Applied Stats - Problem Set 3

Yana Konshyna

November 18, 2023

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.

```
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")
View(inc.sub)
head(inc.sub)
```

After loading incumbents_subset.csv dataset into the working environment, I execute the regression model in which the vote share of the presidential candidate of the incumbent's party (voteshare) is explained by the difference in campaing spending between incumbent and challenger (difflog). Then I investigate the estimated coefficients of the model using summary().

Code in R:

```r
# 1.1. Running a regression where the outcome variable is voteshare and the explanatory variable is difflog
model_q1<- lm(voteshare ~ difflog, data = inc.sub)
model_summary_q1 <- summary(model_q1)
print(model_summary_q1)
```

Output:

```
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
```

Conclusion: Increasing the difference in campaing spending between incumbent and challenger by 1 unit, on average, will increase the incumbent's vote share by 0.041 units. The estimated coefficient is statistically differentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 ($\approx$2e-16).

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

   Making a scatterplot of the two variables and adding the regression line by using plot() and abline().

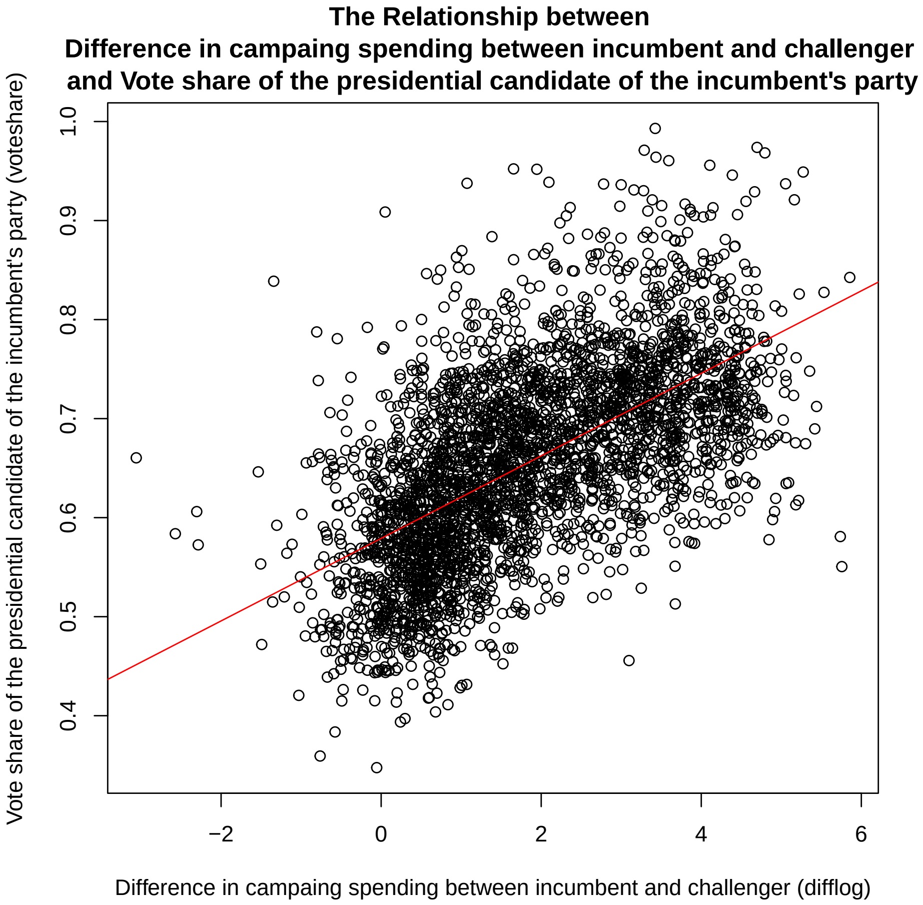
Code in R:

```r
pdf("scatterplot_q1.pdf")
plot(inc.sub$difflog,
     inc.sub$voteshare,
     xlab="Difference in campaing spending between incumbent and challenger (difflog)",
     ylab="Vote share of the presidential candidate of the incumbent's party (voteshare)",
     main="The Relationship between \nDifference in campaing spending between incumbent and challenger \nand Vote share of the presidential candidate of the incumbent's party")

# Adding the regression line
```

```
abline(model_q1, col = "red")
dev.off()
```

Scatterplot 1 with the regression line.



