

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`.

```
1 #1.1 run a regression about voteshare and difflog
2 reg_1 <- lm(voteshare ~ difflog, data=inc.sub)
3 summary(reg_1)
```

The regression about voteshare and difflog:

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

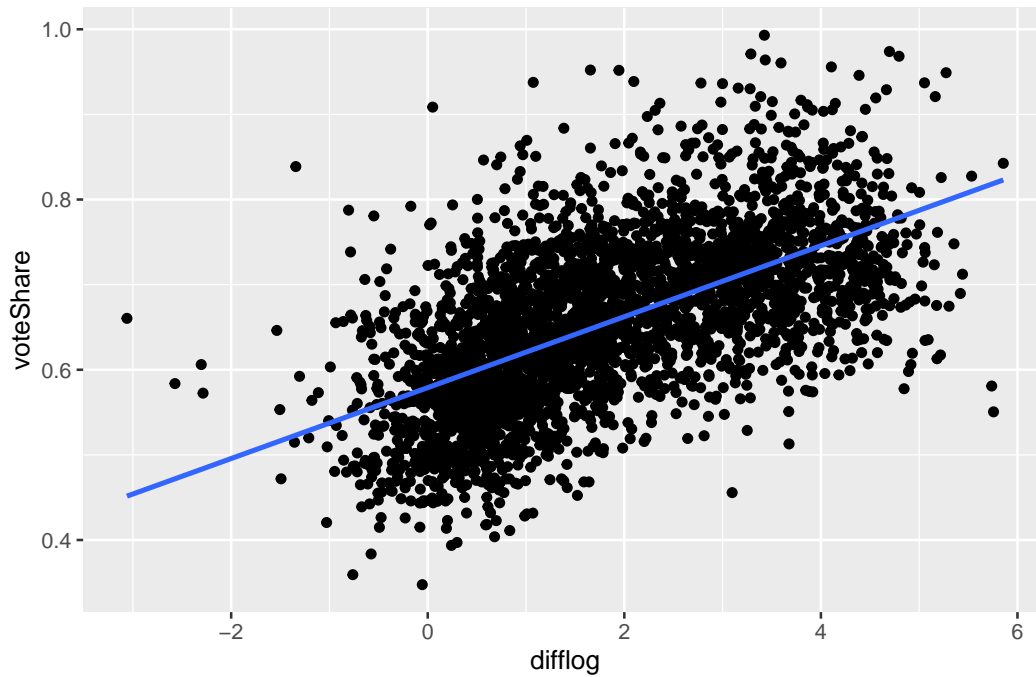
The intercept is estimated to be 0.579031, and the coefficient for difflog is estimated to be 0.041666. Both coefficients are highly statistically significant (p-value < 0.001), indicating a significant relationship between the explanatory variable difflog and the outcome variable voteshare. The R-squared values suggest that approximately 36.73% of the variance in voteshare is explained by the model. The F-statistic is highly significant, indicating that the model as a whole is statistically significant.

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

```
1 #1.2 Make a scatterplot of the two variables
2 library(ggplot2)
3 # Create scatterplot with regression line
4 scatter_1 <- ggplot(data = inc.sub, mapping = aes(x = difflog, y =
  voteshare)) +
5   geom_point() +
6   geom_smooth(method = "lm", se = FALSE) +
7   labs(title = "Scatterplot with Regression Line",
8        x = "difflog",
9        y = "voteShare")
10 # Print the scatterplot
11 print(scatter_1)
```

The scatterplot illustrates the relationship between campaign spending (difflog) and the incumbent's voteshare. The regression line suggests a positive correlation, indicating that as campaign spending increases, the incumbent's vote share tends to increase. The points are tightly clustered around the regression line, suggesting a strong linear relationship. However, a few outliers are noticeable, warranting further investigation.

Figure 1: Scatterplot of relationship between voteshare and difflog.
Scatterplot with Regression Line



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

```
1 #1.3 save the residuals in a separate object
2 residuals_1 <- resid(reg_1)
3 summary(str(residuals_1))
```

Save the residuals in a separate object:

```
Named num [1:3193] -0.000423 -0.031684 -0.004551 0.038669 0.035529 ...
- attr(*, "names")= chr [1:3193] "1" "2" "3" "4" ...
Length Class Mode
0 NULL NULL
```

4. Write the prediction equation.

```
1 #1.4 Write the prediction equation
2 # Extract coefficients
3 coefficients_1 <- coef(reg_1)
4 # Write prediction equation
5 prediction_equation_1 <- paste("voteshare =",
6                               round(coefficients_1[1], 4), "+",
7                               round(coefficients_1[2], 4), " * difflog")
8 # Print prediction equation
9 cat(prediction_equation_1)
```

The prediction equation:

$$\text{voteshare} = 0.579 + 0.0417 * \text{difflog}$$

The intercept of 0.579 indicates that the predicted vote share is 0.579 when the campaign spending difference is zero. The coefficient for difflog (0.0417) predicts a unit increase in difflog is associated with 0.0417 increase in vote share.

