Problem Set 3

Applied Stats/Quant Methods 1
Dan Zhang 23335541

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.

   We can use the lm() function in R to see the regression between voteshare and difflog. Run the following codes in R then we can get the results of regression1.

```r
# Run a regression considering voteshare and difflog variables
regression1<-lm(voteshare ~ difflog , data=inc.sub)
summary(regression1)
```

   We can get the following results:

```
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
```

From the results of regression1, we can see that the P-value (<2.2e-16) is extremely low, providing strong evidence in favor of a significant relationship between `voteshare` and `difflog`

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

   First, we draw the scatter plot of the two variables, and then using the abline() function to add the regression line. By running the following codes in R. We can get the graphic.

```r
# Draw a scatter plot of voteshare and difflog
pdf("Scatter plot and regression line of voteshare and difflog in R.pdf")
plot(inc.sub$difflog,inc.sub$voteshare,main="Relationship between voteshare and difflog in R",
     xlab="difflog",
     ylab="voteshare",
     pch=20,
     col="blue")
# Add the regression line
abline(regression1$coefficients[1],regression1$coefficients[2],col="red")
dev.off()
```

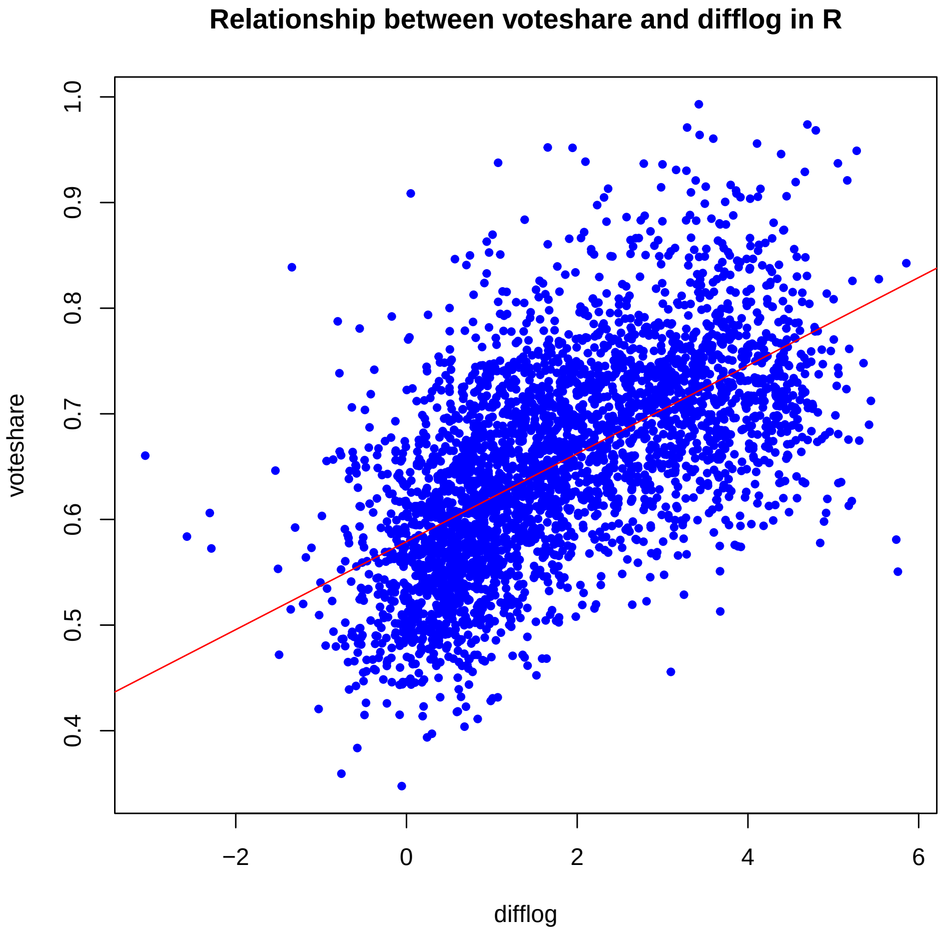

Figure 1: Scatter plot and regression line of voteshare and difflog in R

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

   Save the residuals of the regression1 in `residuals_reg1` by running the following codes:

```
# Save the residuals of the model
residuals_reg1<-residuals(regression1)
head(residuals_reg1)
```

