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`.

   let's use the lm() function in R to run a simple linear regression. In question 1: the dependent variable (outcome) is voteshare and the independent variable (explanatory) is difflog; Then, use the summary() function to see the regression results. The code and result showed as below:

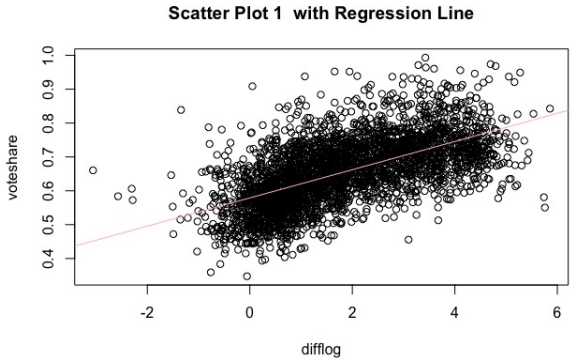
```r
model1 <- lm(voteshare ~ difflog , data = inc.sub)
summary(model)

```

```
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
```

   In summary: The coefficient for difflog is 0.041666 and it is positive, There is a positive relationship between difflog and voteshare. As difflog increases, voteshare is expected to increase. The p-value associated with the F-statistic: 2.2×10^{-16}, which is close to zero, so the linear regression model is statistically significant. The increase of the difference in campaign spending between incumbents will result in an increase of the challenger the incumbent's vote share.

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

   Create a scatter plot with a regression line using the plot() first: with difflog on the x-axis and voteshare on the y-axis. and then add pink color line by : abline() functions , code in R and plot shows below:



   The scatter plot1 of Voteshare vs. Difflog shows a positive linear relationship between the two variables. As Difflog increases, Voteshare tends to increase, there is a tight cluster of points around the major regression line.

```
plot(inc.sub$difflog, inc.sub$voteshare, main = "Scatter Plot 1 with
    Regression Line",
     xlab = "difflog", ylab = "voteshare")
abline(lm(voteshare ~ difflog, data = inc.sub), col = "pink")
```

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

```
residuals1 <- residuals(model1)
```

   Save residuals from regression model to a vector in R

4. Write the prediction equation.

   voteshare= 0.579031 (intercept) +0.041666 $\times$ difflog

   This equation can use to predict the voteshare based on the difflog: 0.579031 is the intercept which represents : the estimated value of voteshare when difflog is zero; 0.041666 represents: on average, for each one-unit increase in the difflog variable, the predicted value of voteshare is expected to increase by 0.041666

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`.

   let's use the lm() function in R to run a simple linear regression. In question 2: the dependent variable (outcome) is presvote: the vote share of the presidential candidate of the incumbent's party; and the independent variable (explanatory) is difflog: the difference between incumbent and challenger's spending. Then, use the summary() function to see the regression results. The code and result showed as below:

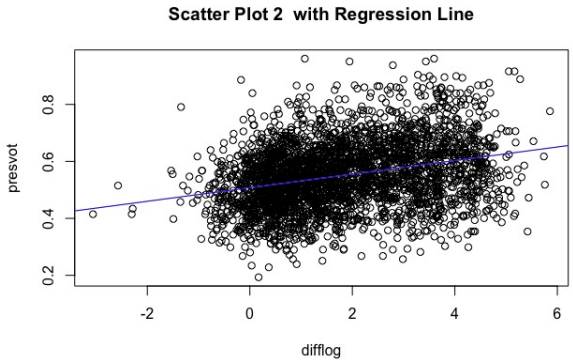
```
model2 <- lm(presvote ~ difflog , data = inc.sub)
summary(model2)

```

```
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
```

   In summary: The coefficient for difflog is 0.023837 and it is positive, There is a positive relationship between difflog and voteshare. As difflog increases, voteshare is expected to increase. The p-value associated with the F-statistic: 2.2×10^{-16}, which is close to zero, so the linear regression model is statistically significant. The difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are positively related.

