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.

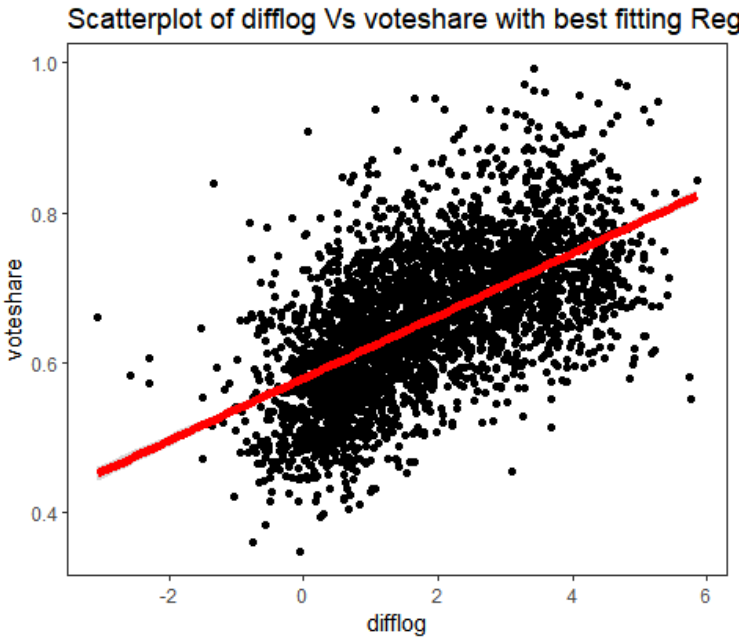
1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `difflog`.

```
fit_reg1=lm(voteshare ~ difflog, data=data_ps3)
summary(fit_reg1)
###
#### Answer/ R_code output ##
###
#Call:
#lm(formula = voteshare ~ difflog, data = data_ps3)

#Residuals:
#    Min      1Q  Median      3Q     Max
#-0.26832 -0.05345 -0.00377  0.04780  0.32749
```

```
12
13 #Coefficients:
14 #  Estimate Std. Error t value Pr(>|t|)
15 #(Intercept) 0.579031   0.002251  257.19   <2e-16 ***
16 #  difflog     0.041666   0.000968   43.04   <2e-16 ***
17 #  ---
18 #  Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
     0.1          1
19
20 #Residual standard error: 0.07867 on 3191 degrees of freedom
21 #Multiple R-squared:  0.3673, Adjusted R-squared:  0.3671
22 #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.

```
1 # Method: Using ggplot2
2
3 ggplot(data_ps3, aes(x=difflog, y=voteshare)) +
4   geom_point() +
5   geom_jitter() +
6   geom_smooth(method="lm", formula=y ~ x, se=T, color="red", lwd=2) +
7   theme_bw() + theme(panel.grid=element_blank()) +
8   ggtitle("Scatterplot of difflog Vs voteshare with best fitting
      Regression Line")
```



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

```
# 1.3 Save the residuals of the model in a separated object:

my_residual_q1 <-residuals(fit_reg1)
head(my_residual_q1) # The first 6 observations of residuals:
# Answer:
#          1             2             3             4             5
          6
#-0.0004227622 -0.0316840149 -0.0045514943  0.0386688767  0.0355287965
    0.0322832521
##
tail(my_residual_q1) # The last 6 observations of residuals:
### Answer:
#     3188          3189          3190          3191          3192
    3193
# 0.018604721  0.048283877  0.023159323 -0.040639860 -0.065834625
    0.007829042
```

4. Write the prediction equation.