Conclusion: There is a positive relationship between `difflog` (the difference in campaign spending between incumbent and challenger) and `voteshare` (the incumbent's vote share).

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

After execution of the regression model I can save the residuals of the model in a separate object by using residuals().

Code in R:

```r
residuals_q1 <- residuals(model_q1)
head(residuals_q1)
```

Output:

```
1             2             3
-0.0004227622 -0.0316840149 -0.0045514943
4             5             6
0.0386688767  0.0355287965  0.0322832521
```

4. Write the prediction equation.

   The formula of prediction equation is:

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{X}_{1i} + \beta_2 \mathrm{X}_{2i} + ... + \beta_k \mathrm{X}_{ki} + \epsilon_i$$

   Or

$$Y = Xb + e$$

   Getting the coefficients for writing the prediction equation.

   Code in R:

```r
coefficient_q1 <- model_summary_q1$coefficients
print(coefficient_q1)
```

   Output:

```
              Estimate   Std. Error   t value      Pr(>|t|)
(Intercept) 0.57903071 0.0022513886 257.18826  0.000000e+00
difflog     0.04166632 0.0009679924  43.04406 1.359767e-319
```

   Writing the prediction equation.

   Code in R:

```r
cat("Prediction Equation:\n voteshare =", coefficient_q1[1], "+",
   coefficient_q1[2], "* difflog\n")
```

Output:

```
Prediction Equation:
voteshare = 0.5790307 + 0.04166632 * 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`.

   Executing the regression model in which the incumbent's electoral success (`presvote`) is explained by the difference between incumbent and challenger's spending (`difflog`). Then I investigate the estimated coefficients of the model using summary().

   Code in R:

   ```r
   # 2.1.Running a regression where the outcome variable is presvote and the explanatory variable is difflog.
   model_q2<- lm(presvote ~ difflog , data = inc.sub)
   model_summary_q2 <- summary(model_q2)
   print(model_summary_q2)
   ```

   Output:

   ```
   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
   ```

   Conclusion: Increasing the difference in campaing spending between incumbent and challenger by 1 unit, on average, will increase the incumbent's electoral success by 0.0238 units. The estimated coefficient is statistically diferentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 ($\approx$2e-16).

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

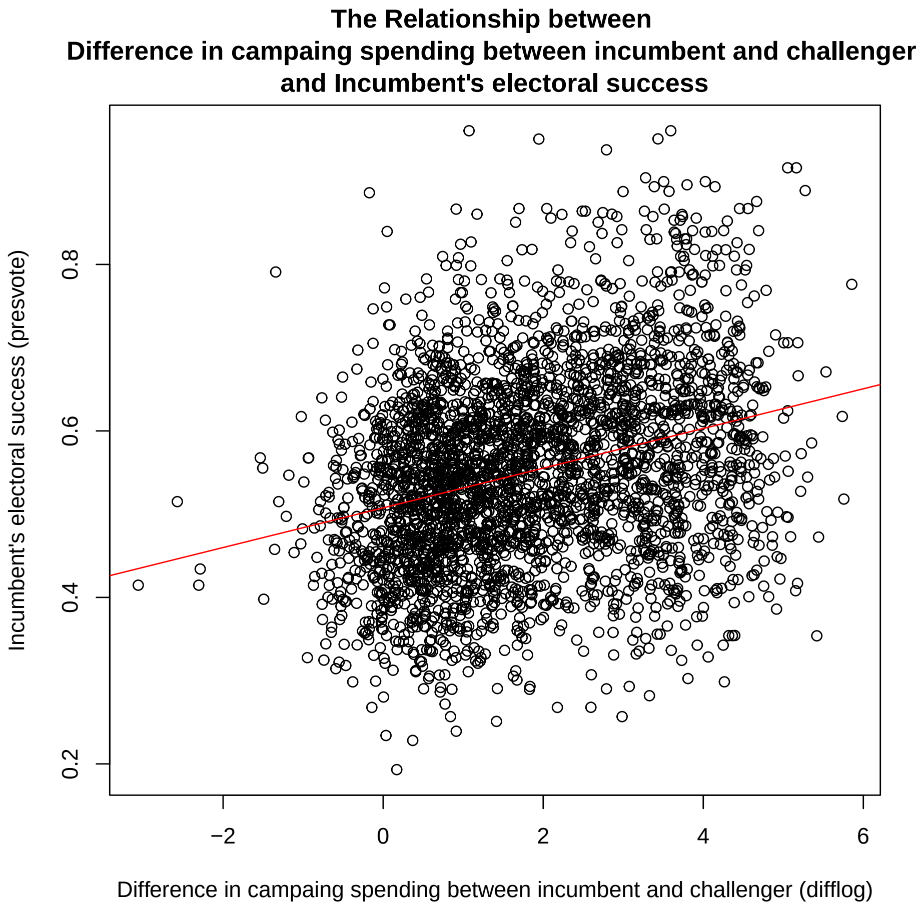
Making a scatterplot of the two variables and adding the regression line by using plot() and abline().

Code in R:

