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 in data
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")
#1. regression:
model <- lm(voteshare ~ difflog, data = inc.sub)
summary(model)
```

```
Call:lm(formula = voteshare ~ difflog, data = inc.sub)
Residuals:     Min       1Q   Median       3Q      Max
-0.26832 -0.05345 -0.00377  0.04780  0.32749
```

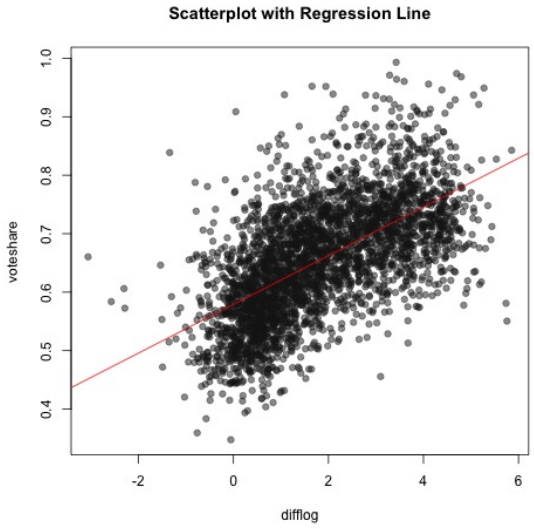
```
Coefficients:             Estimate Std. Error t value Pr(>|t|)
  (Intercept) 0.579031   0.002251  257.19   <2e-16 ***
  difflog     0.041666   0.000968   43.04   <2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  Residual standard error: 0.07867 on 3191 degrees of freedom
  Multiple R-squared:  0.3673,Adjusted R-squared:  0.3671
  F-statistic:  1853 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```r
# Scatterplot with regression line

jpeg(file='voteshare_difflog.jpeg')
plot(inc.sub$difflog, inc.sub$voteshare, main = "Scatterplot with
    Regression Line", xlab = "difflog", ylab = "voteshare", pch = 19, col
    = rgb(0.1, 0.1, 0.1, 0.5))
abline(model, col = "red")
dev.off()
```

Figure 1: scatter plot with line.



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

```r
residuals_voteshare <- resid(model)
```

4. Write the prediction equation. voteshare=intercept+slope ×difflog

```r
#prediction
intercept <- coef(model)[1]
slope <- coef(model)[2]
#prediciton: voteshare = intercept + slope * difflog
paste("voteshare =", intercept, "+", slope, "* difflog ")
```

voteshare = 0.579030710920674 + 0.0416663238227399 * 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`.

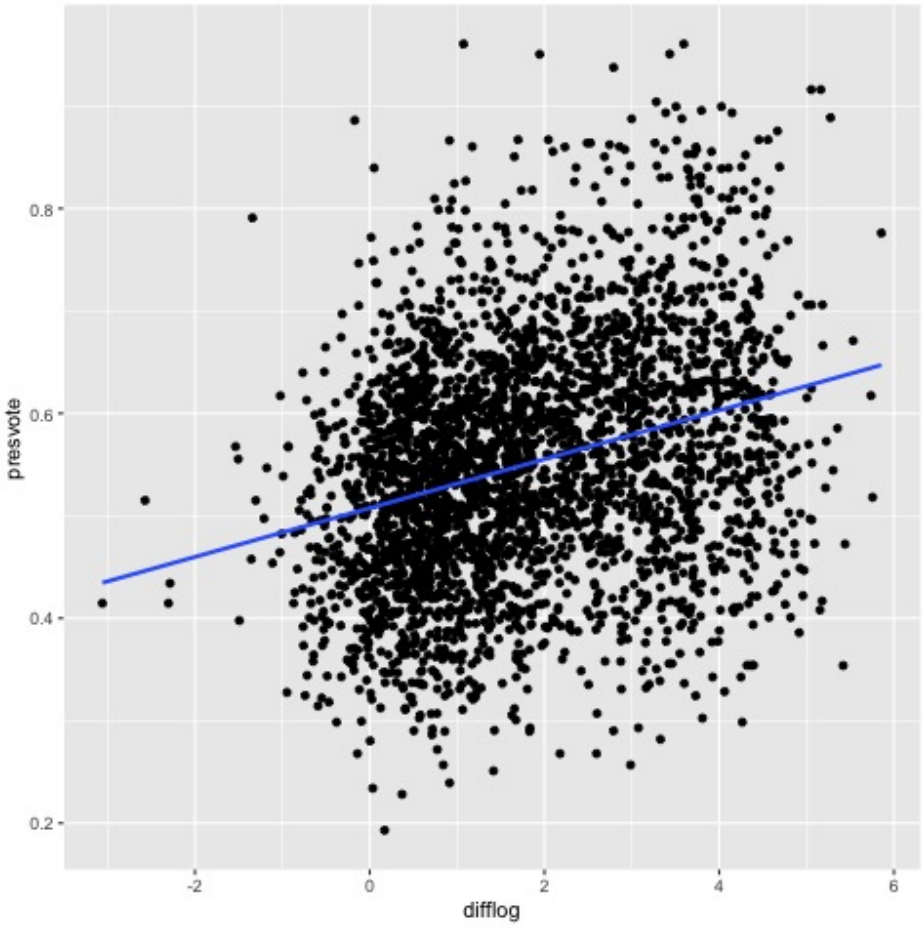
```
#q2:
lm_model_presvote <- lm(presvote ~ difflog, data = inc.sub)
summary(lm_model_presvote)
```