4. Write the prediction equation.

   From the results of regression1 in Part 1.1, we can get the intercept and slope respectively. So the prediction equation would be:

```
voteshare= 0.579031 + 0.041666*difflog
```

In this model, difflog has a positive effect on voteshare, with each unit difflog increase, increasing the voteshare estimate by 0.04167. The voteshare estimate by 0.57903 when difflog is zero.

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 can use the lm() function in R to see the regression between `presvote` and `difflog`. Run the following codes in R then we can get the results of regression2.

   ```
   # Run a resgression considering presvote and difflog variables
   regression2<-lm(presvote ~ difflog , data=inc.sub)
   summary(regression2)
   ```

   We can get the following results:

   ```
   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
   ```

   From the results of regression2, we can see that the P-value (<2.2e-16) is extremely low, providing strong evidence in favor of a significant relationship between `presvote` and `difflog`

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

   First, we draw the scatter plot of the two variables, and then using the abline() function to add the regression line. By running the following codes in R. We can get the graphic.

```
# Draw a scatter plot of presvote and difflog
pdf("Scatter plot and regression line of presvote and difflog in R.pdf")
plot(inc.sub$difflog,inc.sub$presvote,main="Relationship between presvote and difflog in R",
      xlab="difflog",
      ylab="presvote",
      pch=20,
      col="blue")
# Add the regression line
abline(regression2$coefficients[1],regression2$coefficients[2],col="red")
dev.off()
```