```
pdf("scatterplot_q2.pdf")
plot(inc.sub$difflog,
     inc.sub$presvote,
     xlab="Difference in campaing spending between incumbent and challenger (difflog)",
     ylab="Incumbent's electoral success (presvote)",
     main="The Relationship between \nDifference in campaing spending between incumbent and challenger \nand Incumbent's electoral success")

# Adding the regression line
abline(model_q2, col = "red")
dev.off()
```

Scatterplot 2 with the regression line.



Conclusion: There is a positive relationship between `difflog` (the difference in campaign spending between incumbent and challenger) and `presvote` (the incumbent's electoral success).

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

   After execution of the regression model I can save the residuals of the model in a separate object by using residuals().

   Code in R:

```
residuals_q2 <- residuals(model_q2)
head(residuals_q2)
```

   Output:

```
1             2             3
0.005605594   0.037578519 -0.053134788
4             5             6
-0.052993694 -0.045842994  0.074339701
```

4. Write the prediction equation.

   Getting the coefficients for writing the prediction equation.

   Code in R:

```
coefficient_q2 <- model_summary_q2$coefficients
print(coefficient_q2)
```

   Output:

```
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 0.50758333 0.003160529 160.60077 0.000000e+00
difflog     0.02383723 0.001358880  17.54182 7.681359e-66
```

   Writing the prediction equation.

   Code in R:

```
cat("Prediction Equation:\n presvote =", coefficient_q2[1], "+",
    coefficient_q2[2], "* difflog\n")
```

Output:

```
Prediction Equation:
presvote = 0.5075833 + 0.02383723 * 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`.

   Executing the regression model in which the vote share of the presidential candidate of the incumbent's party (`voteshare`) is explained by the incumbent's electoral success (`presvote`). Then I investigate the estimated coefficients of the model using summary().

   ### Code in R:

```r
model_q3<- lm(voteshare ~ presvote, data = inc.sub)
model_summary_q3 <- summary(model_q3)
print(model_summary_q3)
```

   ### Output:

```
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
```

   Conclusion: Increasing the incumbent's electoral success by 1 unit, on average, will increase the vote share of the presidential candidate of the incumbent's party by 0.388 units. The estimated coefficient is statistically diferentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 ($\approx$2e-16).

