# Problem Set 3

Applied Stats/Quant Methods 1

Due: November 12, 2021

## 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 in `.pdf` form.
- This problem set is due before class on Friday November 12, 2021. No late assignments will be accepted.
- Total available points for this homework is 80.

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`.**

   To Run the regression we first have to make some assumptions:

   1. Randomised Data Generation
   2. Independent observations
   3. Linearity (a straight line relationship between x and y)
   4. Normality and constant variance.

Once the above is assumed. We then make our hypotheses. In this case as we are looking at if the difference in campaign spending between incumbent and challenger affects the incumbents vote share. The null hypothesis would be that there is no affect on the difference in spending and the alternative hypothesis would be that there is a difference and therefore the slope would not be zero.

H0 = Slope is 0

Ha = Slope is not 0.

We then run the regression using the below code where `voteshare` is the outcome variable (Y) and `difflog` is the explanatory Variable (X):

```
summary(lm(data = incumbents, voteshare ~ difflog))
```

This output the following:

```
Call:
lm(formula = voteshare ~ difflog, data = incumbents)

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 <0.0000000000000002 ***
difflog     0.041666   0.000968   43.04 <0.0000000000000002 ***
---
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: < 0.00000000000000022
```

What this is showing is that for every one unit increase in Y the voteshare (the outcome variable) there is a 0.04167 increase in the X the diffshare. We can see that the p-value is significant so we reject the null hypothesis that there the slope is zero.

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

The following code was input into R:

```
ggplot(data = incumbents, aes(x = difflog, y = voteshare)) +
  geom_point(alpha = 0.2) + #add a scatterplot
  geom_smooth(method = lm) #add a linear regression line
```

The above produced the following graph:

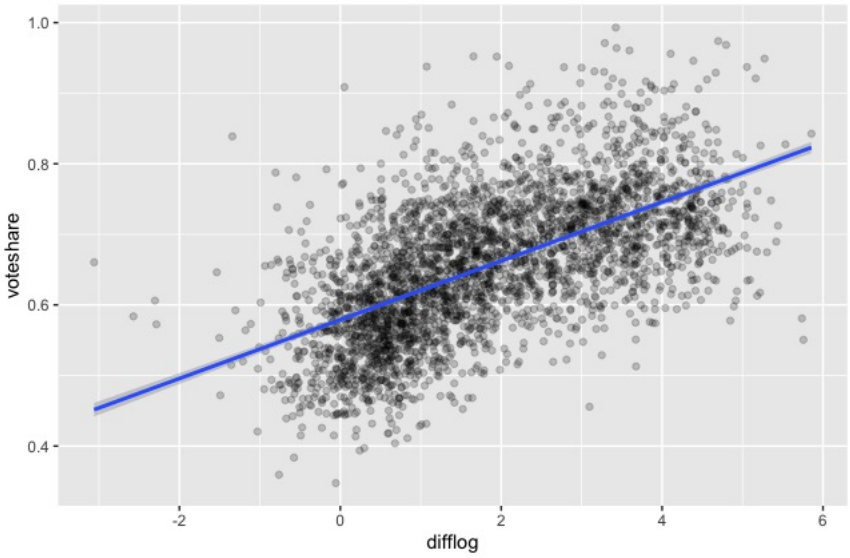


Figure 1: Scatterplot: voteshare and difflog with the regression line

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

   The following code was used to save the residuals as a separate object, I called these **q1_residuals**

```
q1_residuals <- residuals(lm(data = incumbents, voteshare ~ difflog))
```

4. **Write the prediction equation.**

   The general prediction equation is as follow: `y = a + bx`. For the regression model above the prediction model is found by using the regression outputs:

```
lm(data = incumbents, voteshare ~ difflog)

(Intercept)      difflog
    0.57903      0.04167
```

   The intercept and slope can be input to produce the following prediction equation:

   `y = 0.57903 + 0.04167x`

# 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`.**

   To run the regression we first have to make some assumptions:

   1. Randomised Data Generation
   2. Independent observations
   3. Linearity (a straight line relationship between x and y)
   4. Normality and constant variance.

