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

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

Residuals:
Min       1Q   Median       3Q      Max
```

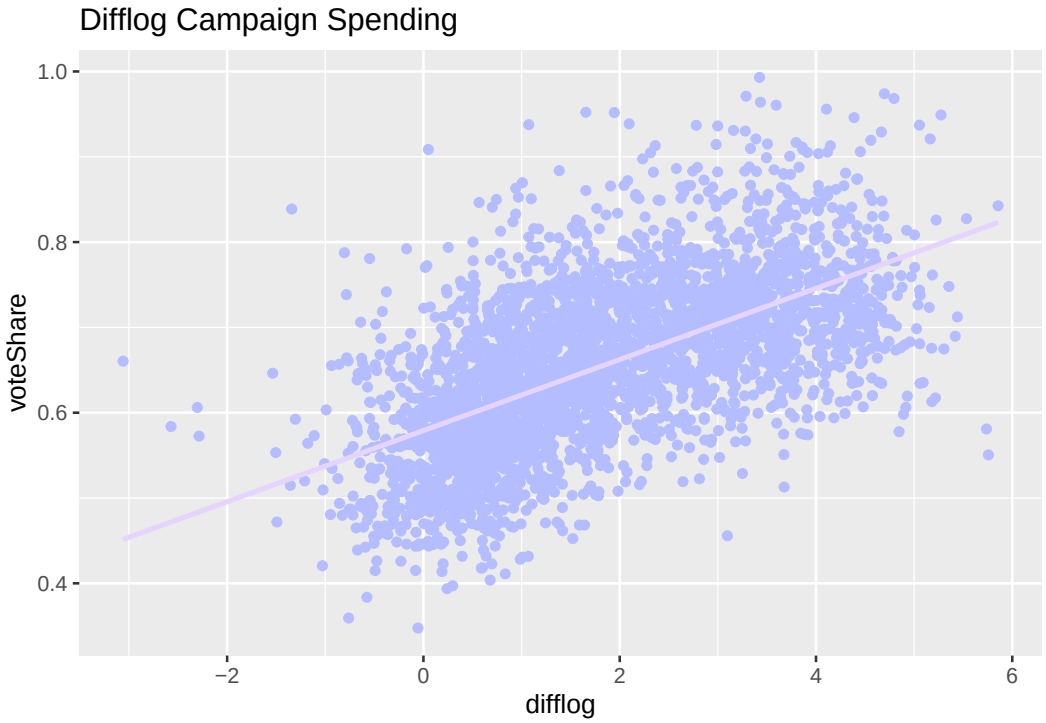
```
-0.26832 -0.05345 -0.00377  0.04780  0.32749

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

According to the results, the F statistic is 1853,
the degrees of freedom are 1 and 3191,
and the p-value is less than 2.2e-16,
indicating that the overall model is highly significant
and we have high confidence to reject the null hypothesis.
```

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

```r
# Create scatterplot with regression line and customized colors
ggplot(data = inc.sub, aes(x = difflog, y = voteshare)) +
geom_point(color = "#B4BDFF") +
#Add scatterplot points
geom_smooth(method = "lm", se = FALSE, color = "#E5D4FF") +
  #Add regression line
  labs(title = "Difflog Campaign Spending",
       x = "difflog",
       y = "voteShare")
```

```
According to the results, the linear correlation is positive,
indicating that as campaign spending increases,
the incumbent's vote share tends to increase.
Moreover, the dotted plots are concentrated on both sides of the linear plot,
indicating that there is a strong linear relationship.
```

Figure 1: Scatterplot of relationship between `voteshare` and `difflog`.



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

```r
#save the residuals in a separate object
res_1 <- residuals(reg_model)
res_sep_1 <- str(res_1)
summary(res_sep_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.

```r
# Extracting the intercept and slope
intercept <- round(coef(reg_model)[1],4)
slope <- round(coef(reg_model)[2],4)

# Printing the intercept and slope
cat("Final Model: votehare=", intercept, "+", slope, "* difflog\n")
```

```
Final Model: votehare= 0.579 + 0.0417 * difflog

According results,
For each one-unit increase in difflog,
we expect voteshare to increase by approximately 0.0417,
assuming all other factors remain constant.
The intercept of 0.579 represents the estimated voteshare when difflog is zero.
```

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`.