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

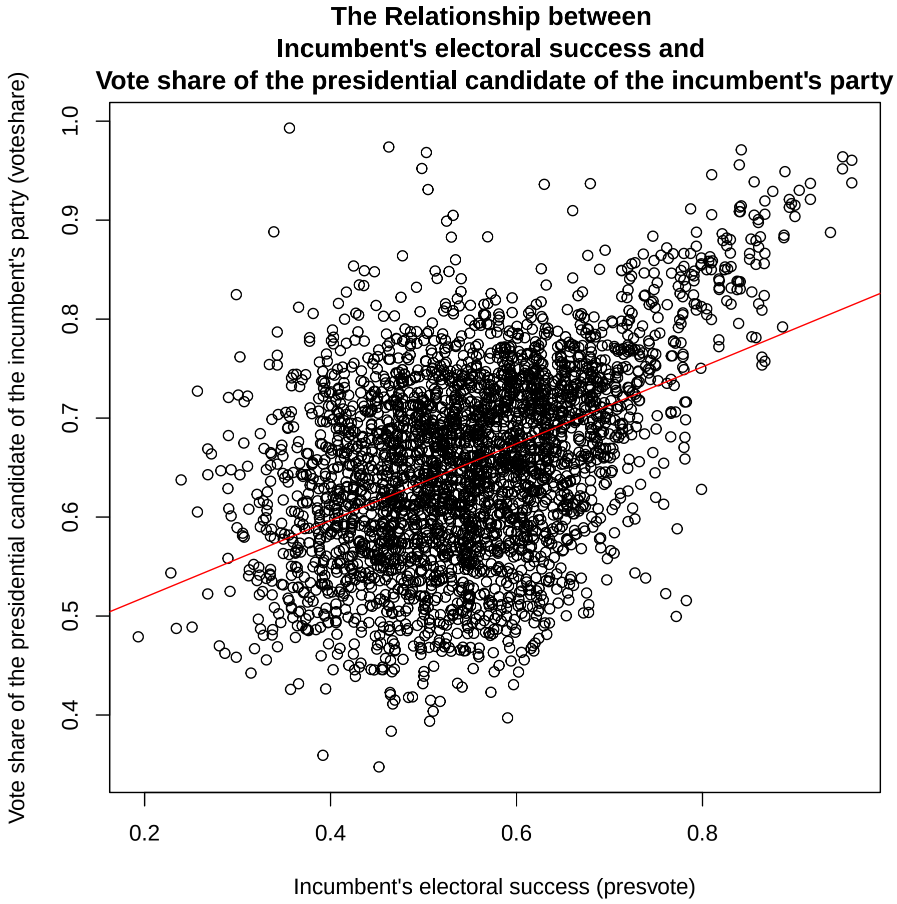
   Making a scatterplot of the two variables and adding the regression line by using plot() and abline().

Code in R:

```r
pdf("scatterplot_q3.pdf")
plot(inc.sub$presvote,
     inc.sub$voteshare,
     xlab="Incumbent's electoral success (presvote)",
     ylab="Vote share of the presidential candidate of the incumbent's
    party (voteshare)",
     main="The Relationship between \nIncumbent's electoral success and \
    nVote share of the presidential candidate of the incumbent's party")

# Adding the regression line
abline(model_q3, col = "red")
dev.off()
```

Scatterplot 3 with the regression line.



Conclusion: There is a positive relationship between `presvote` (the incumbent's electoral success) and `voteshare` (the vote share of the presidential candidate of the incumbent's party).

3. Write the prediction equation.

   Getting the coefficients for writing the prediction equation.

   Code in R:

```
coefficient_q3 <- model_summary_q3$coefficients
print(coefficient_q3)
```

   Output:

```
             Estimate  Std. Error  t value      Pr(>|t|)
(Intercept) 0.4413299 0.007598612 58.08033  0.000000e+00
presvote    0.3880184 0.013493130 28.75674 6.586314e-162
```

   Writing the prediction equation.

   Code in R:

```
cat("Prediction Equation:\n voteshare =", coefficient_q3[1], "+",
    coefficient_q3[2], "* presvote\n")
```

   Output:

```
Prediction Equation:
voteshare = 0.4413299 + 0.3880184 * presvote
```

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.