```
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.

```
jpeg(file='presvote_difflog.jpeg')
ggplot(inc.sub, aes(x = difflog, y = presvote)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
dev.off()
```

Figure 2: scatter plot presvote and difflog.



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

```r
residuals_presvote <- resid(lm_model_presvote)
```

4. Write the prediction equation.

```r
#prediciton: presvote = intercept + slope * difflog
intercept2 <- coef(lm_model_presvote)[1]
slope2 <- coef(lm_model_presvote)[2]
paste("presvote =", intercept2, "+", slope2, "* difflog ")
```

"presvote = 0.507583328405015 + 0.023837233841334 * 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`.

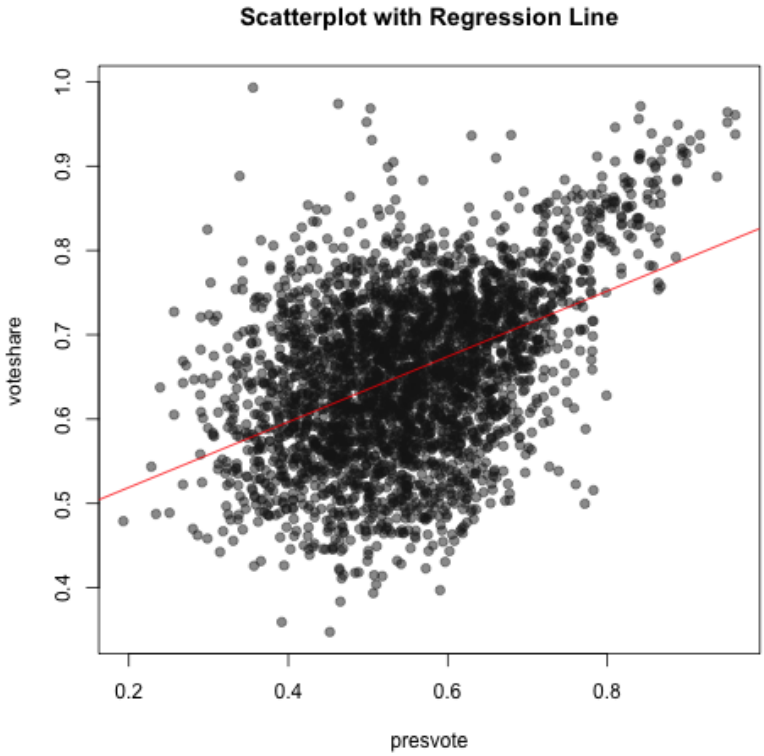
```r
#q3
model_voteshare <- lm(voteshare ~ presvote, data = inc.sub)
summary(model_voteshare)
```

```
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.

