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

```
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
```

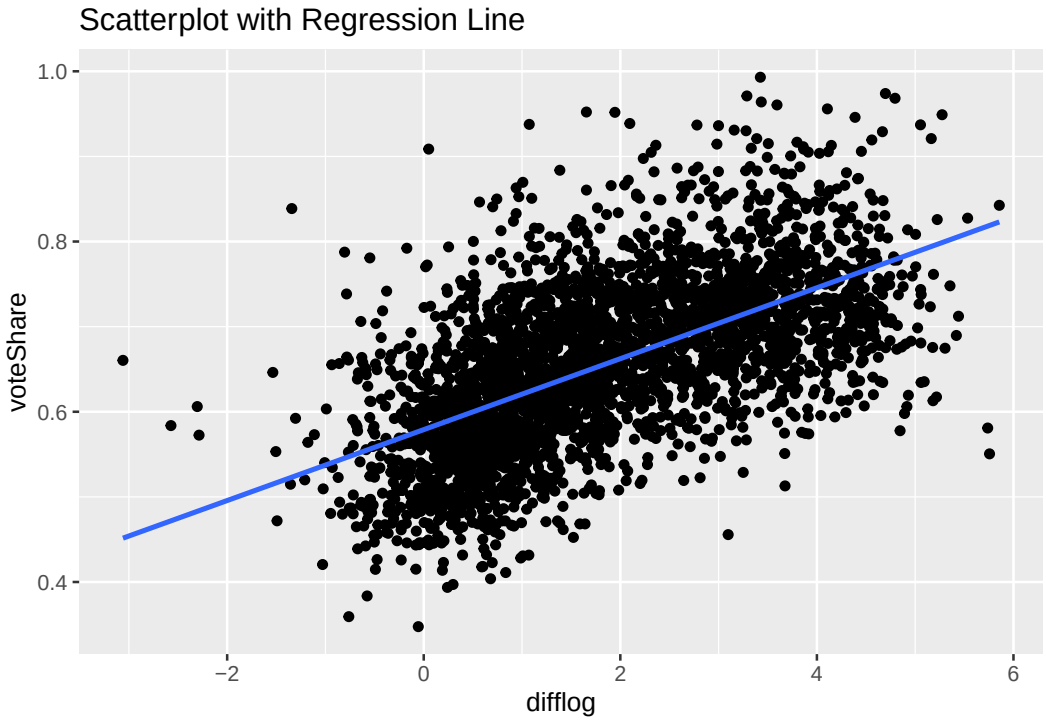
The intercept is estimated to be 0.579031, and the coefficient for difflog is estimated to be 0.041666.Both coefficients are highly statistically significant (p-value ¡ 0.001), indicating a significant relationship between the explanatory variable difflog and the outcome variable voteshare. The R-squared values suggest that approximately 36.73The 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.

```r
#1.2 Make a scatterplot of the two variables
library(ggplot2)
# Create scatterplot with regression line
scatter_1 <- ggplot(data = inc.sub, mapping = aes(x = difflog, y =
    voteshare)) +
            geom_point() +
            geom_smooth(method = "lm", se = FALSE) +
            labs(title = "Scatterplot with Regression Line",
                 x = "difflog",
                 y = "voteShare")
# Print the scatterplot
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`.



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

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

```
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.

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

Thr prediction equation:
voteshare = 0.579 + 0.0417 * 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`.

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

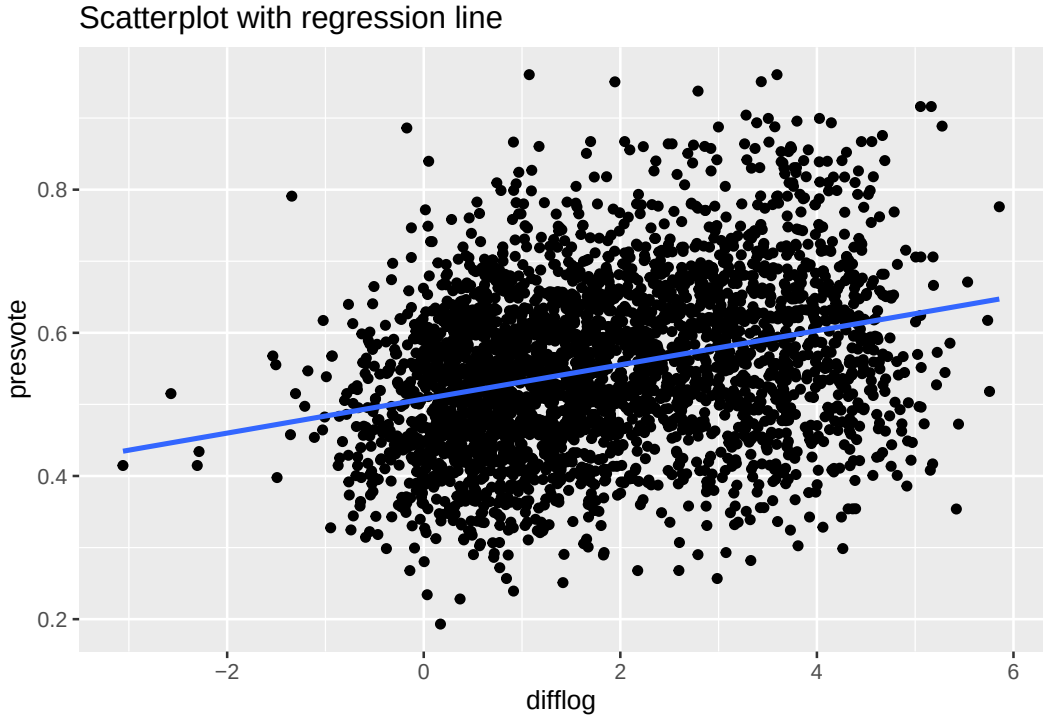
```
Call:
lm(formula = presvote ~ difflog, data = inc.sub)

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.