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**.

```
1 #2.1 run a regression about presvote and difflog
2 reg_2 <- lm(presvote ~ difflog, data = inc.sub)
3 summary(reg_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

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

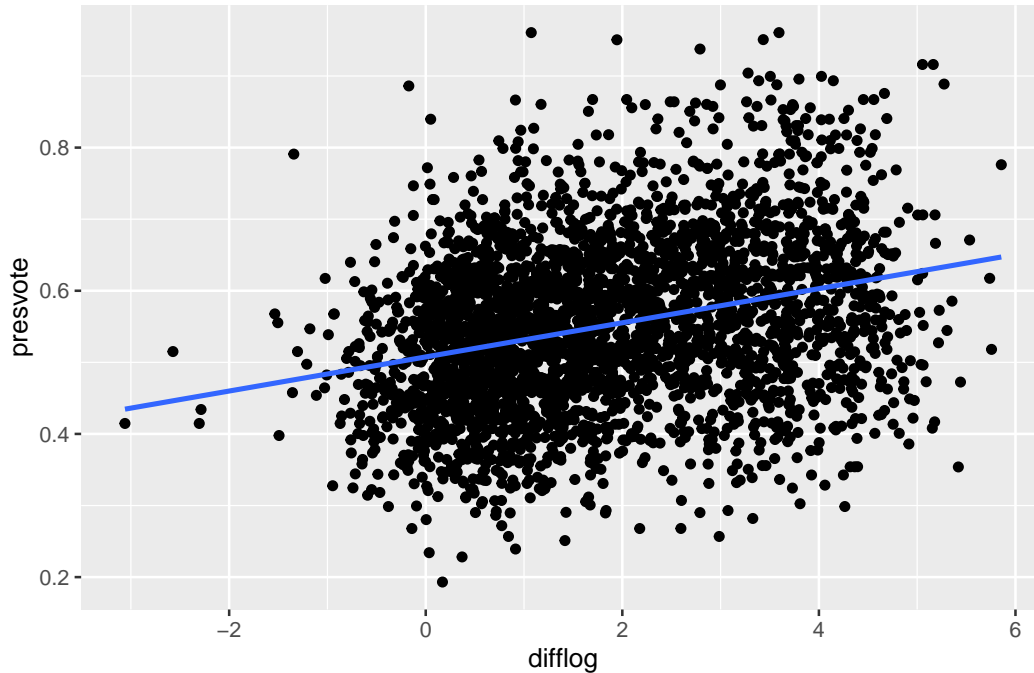
```
1 #2.2 make a scatterplot
2 # Create scatterplot with regression line
3 scatter_2 <- ggplot(data = inc.sub, mapping = aes(x=difflog, y=presvote))+
4   geom_point()+
5   geom_smooth(method = "lm", se = FALSE)+
6   labs(title = "Scatterplot with regression line",
7        x = "difflog",
8        y = "presvote")
```

```

9 # Print the scatterplot
10 print(scatter_2)

```

Figure 2: Scatterplot of relationship between presvotear and difflog.
Scatterplot with regression line



This scatterplot provides a visual representation of the relationship between campaign spending difference (difflog) and presidential vote (presvote). The regression line helps to identify the general trend in the data.

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

```

1 #2.3 save the residuals in a separate object
2 residuals_2 <- resid(reg_2)
3 print(str(residuals_2))

```

Save the residuals in a separate object residuals_2:

```

Named num [1:3193] 0.00561 0.03758 -0.05313 -0.05299 -0.04584 ...
- attr(*, "names")= chr [1:3193] "1" "2" "3" "4" ...
NULL

```

4. Write the prediction equation.

```

1 #2.4 write the prediction equation
2 coefficients_2 <- coef(reg_2)
3 prediction_equation_2 <- paste("presvote=", round(coefficients_2[1], 4), "+",
  round(coefficients_2[2], 4), "*difflog")

```

```
4 cat(prediction_equation_2)
```

The prediction equation between presvote and difflog:

$\text{presvote} = 0.5076 + 0.0238 * \text{difflog}$ The intercept of 0.5076 indicates that the presidential vote is 0.5076 when the campaign spending difference is zero. The coefficient for difflog (0.0238) predicts a unit increase in difflog is associated with 0.0238 increase in presidential vote.

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**.

