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`.

```r
# Read the dataset
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")
head(inc.sub, n=10)
dim(inc.sub)
names(inc.sub)

# Question 1

```

```r
# 1-1

# Run linear regression model
regression_model <- lm(voteshare ~ difflog, data = inc.sub)

# Check the result of regression model
summary(regression_model)

# Take the coefficients from model
coefficients_package <- coef(regression_model)

# Sort the coefficients from coefficients package
intercept <- coefficients_package[1]
print(round(intercept, digits = 2))

coef_diddlog <- coefficients_package[2]
print(round(coef_diddlog, digits = 2))
```

```
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
```

Interpretation:

The mathematical expression of the linear regression model is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

And there are some assumptions regarding regression analysis as follows:

1) Error follows the normal distribution:

$$\epsilon_i \sim N(0, \sigma^2)$$

2) Randomized data generation
3) Independent observations
4) linearity and normality
5) Constant variance

After calculating in R, the intercept is 0.58, the slope is 0.04 , so the regression equation is voteshare = 0.58+0.04*difflog

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

```r
# 1-2

library(ggplot2)

# Make scatterplot of the two variables and add the regression line by
    using ggplot
ggplot(inc.sub, aes(x = difflog, y = voteshare))+
  geom_point() + # Add scatterplot
  geom_smooth(method = "lm", se = FALSE) + # Add the regression line
  labs(x = "Difflog", y = "Voteshare") + # Add axis labels
  ggtitle("Scatterplot with Regression Line (Voteshare ~ difflog *1-2*)")
    # Add title
```

Interpretation:

Intercept(0.58) : When difflog = 0 , voteshare = 0.58
Slope(0.04) : because the slope is more than 0, there is a positive relationship between voteshare and difflog; and 1 unit increase in difflog is associated with 0.04 unit increase in voteshare .

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

```r
# 1-3

# Save the residuals of the model in a separate object
residuals <- resid(regression_model)

# Check the head line of residuals
```

```
head(residuals)
```

4. Write the prediction equation.
   Interpretation:
   Check the validity of this regression equation:
   1)Check the significance of coefficients: from the summary of the regression model, the P-values of the coefficients are both less than 0.05 , so the coefficients are significant.
   2)Check the residuals are valid in 1-3.
   So, The prediction equation is : voteshare-hat = 0.58+0.04*difflog-hat

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
# 2-1

# Perform linear regression analysis
regression_model_2 <- lm(presvote ~ difflog, data = inc.sub)

# Check the result of regression model
summary(regression_model_2)

# Take the coefficients from model
coefficients_package_2 <- coef(regression_model_2)
print(coefficients_package_2)

# Sort the coefficients from coefficients package
intercept_2 <- coefficients_package_2[1]
print(round(intercept_2, digits = 2))

coef_diddlog_2 <- coefficients_package_2[2]
print(round(coef_diddlog_2, digits = 2))
```