   Once the above is assumed. We then make our hypothesis. Here we are investigating the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related. The null hypothesis would be there there is not relation and therefore the slope would be zero and the alternative hypothesis would be that there is a relation and the slope would not be zero.

   H0 = Slope = 0

   Ha = Slope != 0

   The following code was used to run a regression where presvote is the outcome variable (y) and difflog is the explanatory variable (x)

```
summary(lm(data = incumbents, presvote ~ difflog))
```

   This produced the following output:

```
Call:
lm(formula = presvote ~ difflog, data = incumbents)

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 <0.0000000000000002 ***
difflog     0.023837   0.001359   17.54 <0.0000000000000002 ***
---
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: < 0.00000000000000022
Coefficients:
(Intercept)      difflog
    0.50758      0.02384
```

The above output shows that the slope of the regression is 0.02384 and the p value shows that this is significantly different to 0. So we can reject the null hypothesis to suggest that there is a relationship between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party.

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

The following code was used to produce a scatterplot:

```
ggplot(data = incumbents, aes(x = difflog, y = presvote)) +
  geom_point(alpha = 0.2) + #add a scatterplot
  geom_smooth(method = lm) #add a linear regression line
```

The following graph was produced:

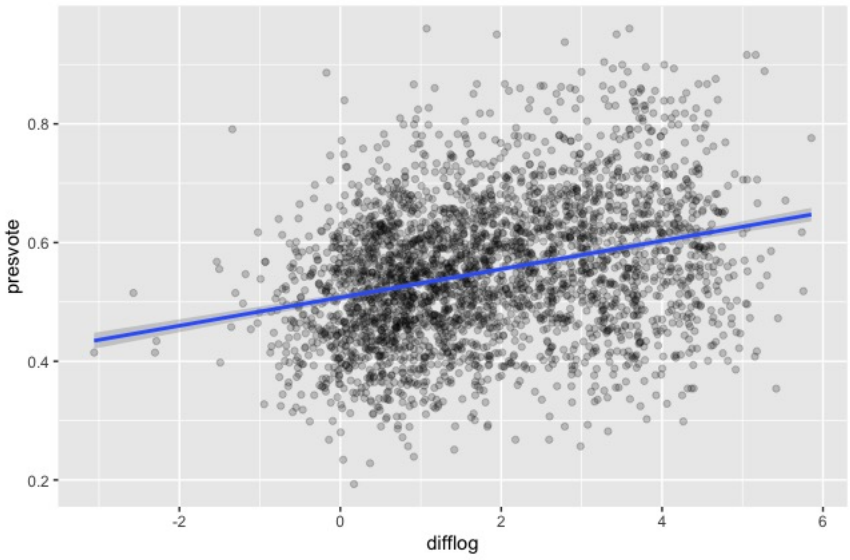


Figure 2: Scatterplot: presvote and difflog with the regression line

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

To save the residuals as a separate object the following code was used:

```
q2_residuals <- residuals(lm(data = incumbents, presvote ~ difflog))
```

4. **Write the prediction equation.**

   The prediction equation for this model was produced using the following outputs from the lm function:

```
(Intercept)      difflog
    0.50758      0.02384
```

   Therefor the prediction equation for this model is:

   `y = 0.50758 + 0.02384x`

# 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`.**

   For the following regression we first need to acknowledge the assumptions.

   1. Randomised Data Generation
   2. Independent observations
   3. Linearity (a straight line relationship between x and y)
   4. Normality and constant variance.

   Now this is assumed. We then make our hypothesis. Here we are investigating how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success. The null hypothesis would be there there is no association and therefore the slope would be zero and the alternative hypothesis would be that there is an association and the slope would not be zero.

   H0 = Slope = 0

   Ha = Slope != 0

   To run a regression equation where the outcome variable (Y) is voteshare and the explanatory variable (x) is presvote I used the following code:

```
summary(lm(data = incumbents, voteshare ~ presvote))
```

   This produced the following output:

```
Call:
lm(formula = voteshare ~ presvote, data = incumbents)

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 <0.0000000000000002 ***
presvote    0.388018   0.013493   28.76 <0.0000000000000002 ***
---
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: < 0.00000000000000022
```

The above summary of the regression can us investigate the association. It shows the slope is 0.388018 and the p-value shows that this is significantly different to zero. This suggests that there is an association between the vote share of the presidential candidate of the incumbent's party and the incumbent's electoral success.