```
1 #3.1 run a regression about voteshare and presvote
2 reg_3 <- lm(voteshare ~ presvote, data = inc.sub)
3 summary(reg_3)
```

The regression about voteshare and prevote:

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

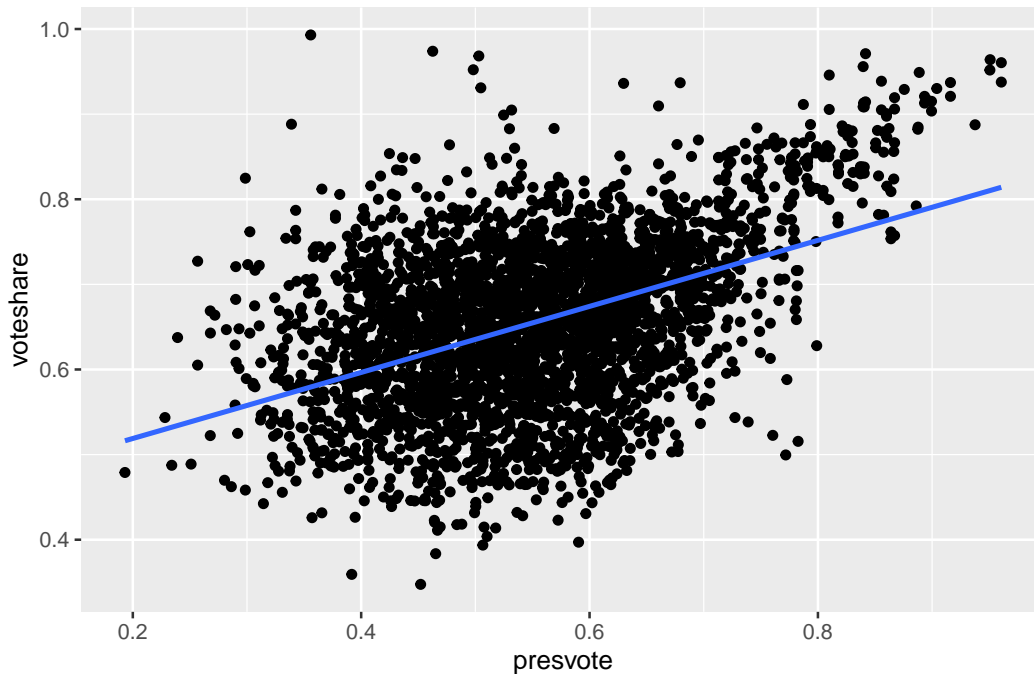
2. Make a scatterplot of the two variables and add the regression line.

```

1 #3.2 make a scatterplot
2 # Create scatterplot with regression line
3 scatter_3 <- ggplot(data = inc.sub, mapping = aes(x = presvote, y =
  voteshare)) +
4   geom_point() +
5   geom_smooth(method = "lm", se = FALSE) +
6   labs(title = "scatterplot with regression line",
7         x = "presvote",
8         y = "voteshare")
9 # Print the scatterplot
10 print(scatter_3)

```

Figure 3: Scatterplot of relationship between voteshare and presvote.
scatterplot with regression line



3. Write the prediction equation.

```

1 #3.3 write the prediction equation
2 coefficients_3 <- coef(reg_3)
3 prediction_equation_3 <- paste("voteshare=", round(coefficients_3[1], 4), "+",
  , round(coefficients_3[2], 4), "*presvote")

```

The prediction equation about voteshare is:
 $\text{voteshare} = 0.4413 + 0.388 \cdot \text{presvote}$

The intercept of 0.4413 indicates that the presidential vote is 0.5076 when the campaign spending difference is zero. The coefficient for `presvote` (0.388) predicts a unit increase in `presvote` is associated with 0.388 increase in `voteshare`.

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.