```r
# Save the plot
png(filename = "scatterplot_voteshare_presvote.png")
plot(inc.sub$presvote, inc.sub$voteshare, main = "Scatterplot with
    Regression Line", xlab = "presvote", ylab = "voteshare", pch = 19, col
    = rgb(0.1, 0.1, 0.1, 0.5))
abline(model_voteshare, col = "red")
dev.off()
```

Figure 3: scatter plot presvote and voteshare.



3. Write the prediction equation.

```
# Getting coefficients
intercept_voteshare <- coef(model_voteshare)[1]
slope_voteshare <- coef(model_voteshare)[2]

# Prediction equation
paste("voteshare =", intercept_voteshare, "+", slope_voteshare, "*
    presvote")
```

"voteshare = 0.441329881204298 + 0.38801844338744 * 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.

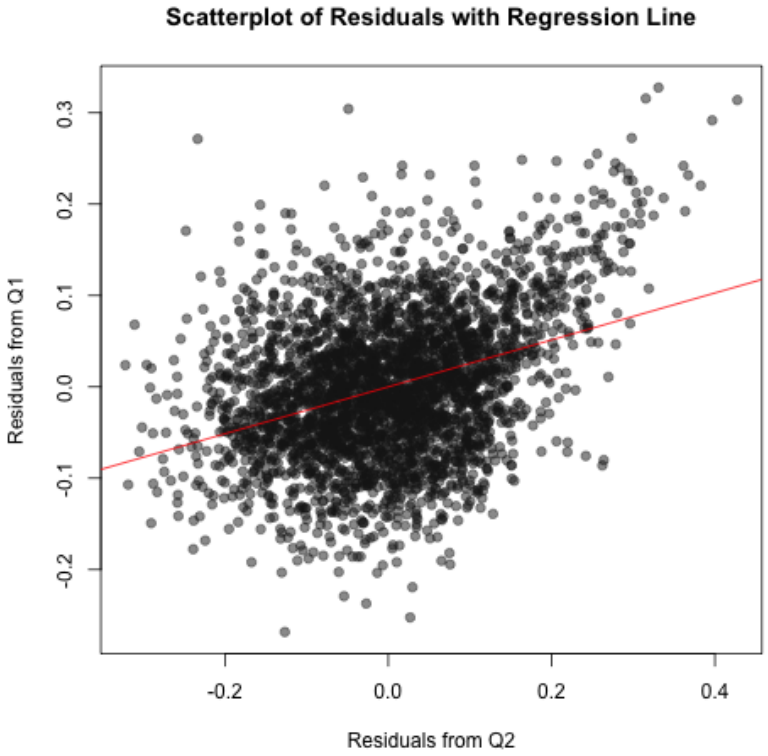
```
#q4
# Fit the linear model with residuals
model_residuals <- lm(residuals_voteshare ~ residuals_presvote, data =
    inc.sub)
summary(model_residuals)
```

```
Call:lm(formula = residuals_voteshare ~ residuals_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)
  -1.942e-18  1.299e-03    0.00   1
 residuals_presvote  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.

```
# Save the plot
png(filename = "scatterplot_residuals_q1_q2.png")
plot(residuals_presvote, residuals_voteshare, main = "Scatterplot of
    Residuals with Regression Line", xlab = "Residuals from Q2", ylab = "
    Residuals from Q1", pch = 19, col = rgb(0.1, 0.1, 0.1, 0.5))
abline(model_residuals, col = "red")
dev.off()
```

Figure 4: scatterplot two residuals.



3. Write the prediction equation.

```
intercept_residuals <- coef(model_residuals)[1]
slope_residuals <- coef(model_residuals)[2]

# Prediction equation for residuals
paste("Residuals_Q1(voteshare) =", intercept_residuals, "+", slope_
    residuals, "* Residuals_Q2 (presvote)")
#q5
```

```
"Residuals_Q1(voteshare) = -1.94153862344556e-18 +
0.256877012700097 * Residuals_Q2 (presvote)"
```

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`.

```
# Fit the multiple linear regression model
model_voteshare_multi <- lm(voteshare ~ difflog + presvote, data = inc.sub)
summary(model_voteshare_multi)
# Getting coefficients for multiple regression
```

```
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.