```
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
```

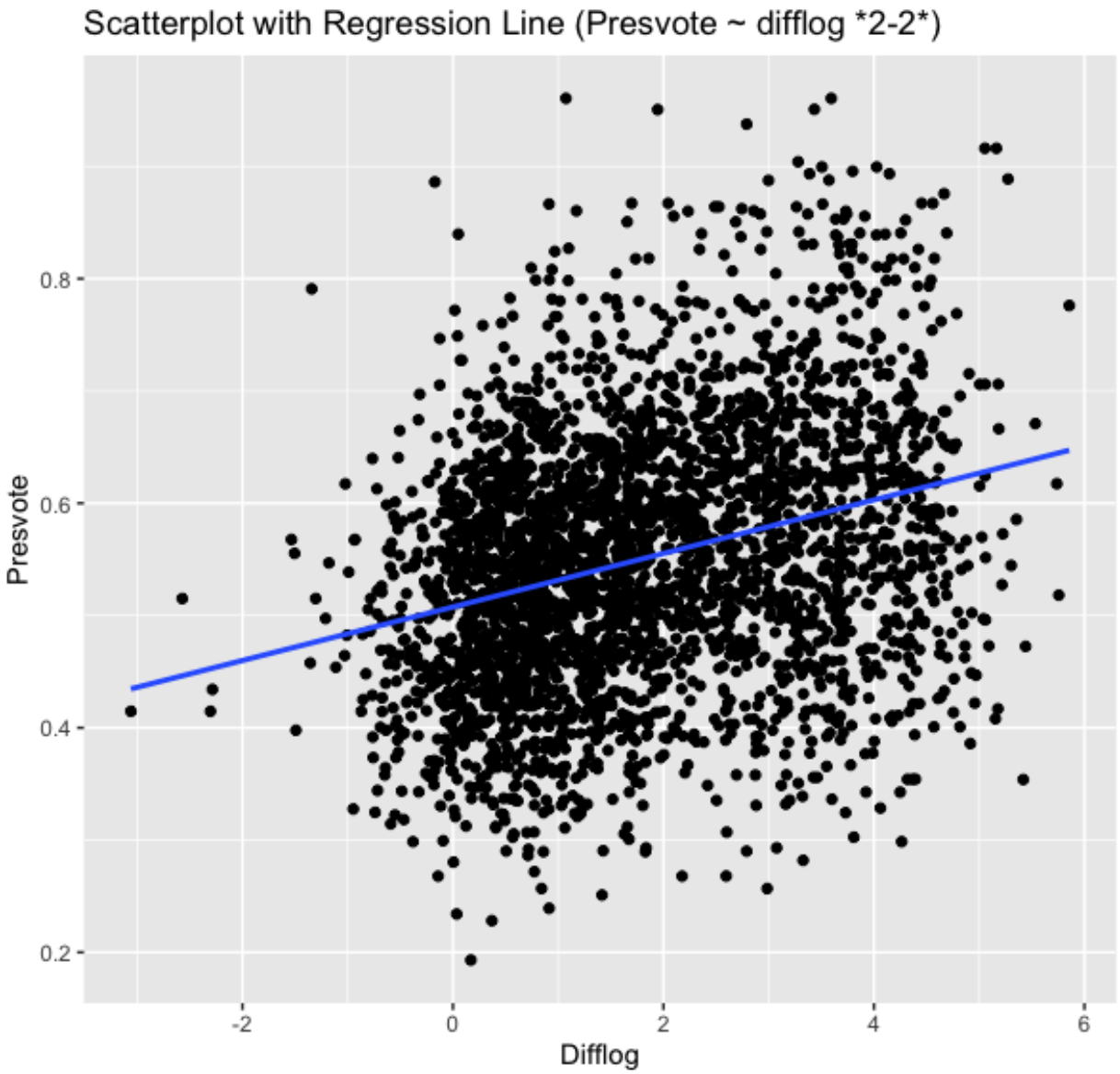
Interpretation:

After calculating in R, the intercept is 0.51, the slope is 0.02, so the regression equation is presvote = 0.51+0.02*difflog

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

```r
library(ggplot2)

# Make scatterplot of the two variables and add the regression line by
    using ggplot
ggplot(inc.sub, aes(x = difflog, y = presvote))+
  geom_point() + # Add scatterplot
  geom_smooth(method = "lm", se = FALSE) + # Add the regression line
  labs(x = "Difflog", y = "Presvote") + # Add axis labels
  ggtitle("Scatterplot with Regression Line (Presvote ~ difflog *2-2*)")
   # Add title
```



Interpretation:

Intercept(0.51): When difflog = 0 , presvote = 0.51
Slope(0.02): because the slope is more than 0, there is a positive relationship between presvote and difflog; and 1 unit increase in difflog is associated with 0.02 unit increase in presvote .

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