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

   Create a scatter plot with a regression line using the plot() first: with difflog on the x-axis and presvote on the y-axis. and then add blue color line by : abline() functions , code in R and plot shows below:



   The scatter plot2 of presvote vs. difflog shows a positive linear relationship between the two variables. As difflog increases, presvote tends to increase, there is a tight cluster of points around the regression line when difflog is postive.

```r
plot(inc.sub$difflog, inc.sub$presvot, main = "Scatter Plot 2 with
    Regression Line",
     xlab = "difflog", ylab = "presvot")
abline(lm(presvote ~ difflog, data = inc.sub), col = "blue")
```

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

```r
residuals2 <- residuals(model2)
```

   Save residuals from regression model to a vector in R

4. Write the prediction equation.

   presvote $= 0.507583 + 0.023837 \times$ difflog

   This equation can use to predict the presvote based on the difflog: 0.507583 is the intercept which represents : the estimated value of presvote when difflog is zero; 0.041666 represents: on average, for each one-unit increase in the difflog variable, the predicted value of presvote is expected to increase by 0.023837. In general, an increase in the difference in campaign spending between incumbents will result in an increase of the vote share of the presidential candidate of the incumbent's party.

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`. let's use the lm() function in R to run a simple linear regression. In

   question 3: the dependent variable (outcome) is voteshare: the incumbent's vote share; and the independent variable (explanatory) is presvote: vote share of the presidential candidate of the incumbent's party. Then, use the summary() function to see the regression results. The code and result showed as below:

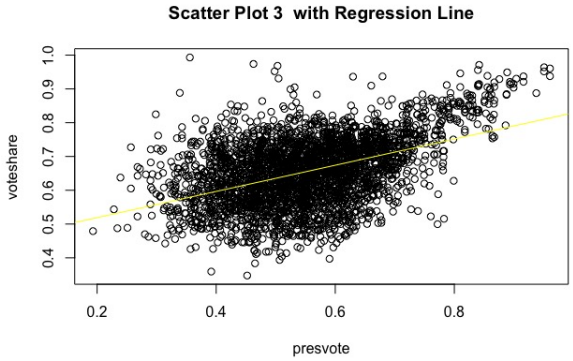
```r
model3 <- lm(voteshare ~ presvote, data = inc.sub)
summary(model3)

```

```
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
```

   In summary: The coefficient for presvote is 0.388018 and it is positive, There is a positive relationship between presvote and voteshare. As presvote increases, voteshare is expected to increase. The p-value associated with the F-statistic: 2.2×10^{-16}, which is close to zero, so the linear regression model is statistically significant. The vote share of the presidential candidate of the incumbent's party and the incumbent's vote share are positively related.

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

   Create a scatter plot with a regression line using the plot() first: with presvote on the x-axis and voteshare on the y-axis. and then add yellow color line by : abline() functions , code in R and plot shows below:



   The scatter plot3 of voteshare vs. presvote shows a positive linear relationship between the two variables. As presvote increases, voteshare tends to increase, there is a tight cluster of points around the regression line.

```
plot(inc.sub$presvote, inc.sub$voteshare, main = "Scatter Plot 3 with
    Regression Line",
     xlab = "presvote", ylab = "voteshare")
abline(lm(voteshare ~ presvote, data = inc.sub), col = "yellow")
```

3. Write the prediction equation.

   voteshare $= 0.441330 + 0.388018 \times$ presvote

   This equation can use to predict the voteshare based on the presvote: 0.441330 is the intercept which represents : the estimated value of voteshare when presvote is zero; 0.388018 represents: on average, for each one-unit increase in the presvote variable, the predicted value of voteshare is expected to increase by 0.388018. In general, an increase in the vote share of the presidential candidate of the incumbent's party will result in an increase of the incumbent's vote share.