```r
#2.2 make a scatterplot
# Create scatterplot with regression line
scatter_2 <- ggplot(data = inc.sub,mapping = aes(x=difflog,y=presvote))+
             geom_point()+
             geom_smooth(method = "lm",se = FALSE)+
```

```
            labs(title = "Scatterplot with regression line",
                 x = "difflog",
                 y = "presvote")
# Print the scatterplot
print(scatter_2)
```

Figure 2: Scatterplot of relationship between `presvote`and `difflog`.



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.

```
#2.3 save the residuals in a separate object
residuals_2 <- resid(reg_2)
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.

```r
#2.4 write the prediction equation
cofficients_2 <- coef(reg_2)
prediction_equation_2 <- paste("presvote=",round(cofficients_2[1],4),"+",
    round(cofficients_2[2],4),"*difflog")
cat(prediction_equation_2)
```

The prediction equation between presvote and difflog:
presvote= 0.5076 + 0.0238 *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`.

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

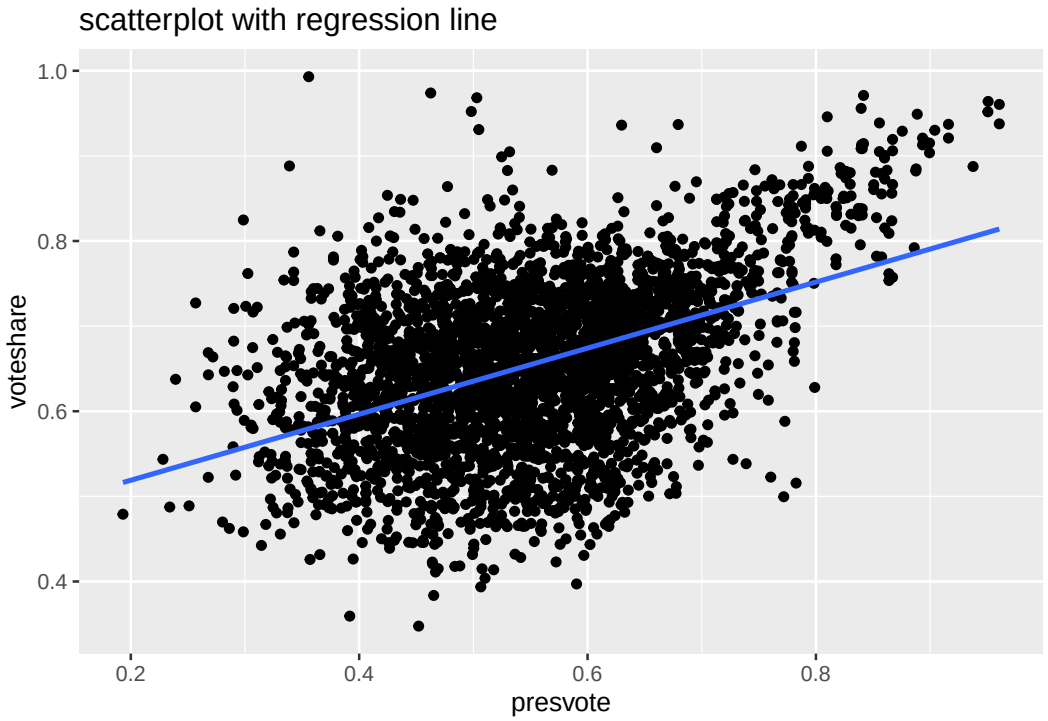
The regression about voteshare and prevote:

```
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.

```
#3.2 make a scatterplot
# Create scatterplot with regression line
scatter_3 <- ggplot(data = inc.sub,mapping = aes(x = presvote,y =
    voteshare))+
              geom_point()+
              geom_smooth(method = "lm",se = FALSE)+
              labs(title = "scatterplot with regression line",
                   x = "presvote",
                   y = "voteshare")
# Print the scatterplot
print(scatter_3)
```

Figure 3: Scatterplot of relationship between `voteshare` and `presvote`.



3. Write the prediction equation.