```
# 2-3
```

```r

# Save the residuals of the model in a separate object
residuals_2 <- resid(regression_model_2)

# Check the head line of residuals_2
head(residuals_2)
```

4. Write the prediction equation.
   Interpretation:
   Check the validity of this regression equation:
   1)Check the significance of coefficients: from the summary of the regression model, the P-values of the coefficients are both less than 0.05 , so the coefficients are significant.
   2)Check the residuals are valid in 2-3.
   So, The prediction equation is : presvote-hat = 0.51+0.02*difflog-hat

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
# 3-1

# Perform linear regression analysis
regression_model_3 <- lm(voteshare ~ presvote, data = inc.sub)

# Check the result of regression model
summary(regression_model_3)

# Take the coefficients from model
coefficients_package_3 <- coef(regression_model_3)
print(coefficients_package_3)

# Sort the coefficients from coefficients package
intercept_3 <- coefficients_package_3[1]
print(round(intercept_3, digits = 2))

coef_presvote_3 <- coefficients_package_3[2]
print(round(coef_presvote_3, digits = 2))
```

```
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
```

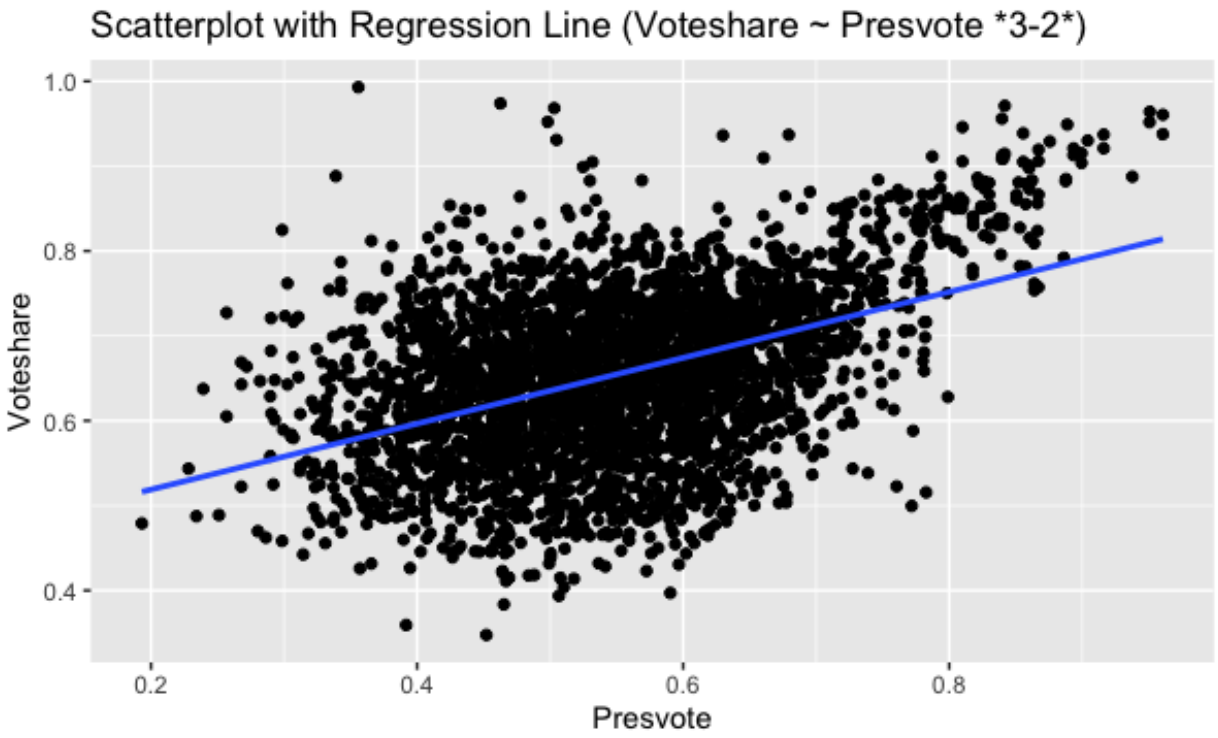
Interpretation:

After calculating in R, the intercept is 0.44,the slope is 0.39, so the regression equation is voteshare = 0.44+0.39*presvote