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

The following code was used to produce a scatterplot:

```
ggplot(data = incumbents, aes(x = presvote, y = voteshare)) +
  geom_point(alpha = 0.2) + #add a scatterplot
  geom_smooth(method = lm) #add a linear regression line
```

This produced the following graph:

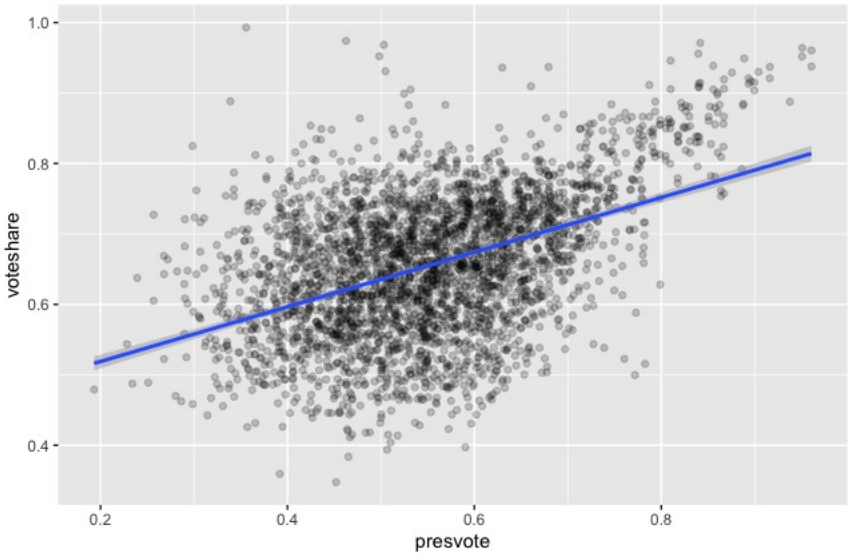


Figure 3: Scatterplot: voteshare and presvote with the regression line

3. **Write the prediction equation.**

The following output from the regression formula was used to produce the prediction line:

```
(Intercept)     presvote
     0.4413       0.3880
```

The prediction equation is `y = 0.4413 + 0.3880x`

# 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.**

   To run this regression we need to make the following assumptions about the residuals from Question 1 & 2.

   1. Randomised Data Generation
   2. Independent observations
   3. Linearity (a straight line relationship between x and y)
   4. Normality and constant variance.

   Question 1 residuals are showing variation not explained by difference in spending (incumbent and challenger)

   Question 2 residuals are showing variation not explained by difference in spening in the district (incumbent and challenger)

   This regression is looking to see if there is a relationship between the variation that was not explained.

   The null hypothesis would be there there is no relationship and therefore the slope would be zero and the alternative hypothesis would be that there is a relationship and the slope would not be zero.

   H0 = Slope = 0

   Ha = Slope != 0

   The following code was used to run a regression using the residuals from Question 1 and 2.

   Where the residuals from Question 1 were the outcome variable (y) and residuals from Question 2 were the explanatory variables (x) .

```
summary(lm(data = incumbents, q1_residuals ~ q2_residuals))
```

   This produced the following output:

```
Call:
lm(formula = q1_residuals ~ q2_residuals, data = incumbents)

Residuals:
      Min        1Q    Median        3Q       Max
-0.25928  -0.04737  -0.00121   0.04618   0.33126
```

```
Coefficients:
                              Estimate                 Std. Error t value
(Intercept)   -0.00000000000000005207   0.00129860494954970010916    0.00
q2_residuals   0.256877012700097828724   0.011761902396021798115   21.84
                         Pr(>|t|)
(Intercept)                     1
q2_residuals <0.0000000000000002 ***
---
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: < 0.00000000000000022
```

The regression output shows that there is a relationship between the residuals. So we reject the null hypothesis that there is no relationship.