```
######  1.4 Write the Predicted Equation:

# y_hat= Beta_0_hat + Beta_1_hat* difflog imply that voteshareˆ =
    0.579031 + 0.041666*difflog_i
##
# Here We can see clearly that the estimated slope Beta1_hat=0.041666,
    means that for every one unit increase in the
# difference in campaign spending in favor of the incumbent, the
    estimated vote share for the incumbent is expected
# to increase on average by 0.041666.
##
#The estimated y-intercept or beta0_hat=0.579031 means when the spending
    difference is zero (difflog=0) the estimated
#vote share(vote_share) for the incumbent is equal to 0.579031.
###
# Note: This implies that on average, an increase in campaign spending by
     the incumbent compared to the challenger is
# associated with an increase in the incumbent's vote share.
```

Question 2

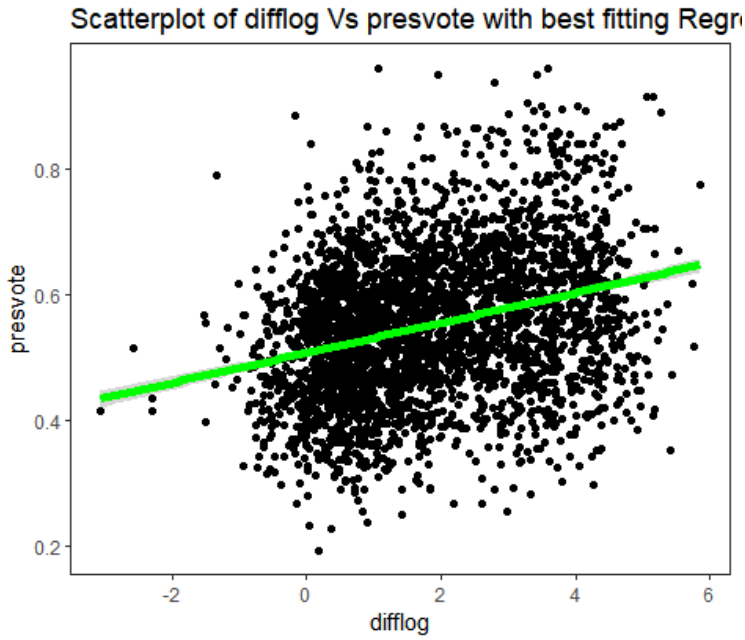
We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

1. Run a regression where the outcome variable is `presvote` and the explanatory variable is `difflog`.

```
# Question 2.1 Run a regression where the outcome variable is presvote and the explanatory variable is difflog.
### Answer:
fit_reg2=lm(presvote ~ difflog, data=data_ps3)
summary(fit_reg2)
#### Answer/ R_code output ##
#####
#Call:
#lm(formula = presvote ~ difflog, data = data_ps3)

#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.

```
#Question 2.2 Make a scatterplot of the two variables and add the regression line
### Answer:
# Method: Using ggplot2

ggplot(data_ps3, aes(x=difflog, y=presvote)) +
  geom_point() +
  geom_jitter() + # I use the jitter() function to avoid overlapping the points
  geom_smooth(method="lm", formula=y ~ x, se=T, color="green", lwd=2) +
  theme_bw() + theme(panel.grid=element_blank()) +
  ggtitle("Scatterplot of difflog Vs presvote with best fitting Regression Line")
```



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

```
# Question 2. 3 Save a Residuals of the model in a separate object.
###
my_residual_q2 <-residuals(fit_reg2)
head(my_residual_q2) # The first 6 observations of residuals:
### Answer/R-Output of Residuals

#      1             2             3             4             5
            6
# 0.005605594    0.037578519   -0.053134788   -0.052993694
    -0.045842994    0.074339701
#######
# The Last six observations of the residuals models:
###
tail(my_residual_q2)
## Answer/ R output
##
```

```
#      3188           3189           3190           3191           3192
            3193
# -0.017727276   -0.033198949   -0.002119851    0.032545042
   0.036938994    0.035795200
```

4. Write the prediction equation.

```
# Question 2. 4 Write the prediction equation.
###
### Answer:
##
# presvote_hat=0.507583 + 0.023837*difflog
##
# The estimated slope beta1_hat=0.023837 means that for every one unit
    increase in the difference in campaign spending
#in favor of the incumbent, the estimated vote share for the presidential
     candidate of the incumbent's party is
# expected to increase on average by 0.023837.
####
#The estimated y-intercept=beta0_hat=0.507583 means when the spending
    difference is zero (i.e., difflog=0), the
# estimated vote share for the presidential candidate of the incumbent's
    party (presvote = 0.507583).
###
# Note: This implies that, on average, an increase in campaign spending
    by the incumbent compared to the challenger
# is associated with an increase in the vote share of the incumbent's
    party's presidential candidate.
```

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.