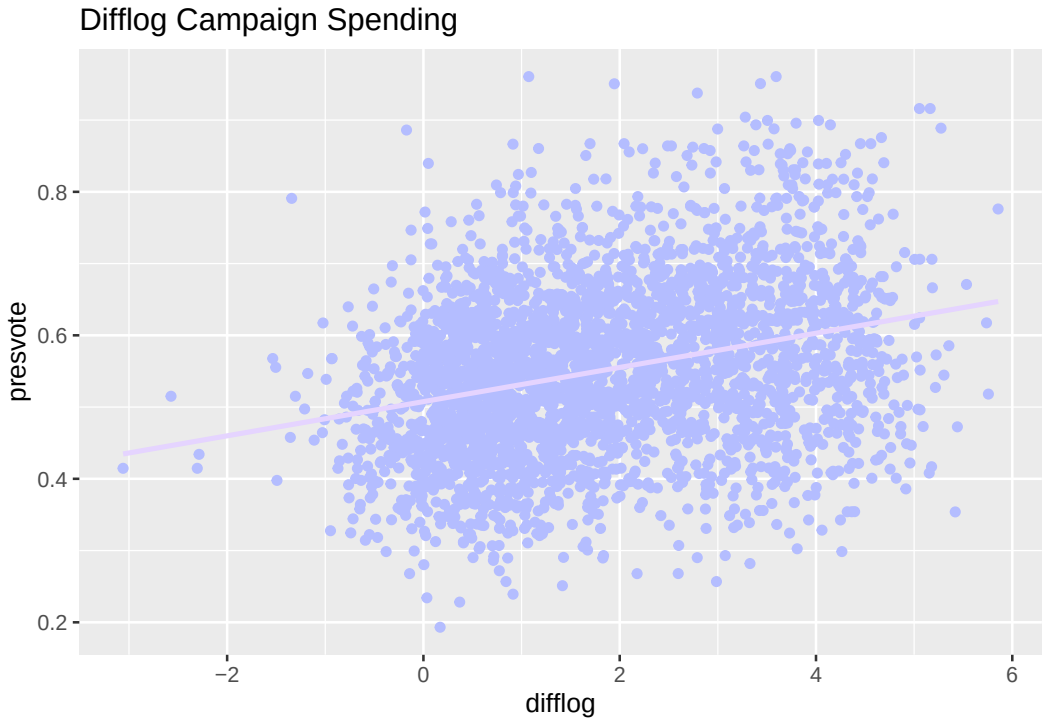
```r
#Creat a regression about presvote and difflog
reg_model_2 <- lm(presvote~difflog,data = inc.sub)
summary(reg_model_2)
```

```
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

According to the results, the F statistic is 307.7,
the degrees of freedom are 1 and 3191,
and the p-value is less than 2.2e-16,
indicating that the overall model is highly significant
and has high confidence to reject the null hypothesis.
```

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

```
# Create scatterplot with regression line and customized colors
ggplot(data = inc.sub, aes(x = difflog, y = presvote)) +
  geom_point(color = "#B4BDFF") +
  #Add scatterplot points
  geom_smooth(method = "lm", se = FALSE, color = "#E5D4FF") +
  #Add regression line
  labs(title = "Difflog Campaign Spending",
       x = "difflog",
       y = "presvote")
```

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




```
According to the results, the linear correlation is positive,
indicating that campaign spending increases,
the presvote tends to increase.
Moreover, the dotted plots are concentrated on both sides of the linear plot,
indicating that there is a strong linear relationship.
```

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

```
#save the residuals in a separate object
res_2 <- residuals(reg_model_2)
res_sep_2 <- str(res_2)
summary(res_sep_2)
```

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

4. Write the prediction equation.

```r
# Extracting the intercept and slope
intercept <- round(coef(reg_model_2)[1],4)
slope <- round(coef(reg_model_2)[2],4)

# Printing the intercept and slope
cat("Final Model: presvote=", intercept, "+", slope, "* difflog\n")
```

```
Final Model: presvote= 0.5076 + 0.0238 * difflog

Based on the results, for each unit increase in the difflog variable,
we expect the pre-voting to increase by approximately 0.0238.
The intercept of 0.5076 represents the estimated baseline value of presvote
when there is no change in the difflog variable.
```