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

The following code was used to produce a scatterplot of the residuals and the regression line:

```
ggplot(data = incumbents, aes(x = q2_residuals, y = q1_residuals)) +
  geom_point(alpha = 0.2) + #add a scatterplot
  geom_smooth(method = lm) #add a linear regression line
```

This produced the following Graph:

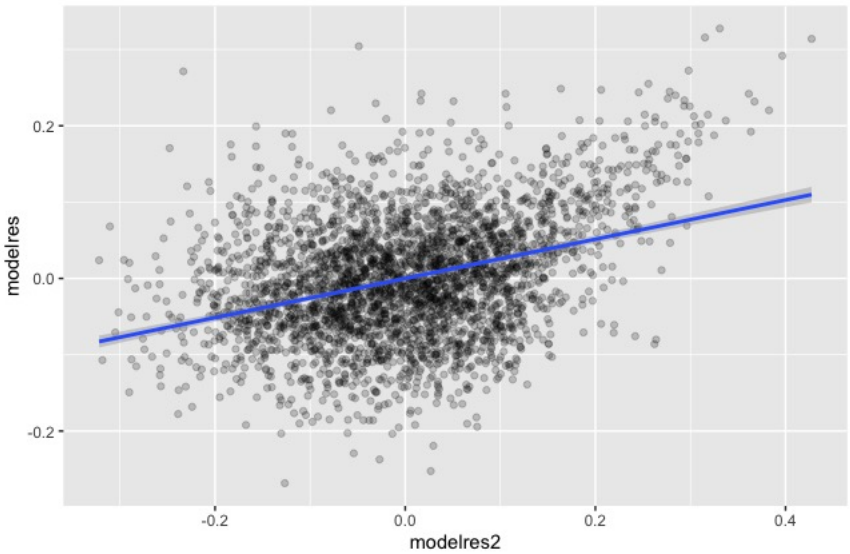


Figure 4: Scatterplot: Residuals from Q1 and Q2 with the regression line

3. **Write the prediction equation.** The following code produces the outputs needed to write the prediction equation:

```
lm(data = incumbents, q1_residuals ~ q2_residuals)
```

The following output is produced which I then input into the prediction equation:

```
Coefficients:
             (Intercept)              q2_residuals
-0.00000000000000005207    0.256877012700097828724
```

This means the prediction equation for the regression model on the residuals is:

`y = -0.00000000000000005207 + 0.256877012700097828724x`

# 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`.**

   To run the following Multivariate regression the following was assumed:

   1. Randomised Data Generation
   2. Independent observations
   3. Linearity (a straight line relationship between x and y)
   4. Normality and constant variance.

   We can then make the hypotheses for this regression. The null hypothesis is that there is no relationship between the Y (voteshare) and X (difflog & presvote) variables, which would be signified by having a slope of 0, suggesting that the relationship is nothing more than chance. The alternative hypothesis suggests there is a relationship between the X and the Y variables.

   *H0 = Slope = 0*

   *Ha = Slope != 0*

   The following code was used to run a regression where the voteshare is the outcome variable and the explanatory variables are the difflog and presvote:

```
summary(lm(data = incumbents, voteshare ~ difflog + presvote))
```

   This produces the following output:

```
Call:
lm(formula = voteshare ~ difflog + presvote, data = incumbents)

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 <0.0000000000000002 ***
difflog     0.0355431  0.0009455   37.59 <0.0000000000000002 ***
presvote    0.2568770  0.0117637   21.84 <0.0000000000000002 ***
---
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: < 0.00000000000000022
```

2. **Write the prediction equation.**

   The prediction equation can be found using the following code:

```
lm(data = incumbents, voteshare ~ difflog + presvote)
```

   This output the following, which can be used to write the prediction equation:

```
Coefficients:
(Intercept)      difflog     presvote
    0.44864      0.03554      0.25688
```

   The prediction equation for this regression model is:

   `y = 0.44864 + 0.03554(difflog) + 0.25688(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?**

   The presvote slope is 0.25688 which is the same as the slope of the Q2 residuals for Question 4 which is 0.256877012700097828724.

   The Q2 residuals slope is showing how much of the variation in presvote is not explained in difflog and this is what is happening when the difflog is added to the regression here between voteshare and presvote. This is because when a second explanatory variable is added into the regression it is accounting for variation. The multivariate regression allows us to account for relationships between variables, to make sure that a relationship is still significant when that variable is accounted for.