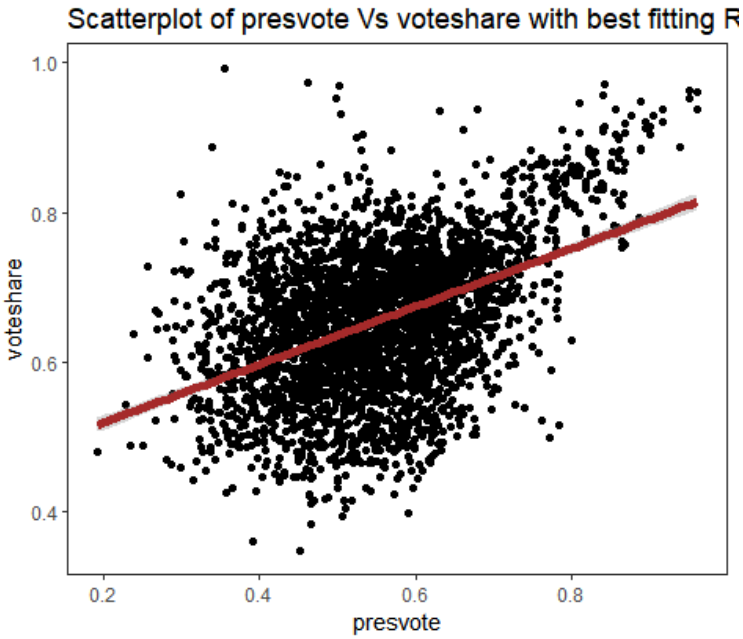
1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`.

```
# Question 3. 1 Run a regression where the outcome variable is voteshare
    and the explanaatory variable is presvote.
###
fit_reg3=lm(voteshare ~ presvote, data=data_ps3)
summary(fit_reg3)
# Answer / R Outcome ###
##
#Call:
#lm(formula = voteshare ~ presvote, data = data_ps3)

#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.

```
# Question 3.2 Make a scatterplot of the two variables and add the
    regression line.
###
## Answer/ R -Output.
#
### Answer:
# Method: Using ggplot2

ggplot(data_ps3, aes(x=presvote, y=voteshare)) +
  geom_point() +
  geom_jitter() +
  geom_smooth(method="lm", formula=y ~ x, se=T, color="brown", lwd=2) +
  theme_bw() + theme(panel.grid=element_blank()) +
```

```
  ggtitle("Scatterplot of presvote Vs voteshare with best fitting
    Regression Line")
####
```



3. Write the prediction equation.

```
### Question 3.3 Write the prediction equation
###
# Answer:

# voteshare_hat= 0.441330 + 0.388018*presvote
###
# The estimated slope=beta1_hat=0.388018 means for every one unit
    increase in the vote share of the presidential
#candidate of the incumbent's party, the estimated vote share for the
    incumbent is expected to increase on average
# by 0.388018
####
# The y-intercept=beta0_hat=0.441330 means when the vote share of the
    presidential candidate is zero (presvote=0),
# the estimate vote share for the incumbent is equal to 0.441330.
```

Question 4

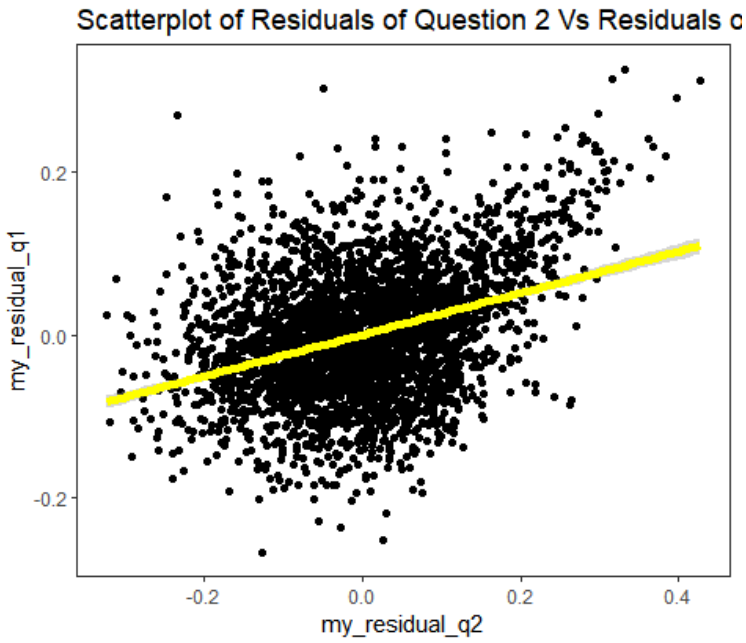
The residuals from part (a) tell us how much of the variation in `voteshare` is *not* explained by the difference in spending between incumbent and challenger. The residuals in part (b) tell us how much of the variation in `presvote` is *not* explained by the difference in spending between incumbent 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.

