a linear relationship between mean of the dependent variable and the value of independent variable
   - The error term is normally distributed with mean 0 and has constant variable for all value of independent variable

   We run regression with `presvote` as outcome variable and `difflog` as the explanatory variable in R. We can then examine the regression coefficients using `summary()`:

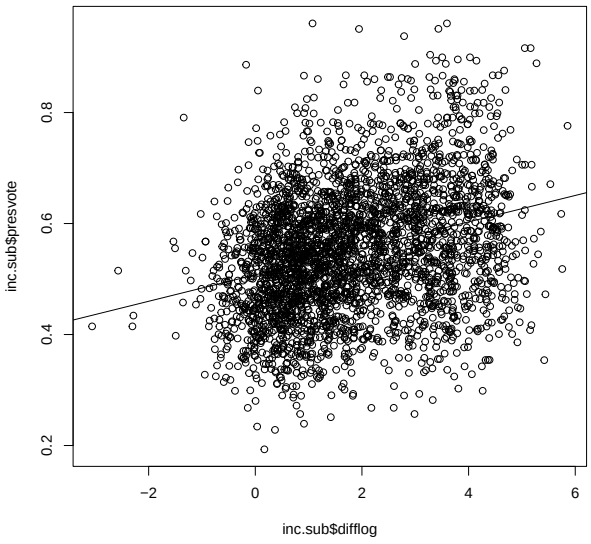
```r
model_q2 <- lm(presvote ~ difflog, data=inc.sub)
summary(model_q2)
```

```
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.
   We draw the scatterplot and regression line in R:

```r
plot(inc.sub$difflog,inc.sub$presvote)
abline(model_q2)
```

Figure 2: Scatterplot between difference in campaign spending and the presidential candidate of the incumbent's party with regression line.



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

   The residual of the model can be obtained in the model object in R.

```r
head(model_q2$residuals)
```

```
          1           2           3           4           5
0.005605594 0.037578519 -0.053134788 -0.052993694 -0.045842994
```

   We will assign it to an object

```r
model_resid_presv_difflog <- model_q2$residuals
```

4. Write the prediction equation.
   General form for prediction equation is:

$$\hat{y} = \beta_0 + \beta_1 x$$

   Based on the coefficients of the regression on section 1 we have the following prediction equation:

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

With $\widehat{presvote}$ as the predicted value for outcome variable and difflog as the explanatory variable

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`.

   We need to declare our assumptions:

   - Each observation is independent
   - Observations are randomly generated
   - There is a linear relationship between mean of the dependent variable and the value of independent variable
   - The error term is normally distributed with mean 0 and has constant variable for all value of independent variable

   We run regression with `voteshare` as outcome variable and `presvote` as the explanatory variable in R. We can then examine the regression coefficients using `summary()`:

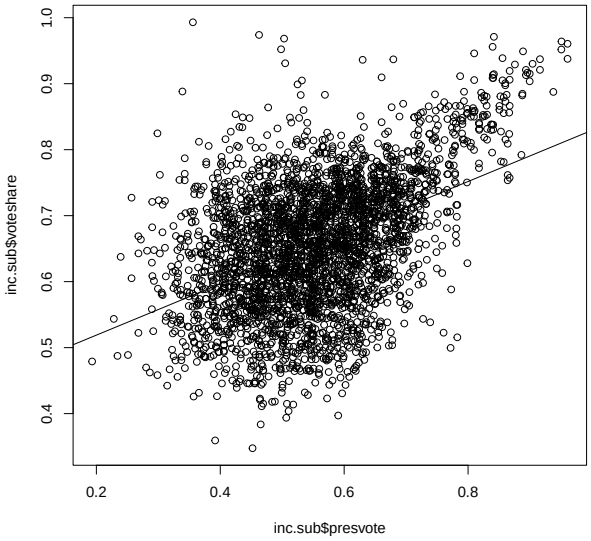
```r
model_q3 <- lm(voteshare ~ presvote, data=inc.sub)
summary(model_q3)
```

```
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    0.388018   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.
   We draw the scatterplot and regression line in R:

```r
plot(inc.sub$presvote, inc.sub$voteshare)
abline(model_q3)
```

Figure 3: Scatterplot between the presidential candidate of the incumbent's party and incumbent's vote share with regression line.