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. let's use the lm() function in R

   to run a simple linear regression. In question 4: the dependent variable (outcome) is residual1 from question 1 and the independent variable (explanatory) is residual2 from question2. Then, use the summary() function to see the regression results. The code and result showed as below:

```r
model4 <- lm(residuals1 ~ residuals2)
summary(model4)

```

```
Call:
lm(formula = residuals1 ~ residuals2)

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) -4.860e-18  1.299e-03    0.00        1
residuals2   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
```

   In summary: The coefficient for residuals2 is 2.569e-01 and it is positive, There is a positive relationship between residuals1 and residuals2. As residuals2 increases, residuals1 is expected to increase. The p-value associated with the F-statistic: 2.2 $\times 10^{-16}$, which is close to zero, so the linear regression model is statistically significant.

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

   Create a scatter plot with a regression line using the plot() first: with residuals2 on the x-axis and residuals1 on the y-axis. and then add a red color line by : abline() functions , code in R and plot shows below:

   The scatter plot4 of residuals1 vs. residuals2 shows a positive linear relationship between the two variables. As residuals1 (the variation in voteshare) increases, residuals2 (variation in presvote) tends to increase, there is a tight cluster of points around the middle of regression line.

```
plot(residuals2 , residuals1 , main = "Scatter Plot 4  with Regression Line
    ",
     xlab = "residuals2", ylab = "residuals1")

abline(lm(residuals1 ~ residuals2), col = "red")
```

3. Write the prediction equation.

   residuals1 = -4.860e-18 (e-18:$\times 10^{-18}$) + 2.569e-01 $\times$ residuals2

   This equation can be used to predict the residuals1 in this question1 based on the residuals2 in question2: -4.860 $\times 10^{-18}$ is the intercept which represents : the estimated value of residuals1 when residuals2 is zero; 2.569e-01 represents: on average, for each one-unit increase in the residuals2 variable, the predicted value of residuals1 is expected to increase by 2.569e-01. In general, an increase in the residuals2 (variation in presvote) will result in an increase of the residuals1 (the variation in voteshare)

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`. let's use lm(outcome variables ex-

   planatory variable1 + explanatory variable2, data = xx) function in R to run a multi variable linear regression. In question 5: the dependent variable (outcome) is voteshare, and two independent variable (explanatory) are difflog and presvote. Then, use the summary() function to see the regression results. The code and result showed as below:

```
model5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
summary(model5 )

```

```
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
```

   In summary: The coefficient for difflog is 0. 0355431; The coefficient for presvote is 0.2568770. there are both positive, There is a positive relationship between voteshare and both the president's popularity and the difference in spending between incumbent and challenge. As voteshare increases, one of or both of the president's popularity and

the difference in spending between incumbent and challenge is expected to increase. The p-value associated with the F-statistic: 2.2×10^{-16}, which is close to zero, so the linear regression model is statistically significant.

2. Write the prediction equation.

   voteshare = 0.4486442+ 0.0355431× difflog + 0.2568770 × presvote

   This equation can be used to predict the voteshare based on difflog and presvote. 0.4486442 is the intercept which represents : the estimated value of residuals1 when difflog and presvote are both zero; The coefficient for difflog is 0.0355431 shows on average, the predictive change is 0.0355431 in voteshare for a one-unit increase in difflog, when holding presvote constant. The coefficient for presvote is 0.2568770 shows on average, the predictive change is 0.2568770 in voteshare for a one-unit increase in presvote, when holding difflog constant.In summary, an increase in any of explanatory variables will result in an increase in the outcome variable(voteshare).

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

   Comparing the regression output in question 4 and question 5, they both have $p - value < 2.2e - 16$ associated with the explanatory variables, there are really small p-values show strong evidence the true coefficient for explanatory is not zero:

   In question 4: p-value with $residuals2 < 2e - 16$

   residuals2 is a statistically significant predictor.

   In question 5: p-value with $difflog < 2e - 16$ p-value with $presvote < 2e - 16$

   difflog and presvote are statistically significant predictors.

   In conclusion: the explanatory variables in question 4 and question 5 are statistically significant predictors in each of the regression model.