```
intercept_multi <- coef(model_voteshare_multi)[1]
slope_difflog <- coef(model_voteshare_multi)[2]
slope_presvote <- coef(model_voteshare_multi)[3]

# Prediction equation for multiple regression
paste("voteshare =", intercept_multi, "+", slope_difflog, "* difflog +", slope_presvote, "* presvote")
```

"voteshare = 0.448644221823622 + 0.0355430864025444 * difflog + 0.256877012700098 * presvote"

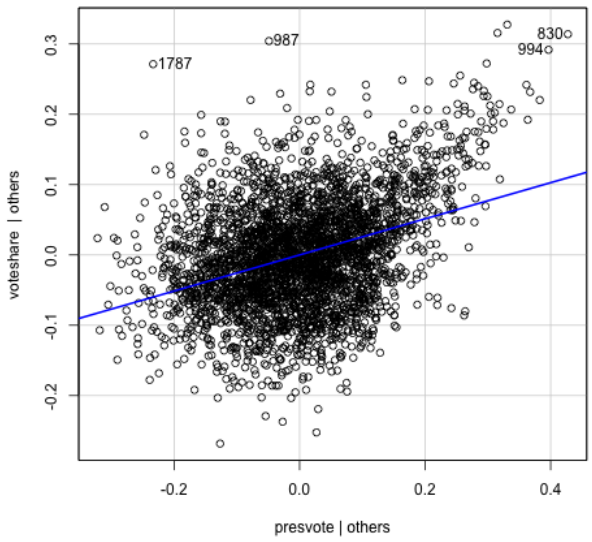


Figure 5: added variable plot

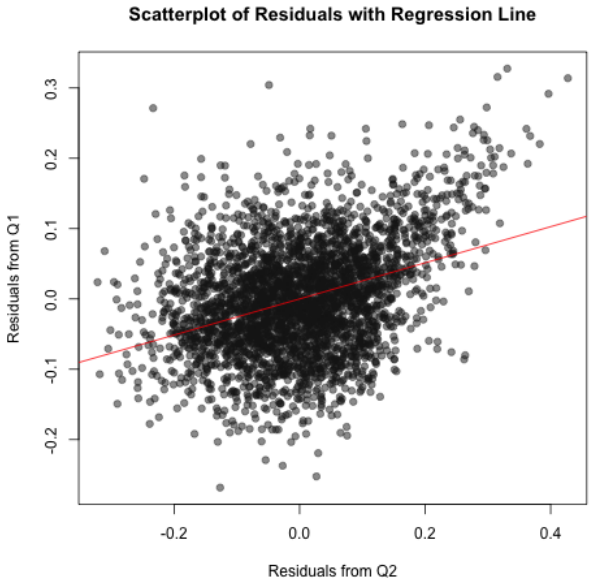


Figure 6: residuals scatterplot

Figure 7: comparison

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 coefficient for presvote is in question 5 is same as the coefficient for residuals(presvote) in the model with residuals in q4. To interpret, this coefficient in q4 context tells us with one unit increase in the residuals from presvore-difflog regression, which is the variation not explained by difference in spending (difflog) with regard to presvote, there is about on average 0.26 unit increase in the residuals from votshare-difflog regression model, which is variation not explained by difflog with regard to voteshare. In other words, if we remove the effect of difflog (holding it constant), one unit increase in the presvote will increase 0.26 unit of voteshare on average, which is also what the coeeficient for presvote in question 5 tells us. To illustrate the point, we can also see a side by side comparison between the added variable plot of presvote to voteshare conditioned on difflog and the scatter plot of residuals and conclude they are showing the same thing. Based on this observation and since the pvalue is very small so it's significant, we can conclude that presvote is a variable that can explain certain variation in voteshare in addition to difflog, so we should probably keep it in our model so as to not commit omitted variable bias. Hence we found evidence that support the incumbent's vote share is affected by both the president's popularity and the difference in spending between incumbent and challenger.