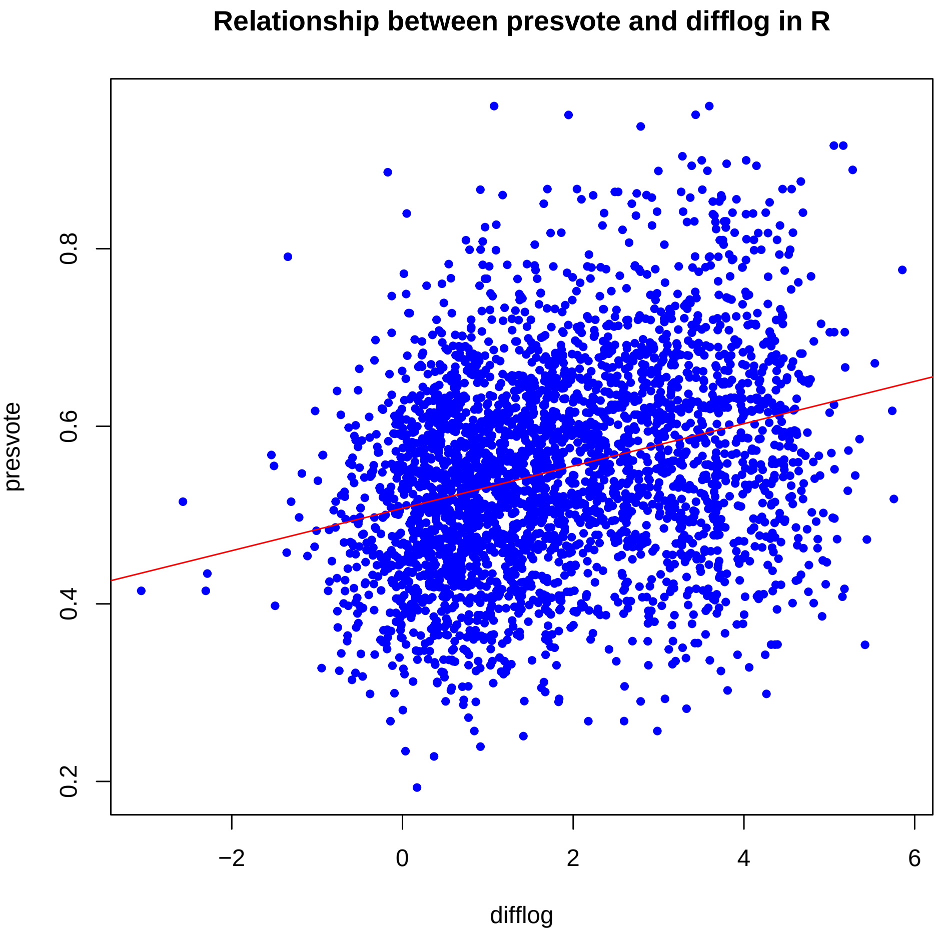

Figure 2: Scatter plot and regression line of presvote and difflog in R

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

   Save the residuals of the regression2 in `residuals_reg2` by running the following codes:

```
# Save the residuals of the model
residuals_reg2<-residuals(regression2)
head(residuals_reg2)
```

4. Write the prediction equation.

   From the results of regression2 in Part 2.1, we can get the intercept and slope respectively. So the prediction equation would be:

   `presvote= 0.507583 + 0.023837*difflog`

   In this model, difflog has a positive effect on presvote, the expected value of presvote would be 0.507583 when difflog is zero, with each addional unit change of difflog, the expected presvote changes by 0.023837.

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 can use the lm() function in R to see the regression between `voteshare` and `presvote`. Run the following codes in R then we can get the results of regression3.

   ```
   # Run a resgression considering voteshare and presvote variables
   regression3<-lm(voteshare ~ presvote, data=inc.sub)
   summary(regression3)
   ```

   We can get the following results:

   ```
   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
   ```

   From the results regression3 we can see that the P-value (<2.2e-16) is extremely low, providing strong evidence in favor of a significant relationship between `voteshare` and `presvote`

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

   First, we draw the scatter plot of the two variables, and then using the abline() function to add the regression line. By running the following codes in R. We can get the graphic.

```
# Draw a scatter plot of voteshare and presvote
pdf("Scatter plot and regression line of voteshare and presvote in R.pdf")
plot(inc.sub$presvote,inc.sub$voteshare,main="Relationship between voteshare and presvote in R",
     xlab="presvote",
     ylab="voteshare",
     pch=20,
     col="blue")
# Add the regression line
abline(regression3$coefficients[1],regression3$coefficients[2],col="red")
dev.off()
```