```
#3.3 write the prediction equation
cofficients_3 <- coef(reg_3)
cat(prediction_equation_3 <- paste("voteshare=",round(cofficients_3[1],4)
    ,"+",round(cofficients_3[2],4),"*presvote"))
```

The prediction equation about voteshare is:
voteshare= 0.4413 + 0.388 *presvote

The intercept of0.4413 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.

```
reg_4 <- lm(residuals_1~residuals_2,data = inc.sub)
summary(reg_4)
```

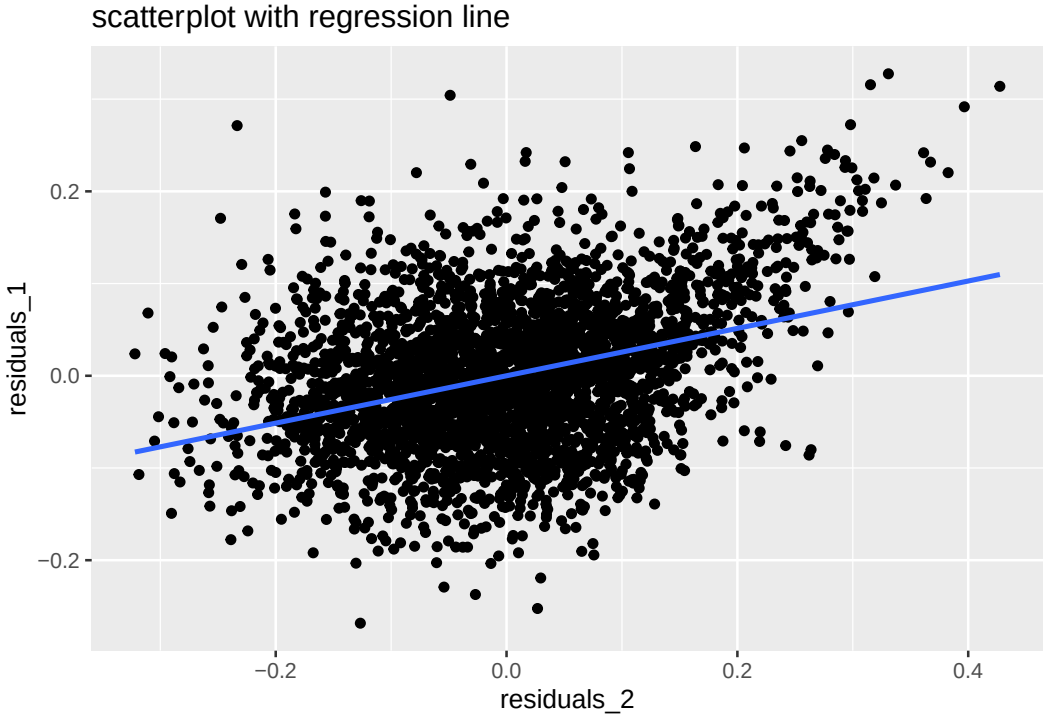
```
Call:
lm(formula = residuals_1 ~ residuals_2, data = inc.sub)

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_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
```

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

```
# Create scatterplot with regression line
scatter_4 <- ggplot(data = inc.sub,mapping = aes(x = residuals_2,y =
    residuals_1))+
              geom_point()+
              geom_smooth(method = "lm",se = FALSE)+
```

```
            labs(title = "scatterplot with regression line",x = "
    residuals_2",y = "residuals_1")
# Print the scatterplot
print(scatter_4)
```

Figure 4: Scatterplot of relationship between `residuals_1` and `residuals_2`.



3. Write the prediction equation.

```
coefficients_4 <- coef(reg_4)
prediction_quation_4 <- paste("residuals_1=",round(coefficients_4[1],4),"
    +",round(coefficients_4[2],4),"*residuals_2")
cat(prediction_quation_4)
```

The prediction equation between residuals_2 and residuals_1:
residuals_1= 0 + 0.2569 *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`.

```
multreg_5 <- lm(voteshare~difflog+presvote,data = inc.sub)
summary(multreg_5)
```

```
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
```

2. Write the prediction equation.

```
coefficients_5 <- coef(multreg_5)
prediction_equation_5 <- paste("voteshare=",round(coefficients_5[1],4),"+
    ",
                               round(coefficients_5[2],4),"*difflog","+",
                               round(coefficients_5[3],4),"*presvote")
cat(prediction_equation_5)
```

The prediction equation about voteshare is:
voteshare= 0.4486 + 0.0355 *difflog + 0.2569 *presvote

The intercept 0.4486 is the predicted voteshare value when both difflog=0 and 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?
   Both models have a similar coefficient for the variable of interest (residuals2 in Model 1 and presvote in Model 2). This may suggest a similarity in their impact on the respective dependent variables. The statistical significance of coefficients is determined by the p-values ($\Pr(>|t|)$). In both models, the coefficients have highly significant p-values (<2e-16), indicating their significance.
   It could be a coincidence that the variable names are similar. If these variables represent different aspects of the data but coincidentally have similar coefficients, the output would reflect this similarity.