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
#Creat a regression about presvote and difflog
reg_model_3 <- lm(voteshare~presvote,data = inc.sub)
summary(reg_model_3)
```

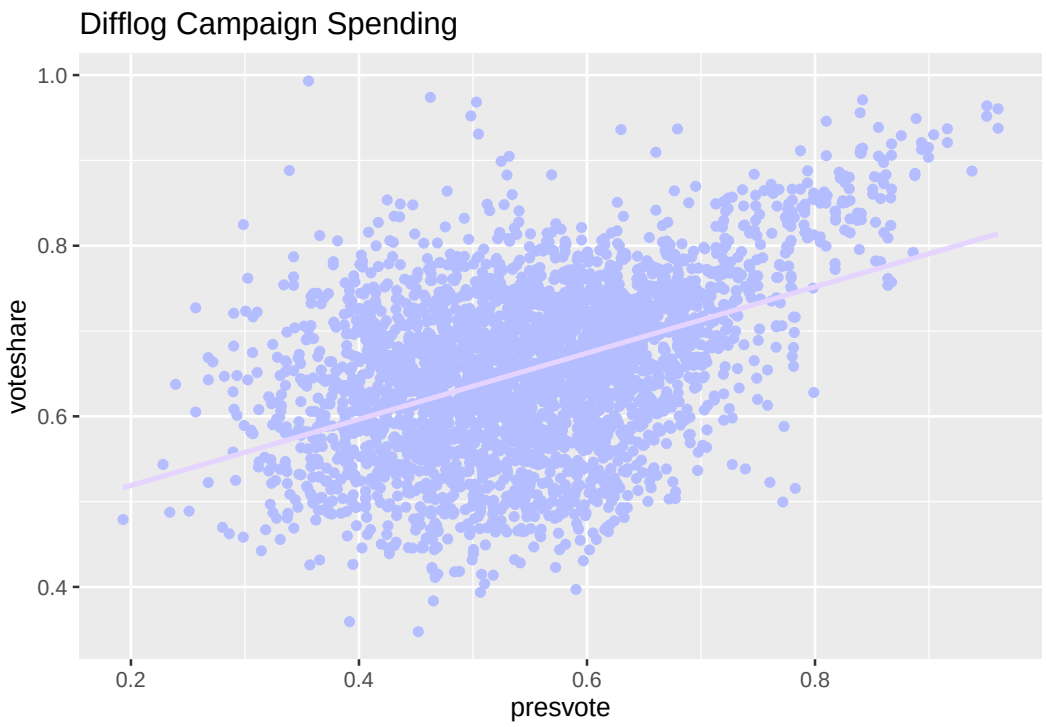
```
Call:
lm(formula = voteshare ~ presvote, data = inc.sub)

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.

```
#Create scatterplot with regression line and customized colors
ggplot(data = inc.sub, aes(x = presvote, y = voteshare)) +
  geom_point(color = "#B4BDFF") +
  #Add scatterplot points
  geom_smooth(method = "lm", se = FALSE, color = "#E5D4FF") +
  #Add regression line
  labs(title = "Difflog Campaign Spending",
       x = "presvote",
       y = "voteshare")
```

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




```
According to the results, the linear correlation is positive,
indicating that campaign spending increases,
```

```
the presvote tends to increase.
Moreover, the dotted plots are concentrated on both sides of the linear plot,
indicating that there is a strong linear relationship.
```

3. Write the prediction equation.

```r
# Extracting the intercept and slope
intercept <- round(coef(reg_model_3)[1],4)
slope <- round(coef(reg_model_3)[2],4)

# Printing the intercept and slope
cat("Final Model: voteshare =", intercept, "+", slope, "* presvote\n")
```

```
Final Model: voteshare = 0.4413 + 0.388 * presvote

For each unit increase in the prevote variable,
we expect the vote share to increase by approximately 0.388.
The intercept 0.4413 represents the estimated baseline value of vote share
without the influence of pre-voting variables.
```

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.

```r
#Create a regression about residual_1 and residual_3
reg_model_4<- lm(res_1~res_2,data = inc.sub)
summary(reg_model_4)
```