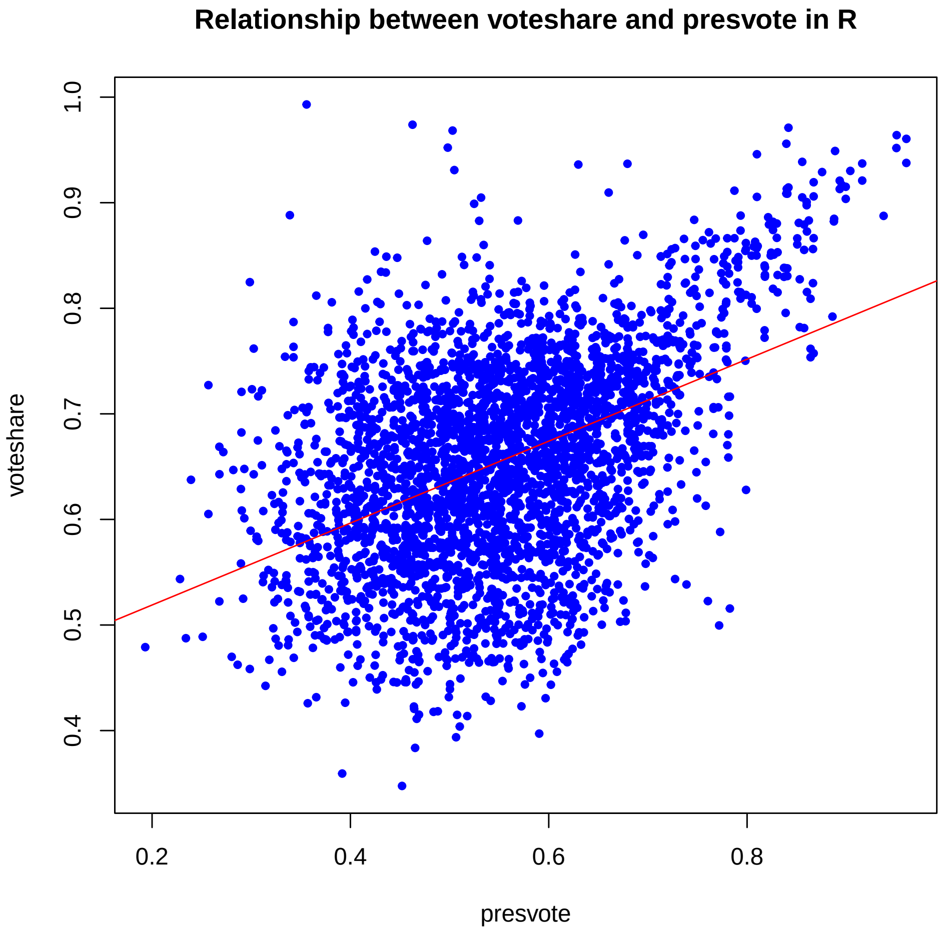

Figure 3: Scatter plot and regression line of voteshare and presvote in R

3. Write the prediction equation.

   From the results of regression3 in Part 3.1, we can get the intercept and slope respectively. So the prediction equation would be:

   ```
   voteshare= 0.441330 + 0.388018*presvote
   ```

   In this model, presvote has a positive effect on voteshare, the estimate value of voteshare would be 0.441330 when presvote is zero, with each addional unit change of presvote, the estimate voteshare changes by 0.388018.

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 can use the lm() function in R to see the regression between `voteshare` and `presvote`. Run the following codes in R then we can get the results of regression4.

   ```r
   # Run a regression considering residuals of regression1 and residuals of regression2
   regression4<-lm(residuals_reg1 ~ residuals_reg2)
   summary(regression4)
   ```

   We can get the following results:

   ```
   Call:
   lm(formula = residuals_reg1 ~ residuals_reg2)
   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
   residuals_reg2  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
   ```

   From the results regression4 we can see that the P-value (<2.2e-16) is extremely low, providing strong evidence in favor of a significant relationship between `residuals_reg1` and `residuals_reg2`

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

   First, we draw the scatter plot of the two variables, and then using the abline() function to add the regression line. By running the following codes in R. We can get the graphic.

```
# Draw a scatter plot of residuals_reg1 and residuals_reg2
pdf("Scatter plot and regression line of residuals_reg1 and residuals_
   reg2 in R.pdf")
plot(residuals_reg2,residuals_reg1,main="Relationship between residuals_
   reg1 and residuals_reg2 in R",
     xlab="residuals_reg2",
     ylab="residuals_reg1",
     pch=20,
     col="blue")
# Add the regression line
abline(regression4$coefficients[1],regression4$coefficients[2],col="red")
dev.off()
```