   Executing the regression model in which the residuals from Question 1 (`residuals1`) are explained by the residuals from Question 2 (`residuals2`). Then I investigate the estimated coefficients of the model using summary().

   Code in R:

   ```
   residual_model<- lm(residuals_q1 ~ residuals_q2)
   residual_summary_q4 <- summary(residual_model)
   print(residual_summary_q4)
   ```

   Output:

   ```
   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
   st_residuals_q2  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
   ```

   Conclusion: Increasing the residuals from Question 2 by 1 unit, on average, will increase the residuals from Question 1 by 0.257 units. The estimated coefficient is statistically diferentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 ($\approx$2e-16).

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

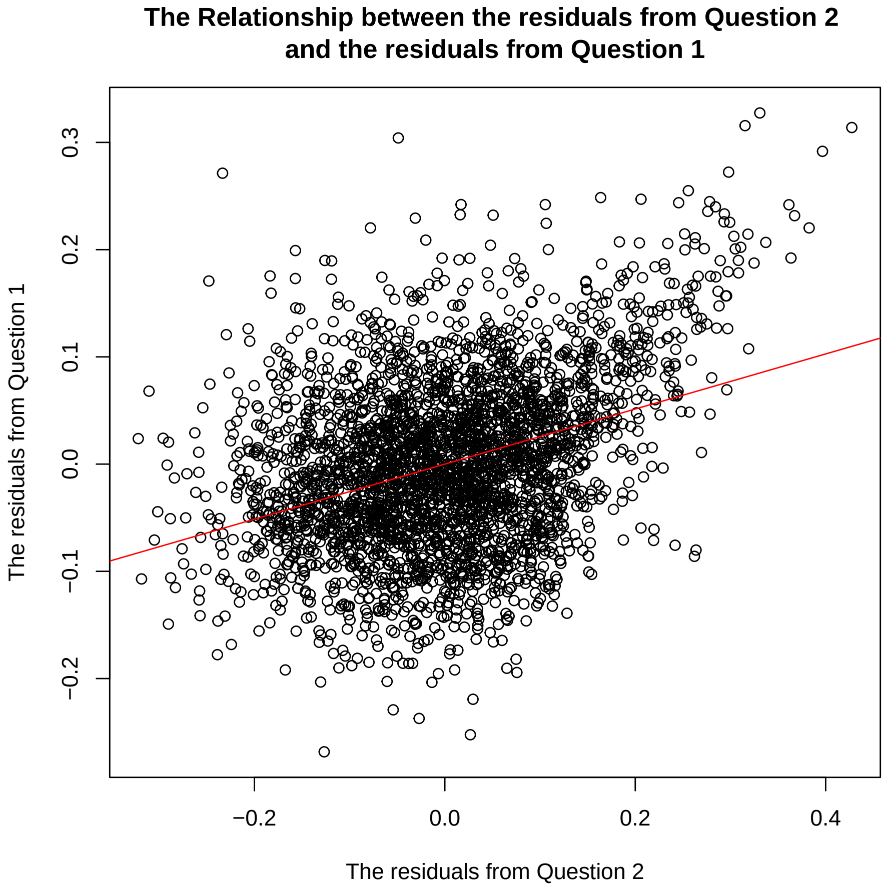
   Making a scatterplot of the two residuals and adding the regression line by using plot() and abline().

Code in R:

```
pdf(”scatterplot_q4.pdf”)
plot(residuals_q2,
     residuals_q1,
     xlab=”The residuals from Question 2”,
     ylab=”The residuals from Question 1”,
     main=”The Relationship between the residuals from Question 2 \nand the residuals from Question 1”)

# Adding the regression line
abline(residual_model, col = ”red”)
dev.off()
```

Scatterplot 4 with the regression line.



Conclusion: There is a positive relationship between residuals from Question 2 and residuals from Question 1.