3. Write the prediction equation.
   General form for prediction equation is:

$$\hat{y} = \beta_0 + \beta_1 x$$

   Based on the coefficients of the regression on section 1 we have the following prediction equation:

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

   With $\widehat{\text{voteshare}}$ as the predicted value for outcome variable and presvote as the explanatory variable

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.
   We need to declare our assumptions:

   - Each observation is independent
   - Observations are randomly generated
   - There is a linear relationship between mean of the dependent variable and the value of independent variable
   - The error term is normally distributed with mean 0 and has constant variable for all value of independent variable

   We will first combine the two residual vector into a dataframe. Lets call the residual from Q1 r_votes_difflog and Q2 r_presv_difflog. We run regression with r_votes_difflog as outcome variable and r_presv_difflog as the explanatory variable in R. We can then examine the regression coefficients using `summary()`:

```r
#combine resid q1 q2 data
df <- data.frame(r_votes_difflog  = model_resid_votes_difflog, r_presv_
    difflog = model_resid_presv_difflog  )
#regression between resid Q1 and resid Q2
model_q4 <- lm(r_votes_difflog ~ r_presv_difflog, data = df)
summary(model_q4)
```

```
Call:
lm(formula = r_votes_difflog ~ r_presv_difflog, data = df)

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)     -1.942e-18  1.299e-03    0.00        1
r_presv_difflog  2.569e-01  1.176e-02   21.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

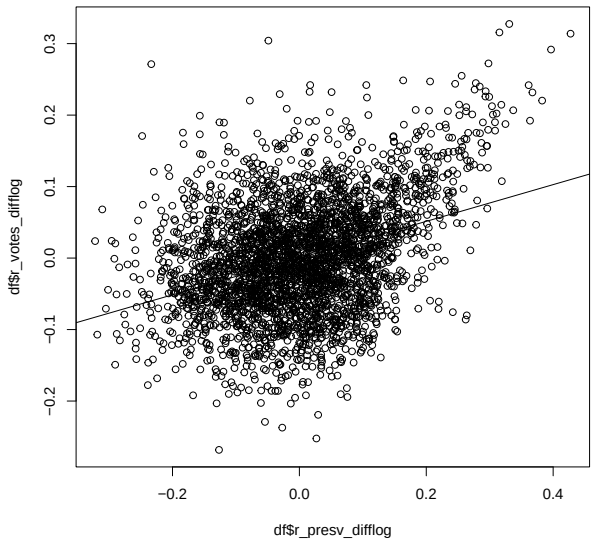
```
Residual standard error: 0.07338 on 3191 degrees of freedom
Multiple R-squared:   0.13, Adjusted R-squared:  0.1298
F-statistic:   477 on 1 and 3191 DF,  p-value: < 2.2e-16
```

One point to note is that the coefficient of the intercept is not statistically significant. Actually it is so small that it is not so much different than 0

2. Make a scatterplot of the two variables and add the regression line.
   We draw the scatterplot and regression line in R:

```r
plot(df$r_presv_difflog,df$r_votes_difflog)
abline(model_q4)
```

Figure 4: Scatterplot between r_votes_difflog and r_presv_difflog with regression line.



3. Write the prediction equation.
   General form for prediction equation is:

$$\hat{y} = \beta_0 + \beta_1 x$$

Based on the coefficients of the regression on section 1 we have the following prediction equation:

$$\widehat{\text{r\_votes\_difflog}} = -1.942 \times 10^{-18} + 0.2569 \times \text{r\_presv\_difflog}$$

With $\widehat{\text{r\_votes\_difflog}}$ as the predicted value for outcome variable and r_presv_difflog as the explanatory variable

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`.
   We need to declare our assumptions:

   - Each observation is independent
   - Observations are randomly generated
   - There is a linear relationship between mean of the dependent variable and the value of independent variables
   - The error term is normally distributed with mean 0 and has constant variable for all value of independent variable

   We run regression with `voteshare` as outcome variable and `presvote` and `difflog` as the explanatory variables in R. We can then examine the regression coefficients using `summary()`:

```r
model_q5 <- lm(voteshare ~ presvote + difflog, data = inc.sub)
summary(model_q5)
```

```
Call:
lm(formula = voteshare ~ presvote + difflog, 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 ***
presvote    0.2568770  0.0117637   21.84   <2e-16 ***
difflog     0.0355431  0.0009455   37.59   <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.

General form for prediction equation is:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Based on the coefficients of the regression on section 1 we have the following prediction equation:

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

With $\widehat{\text{voteshare}}$ as the predicted value for outcome variable and presvote, difflog as the explanatory variables

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 model coefficient, standard error, t-value and p-values for variable presvote in question 5 are identical to the model coefficient, standard error, t-value and p-values for variable r_presv_difflog in question 4.

   The coefficient for `presvote` in question 5 is the effect of the variable to `voteshare` when we have already takes into account the effect of `difflog` on both `presvote` and `voteshare`. At the same time, the residual in question 1 is the unexplained variation of `voteshare` that is not explained by `difflog`. The residual in question 2 is the unexplained variation of `presvote` that is not explained by `difflog`. If we regress the first residual on the second residual it should be the same as the remaining effect of `presvote` on the remaining unexplained variation of `voteshare` that's has taken into account all the effect of `difflog`