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

```r
library(ggplot2)

# Make scatterplot of the two variables and add the regression line by using ggplot
ggplot(inc.sub, aes(x = presvote, y = voteshare))+
  geom_point() + # Add scatterplot
  geom_smooth(method = "lm", se = FALSE) + # Add the regression line
  labs(x = "Presvote", y = "Voteshare") + # Add axis labels
  ggtitle("Scatterplot with Regression Line (Voteshare ~ Presvote *3-2*)") # Add title
```



Interpretation:
Intercept(0.44): When presvote = 0 , voteshare = 0.44
Slope(0.39): because the slope is more than 0, there is a positive relationship between voteshare and presvote; and 1 unit increase in presvote is associated with 0.39 unit increase in voteshare .

3. Write the prediction equation.
   Interpretation:
   Check the validity of this regression equation:
   Check the significance of coefficients: from the summary of the regression model, the P-values of the coefficients are both less than 0.05 , so the coefficients are significant.
   So, the prediction equation is : voteshare-hat = 0.44+0.39*presvote-hat

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.

```r
# 4-1

# Perform linear regression analysis
regression_model_4 <- lm(residuals ~ residuals_2, data = inc.sub)

# Check the result of regression model
summary(regression_model_4)

# Take the coefficients from model
coefficients_package_4 <- coef(regression_model_4)
print(coefficients_package_4)

# Sort the coefficients from coefficients package
intercept_4 <- coefficients_package_4[1]
print(round(intercept_4, digits = 2))

coef_residuals_2_4 <- coefficients_package_4[2]
print(round(coef_residuals_2_4, digits = 2))
```

```
Call:
lm(formula = residuals ~ residuals_2, 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) -1.942e-18  1.299e-03    0.00        1
residuals_2  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
```

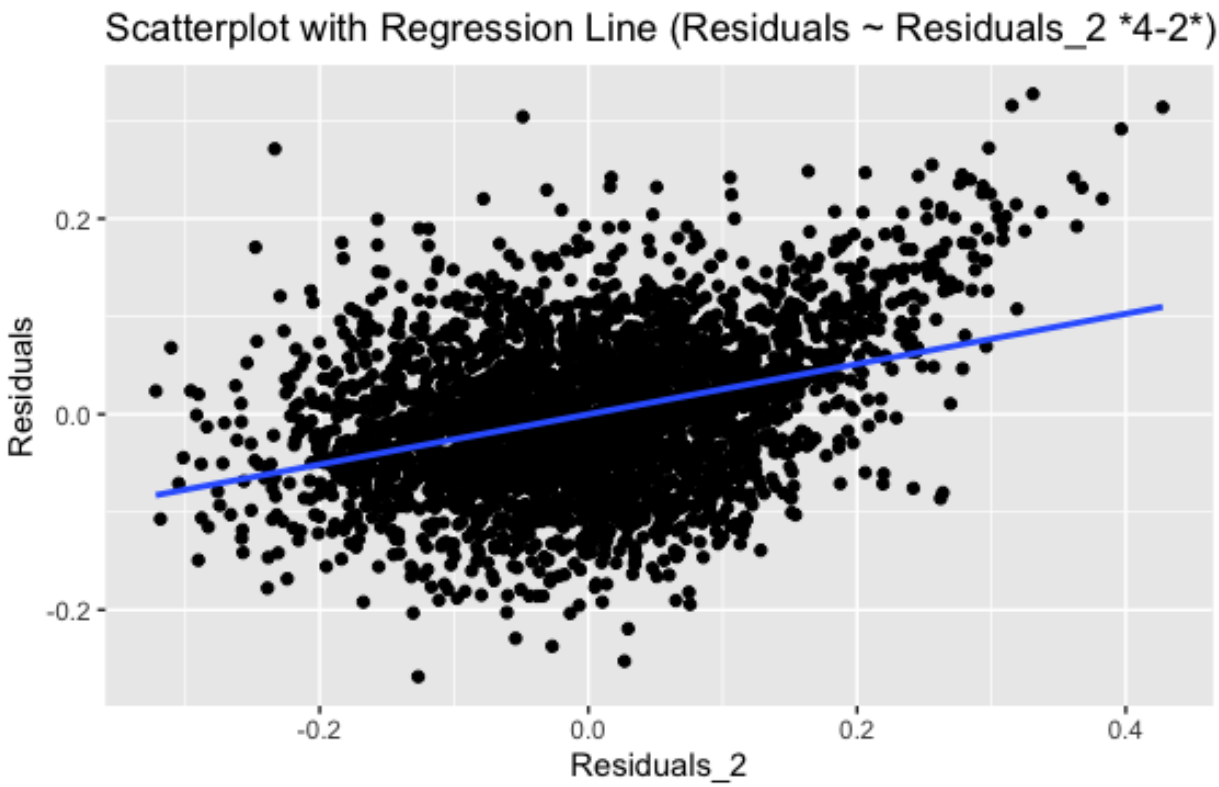
Interpretation:

After calculating in R, the intercept is almost 0, the slope is 0.26, so the regression equation is residuals = 0+0.26*residuals-2

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

```r
library(ggplot2)

# Make scatterplot of the two variables and add the regression line by
    using ggplot
ggplot(inc.sub, aes(x = residuals_2, y = residuals))+
  geom_point() + # Add scatterplot
  geom_smooth(method = "lm", se = FALSE) + # Add the regression line
  labs(x = "Residuals_2", y = "Residuals") + # Add axis labels
  ggtitle("Scatterplot with Regression Line (Residuals ~ Residuals_2 *4-2
    *)") # Add title
```



Interpretation:

Intercept(0) : When residuals-2 = 0 , residuals= 0
Slope(0.26) : because the slope is more than 0, there is a positive relationship between residuals and residuals-2; and 1 unit increase in residuals-2 is associated with 0.26 unit increase in residuals .

3. Write the prediction equation.
   Interpretation:

Check the validity of this regression equation:
Check the significance of coefficients: from the summary of the regression model: the P-value of the intercept is more than 0.05 , so the intercept is not significant, that means the intercept has almost no effect on the model; the P-value of the slope is less than 0.05 , so the slope is significant;

So, the prediction equation is : residuals-hat = 0.26*residuals-2-hat

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`.

```r
# 5-1

# Perform linear regression analysis
regression_model_5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)

# Check the result of regression model
summary(regression_model_5)

# Take the coefficients from model
coefficients_package_5 <- coef(regression_model_5)
print(coefficients_package_5)

# Sort the coefficients from coefficients package
intercept_5 <- coefficients_package_5[1]
print(round(intercept_5, digits = 2))

coef_difflog_5 <- coefficients_package_5[2]
print(round(coef_difflog_5, digits = 2))

coef_presvote_5 <- coefficients_package_5[3]
print(round(coef_presvote_5, digits = 2))
```

```
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
```

Interpretation:

After calculating in R, the intercept is 0.45, the slope of difflog is 0.04 , and the slope of presvote is 0.26, so the regression equation is voteshare = 0.45+0.04*difflog + 0.26*presvote

- Intercept(0.45) : When difflog = 0 and presvote = 0, voteshare = 0.45
- Slope(0.04) : because the slope is more than 0, there is a positive relationship between voteshare and difflog; and with the value of presvote remianing constant, the 1 unit increase in difflog is associated with 0.04 unit increase in voteshare .
- Slope(0.26) : because the slope is more than 0, there is a positive relationship between voteshare and presvote ; and with the value of difflog remianing constant, 1 unit increase in presvote is associated with 0.26 unit increase in voteshare .

2. Write the prediction equation.
   Interpretation:

Check the validity of this regression equation:
Check the significance of coefficients: from the summary of the regression model, the P-values of the coefficients are both less than 0.05 , so the coefficients are significant.

So, the prediction equation is : voteshare-hat = 0.45+0.04*difflog-hat + 0.26*presvote-hat

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

Comparing the residuals of regression model in Q4 and Q5, they are identical, because there is a certain degree of collinearity between difflog and presvote.