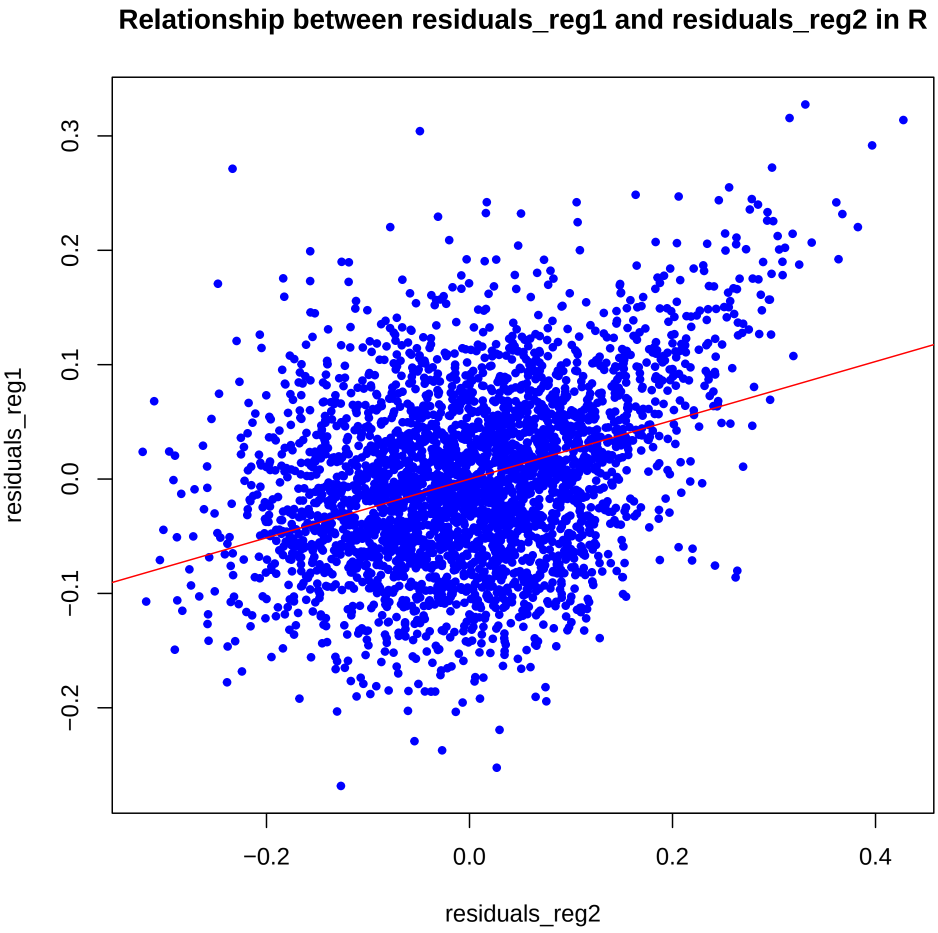

Figure 4: Scatter plot and regression line of residuals_reg1 and residuals_reg2 in R

3. Write the prediction equation.

   From the results of regression4 in Part 4.1, we can get the intercept and slope respectively. So the prediction equation would be:

   ```
   residuals_reg1 = -1.942e-18 + (2.569e-01)*residuals_reg2
   ```

   In this model, residuals_reg2 has a positive effect on residuals_reg1, the estimate value of residuals_reg1 would be -1.942e-18 when residuals_reg2 is zero, with each addional unit change of residuals_reg2, the estimate residuals_reg1 changes by 2.569e-01.

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 can use the lm() function in R to see the regression between `voteshare` , `difflog` and `presvote`. Run the following codes in R then we can get the results of regression5.

   ```
   # Run a regression considering voteshare, difflog and presvote variables
   regression5<-lm(voteshare ~ difflog + presvote, data = inc.sub)
   summary(regression5)
   ```

   We can get the following results:

   ```
   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
   ```

   From the results regression5 we can see that the P-value (<2.2e-16) is extremely low, providing strong evidence in favor of a significant relationship between `voteshare` , `difflog` and `presvote`

2. Write the prediction equation.

   From the results of regression5 in Part 5.1, we can get the intercept and slopes respectively. So the prediction equation would be:

   ```
   voteshare = 0.4486442 + 0.0355431*difflog + 0.2568770*presvote
   ```

   In this model, both difflog and presvote have positive effects on voteshare, the estimate value of voteshare would be 0.4486442 when difflog and presvote are zero. Controlling presvote unchanged, with each addional unit change of difflog, the estimate voteshare changes by 0.0355431. Controlling difflog unchanged, with one unit presvote increase, increasing the voteshare estimate by 0.2568770.

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

   From the regression results of regression5 and regression4, we can see that the Residuals of regression4 and regression5 are the same.

   The reason for this situation is that the model of regression4 is obtained by regressing residuals_reg1 on residuals_reg2, while the model of regression5 is obtained by directly regressing voteshare on difflog and presvote. Since residuals_reg1 and residuals_reg2 are respectively the residuals of voteshare on difflog and the residuals of presvote on difflog, the residuals, in the regression results will be the same. This is also why we see these same values in both regression results.