```
1 #Question 4
2 #4.1 run a regression about residual_1 and residual_3
3 reg_4 <- lm(residuals_1~residuals_2,data = inc.sub)
```

Call: `lm(formula = residuals_1 ~ 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) -5.934e-18 1.299e-03 0.00 1

residuals_12 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: 2.2×10^{-16}

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

```
1
2 #4.2 make a scatterplot
3 # Create scatterplot with regression line
4 scatter_4 <- ggplot(data = inc.sub, mapping = aes(x = residuals_2, y =
  residuals_1)) +
5   geom_point()
```

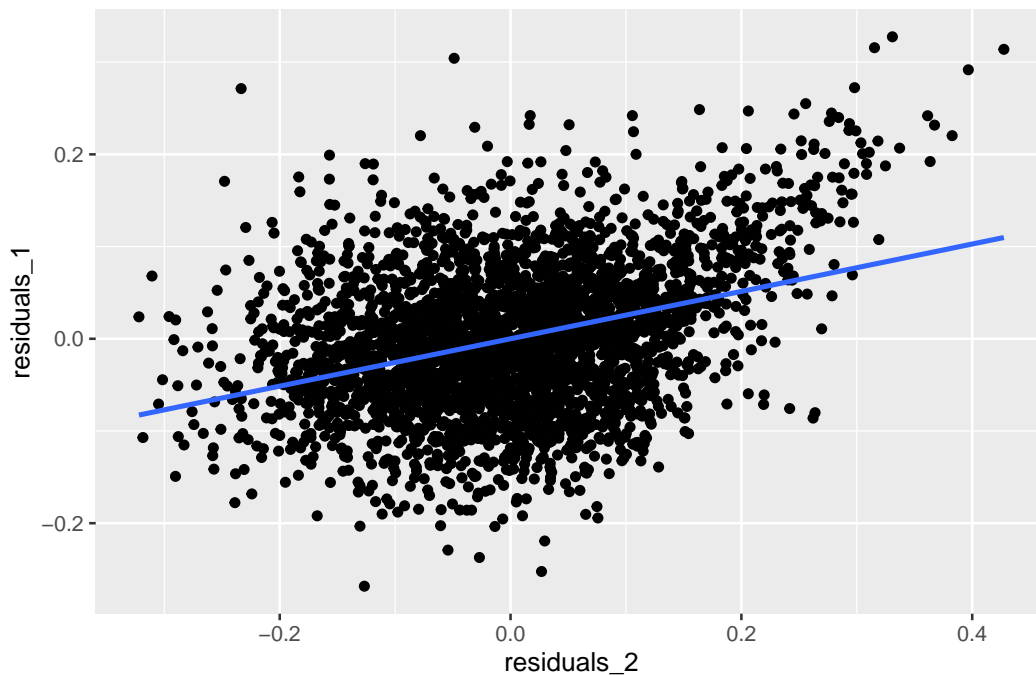


```

6       geom_smooth(method = "lm", se = FALSE)+
7       labs(title = "scatterplot with regression line", x = "
  residuals_2", y = "residuals_1")
8 # Print the scatterplot

```

Figure 4: Scatterplot of relationship between residuals_1 and residuals_2.
scatterplot with regression line



3. Write the prediction equation.

```

1
2 #4.3 write the prediction equation
3 coefficients_4 <- coef(reg_4)
4 prediction_quation_4 <- paste("residuals_1=", round(coefficients_4[1], 4), "
  +", round(coefficients_4[2], 4), "*residuals_2")

```

The prediction equation between residuals_2 and residuals_1:

$$\text{residuals}_1 = 0 + 0.2569 * \text{residuals}_2$$

The intercept of 0 indicates that the presidential vote is 0 when the residuals_2 is zero. The coefficient for residuals_2 (0.2596) predicts a unit increase in residuals_2 is associated with 0.2596 increase in voteshare residuals_1.

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`.

```
1 #Question5
2 #5.1 run a regression about voteshare, difflog and presvote
3 multreg_5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
```

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: 2.2×10^{-16}

2. Write the prediction equation.

```
1
2 #5.2 write the prediction equation
3 coefficients_5 <- coef(multreg_5)
4 prediction_equation_5 <- paste("voteshare=", round(coefficients_5[1], 4), "+",
5                                round(coefficients_5[2], 4), "*difflog", "+",
6                                round(coefficients_5[3], 4), "*presvote")
```

The prediction equation about voteshare is:

$\text{voteshare} = 0.4486 + 0.0355 * \text{difflog} + 0.2569 * \text{presvote}$

The intercept 0.4486 is the predicted voteshare value when both $\text{difflog}=0$ and $\text{presvote}=0$.
The slope of 0.0355 is associated with difflog when controlling for presvote group.
the slope of 0.2569 is associated with presvote when controlling difflog .

The equation can be interpreted as follows: for each one-unit increase in difflog , the expected value of voteshare is expected to increase by 0.0355 and for each one-unit increase in presvote , the expected value of voteshare is expected to increase by 0.2569

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