```
### Answer:
####
# recall both residuals:
my_residual_q1 # Residual from Question 1 (Outcome variable)
my_residual_q2 # Residual from Question 2 (Explanatory variable)
#####
fit_reg4 = lm(my_residual_q1 ~  my_residual_q2, data= data_ps3 )
summary(fit_reg4)
##### Answer/R Output
###
# Call:
#lm(formula = my_residual_q1 ~ my_residual_q2, data = data_ps3)

#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
# my_residual_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
```

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

```
###Question 4.2 Make a scatterplot of the two residuals and add the
    regression line:
###
# Method: ggplot2
#
# Method: Using ggplot2

ggplot(data_ps3, aes(x=my_residual_q2, y=my_residual_q1)) +
  geom_point() +
  geom_jitter() +
  geom_smooth(method="lm", formula=y ~ x, se=T, color="yellow", lwd=2) +
  theme_bw() + theme(panel.grid=element_blank()) +
  ggtitle("Scatterplot of Residuals of Question 2 Vs Residuals of
    Question 1 with best fitting Regression Line")
```



3. Write the prediction equation.

```
# Question 4. 3 Write the prediction equation:
#### Answer:
##
# my_residual_q1_hat = -5.934e**(-18) + 2.569e**(-1)*my_residual_q2
####
# The estimated slope=beta1_hat=2.569e**(-01) means that for each
    increase of one unit in residual of question 2
# (my_residual_q2), the value of residual of question 1 (my_residual_q1)
    is expected to increase by an average of
# 2.569e**(-1) units.
###
```

```
# The y-intercept=beta0_hat=-5.934e**(-18) represent the value when
    residual of question 2 (i.e., my_residual_q2=0).
```

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_reg5 =lm(voteshare ~ difflog + presvote, data=data_ps3)
summary(fit_reg5)
######
# Answer/ RStudio Output
####
#Call:
#lm(formula = voteshare ~ difflog + presvote, data = data_ps3)

#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.

```
# voteshare_hat = 0.4486442 + 0.0355431*difflog + 0.2568770*presvote
#
#The y-intercept=beta0_hat=0.4486442 represent the estimate vote share
    for the incumbent when both variables difflog
# and presvote are zero (i.e., difflog=presvote=0).In this context it
    represents the expected baseline vote share for
# the incumbent##
##
#The coefficient for difflog (beta1_hat=0.0355431) means that for every
    one unit increase in the difference in
# spending in favor of the of the incumbent the estimated vote for the
    incumbent is expected to increase on average
# by 0.0355431 assuming the president's popularity remainn constant.
```

```
######
# The coefficient for presvote (beta2_hat=0.2568770) means for every one unit increase in the president's popularity,
#the estimate vote share for the incumbent is expected to increase by 0.2568770, assuming the difference in spending
# remain constant.
```

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

```
# Here the output of the both residuals are equal:
##Residuals of question 4:
#  Min        1Q    Median        3Q       Max
# -0.25928 -0.04737 -0.00121  0.04618  0.33126
###
##Residuals of Question 5:
#  Min        1Q    Median        3Q       Max
#-0.25928 -0.04737 -0.00121  0.04618  0.33126
##
#One of the possible explanation is related to perfect multicollinearity between the independent variables presvote
# and difflog that are perfectly correlated. On the other hand the difflog variable it's appeared in both models,
# this could lead to identical residuals.
```