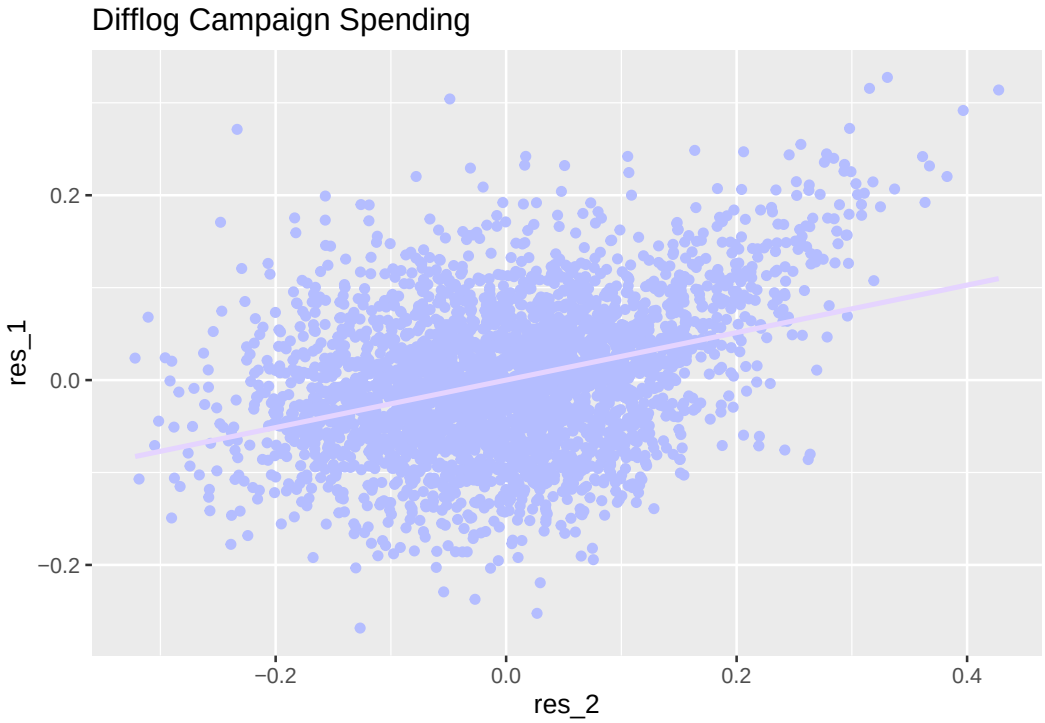
```
Call:
lm(formula = res_1 ~ res_2, data = inc.sub)

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
res_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

According to the results, the F statistic is 477,
the degrees of freedom are 1 and 3191,
and the p value is less than 2.2e-16,
indicating a high degree of confidence in rejecting the null hypothesis.
In summary, the model suggests that there is a
statistically significant relationship between res_1 and res_2.
The F-statistic and its associated p-value support
the overall significance of the model.
The coefficient for res_2 is also significant,
suggesting that it has a meaningful impact on the dependent variable
```

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

```r
#Create scatterplot with regression line and customized colors
ggplot(data = inc.sub, aes(x = res_2, y = res_1)) +
  geom_point(color = "#B4BDFF") +
  #Add scatterplot points
  geom_smooth(method = "lm", se = FALSE, color = "#E5D4FF") +
  #Add regression line
  labs(title = "Difflog Campaign Spending",
       x = "res_2",
       y = "res_1")
```

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



3. Write the prediction equation.

```
# Extracting the intercept and slope
intercept <- round(coef(reg_model_4)[1],4)
slope <- round(coef(reg_model_4)[2],4)

# Printing the intercept and slope
cat("Final Model: res_1=", intercept, "+", slope, "* res_2\n")
```

```
Final Model: res_1= 0 + 0.2569 * res_2
```

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`.

```
#Create a regeression about voteshare, 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

According to the results, the F statistic is 1303,
the degrees of freedom are 2 and 3190,
and the p-value is less than 2.2e-16,
indicating that there is greater credibility to reject the null hypothesis.
In summary, the model suggests a significant relationship
between voteshare, difflog, and presvote.
The R-squared values indicate that the model explains
a substantial proportion of the variance in voteshare.
```

2. Write the prediction equation.

```r
#write the prediction equation
# Extracting the intercept and slope
intercept <- round(coef(multreg_5)[1],4)
slope_1 <- round(coef(multreg_5)[2],4)
slope_2 <- round(coef(multreg_5)[3],4)

# Printing the intercept and slope
cat("Final Model: votehare=", intercept, "+", slope_1, "* difflog", "+",
    slope_2, "* presvote\n")
```

```
Final Model: votehare= 0.4486 + 0.0355 * difflog + 0.2569 * 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?

```
In both instances, the coefficient for the variable of interest
 (presvote in Question 4 and residuals2 in the recent regression)
 coincidentally appears to be approximately 0.2569.
 The associated t-values are both 21.84,
```

```
and both exhibit highly significant p-values (<2e-16).
This similarity might be a statistical coincidence or could suggest
that the variable residuals2 has a comparable impact
on the dependent variable (voteshare in Question 4 and residuals1 in the recent regression).
While investigating further, it is essential to consider
whether there is a theoretical or logical explanation
for this observed similarity.
Additionally, careful scrutiny of the model setup
and data appropriateness is crucial. If the relationship
between residuals2 and the dependent variable
is unexpected from a theoretical standpoint,
it may warrant a closer examination to ensure
the robustness of the analysis and rule out potential data issues.
```