3. Write the prediction equation.

   Getting the coefficients for writing the prediction equation.

   Code in R:

```r
coefficients_res <- residual_model$coefficient
print(coefficients_res)
```

   Output:

```
    (Intercept) st_residuals_q2
-1.941539e-18    2.568770e-01
```

   Writing the prediction equation.

   Code in R:

```r
cat("Prediction Equation:\n residuals_question1 =", coefficients_res[1],
   "+", coefficients_res[2], "* residuals_question2\n")
```

   Output:

```
Prediction Equation:
residuals_question1 = -1.941539e-18 + 0.256877 * residuals_question2
```

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`.

   Executing the regression model in which the vote share of the presidential candidate of the incumbent's party (`voteshare`) is explained by the difference in campaing spending between incumbent and challenger (`difflog`) and the incumbent's electoral success (`presvote`). Then I investigate the estimated coefficients of the model using summary().

   Code in R:

```
model_q5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
model_summary_q4 <- summary(model_q5)
print(model_summary_q4)
```

   Output:

```
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
```

   Conclusion: Controlling the difference in campaing spending between incumbent and challenger, a 1 unit increase in the incumbent's electoral success is associated, on average, with 0.2568 increase in the vote share of the presidential candidate of the incumbent's party. Controlling the incumbent's electoral success, a 1 unit increase in the difference in campaing spending between incumbent and challenger is associated, on average, with 0.0355 increase in the vote share of the presidential candidate of the incumbent's party. The both estimated coefficients is statistically diferentiable from zero at the $\alpha = 0.05$ level because the p-value < 0.05 ($\approx$2e-16).

2. Write the prediction equation.

   Getting the coefficients for writing the prediction equation.

   Code in R:

```r
coefficients_q5 <- model_q5$coefficients
print(coefficients_q5)
```

   Output:

```
(Intercept)     difflog    presvote
0.44864422  0.03554309  0.25687701
```

   Writing the prediction equation.

   Code in R:

```r
cat("Prediction Equation:\n voteshare =", coefficients_q5[1], "+",
    coefficients_q5[2], "* difflog +", coefficients_q5[3], "* presvote\n")
```

   Output:

```
Prediction Equation:
voteshare = 0.4486442 + 0.03554309 * difflog + 0.256877 * presvote
```

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 prediction equation from Question 4 is:

```
residuals_question1 = -1.941539e-18 + 0.256877 * residuals_question2
```

   The prediction equation from Question 5 is:

```
voteshare = 0.4486442 + 0.03554309 * difflog + 0.256877 * presvote
```

The identical part in the two preditiction equations is the estimated coefficient 0.256877. I think this is case because when we do assumption about linear regression we assume that there is an error:

$$\mathrm{Y}_i = \alpha + \beta \mathrm{X}_i + \epsilon_i$$

In Question 4 we have written a prediction equation where:

$$\text{residuals}_1 = -1.941539e - 18 + 0.256877 * \text{residuals}_2 \approx$$

$$\approx 0 + 0.256877 * \text{residuals}_2 = 0.256877 * \text{residuals}_2$$

Having the linear relationship between variables `voteshare` and `difflog` such as:

$$\text{voteshare} = \alpha + \beta * \text{difflog} + \text{residuals}_1$$

which we can rewrite as an equation:

$$\text{voteshare} = \alpha + \beta * \text{difflog} + 0.256877 * \text{residuals}_2$$

Thus, making assupmtion about having error, we include in the regression model the another predictor, in this case `residuals2`, and this predictor is highly correlated with variable `presvote`. We see the same coefficient in the prediction equation from Question 5:

$$\text{voteshare} = 0.4486442 + 0.03554309 * \text{difflog} + 0.256877 * \text{presvote}$$

Checking the correlation coefficient between `presvote` and `residuals2` in R:

Code in R:

```r
cor(inc.sub$presvote, residuals_q2)
```

Output:

```
[1] 0.9550126
```