Problem Set 3

Applied Stats/Quant Methods 1

Due: November 20, 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 20, 2022. 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`.

   The following code was used to run the regression:

```
reg1 <- lm(voteshare ~ difflog , data = inc.sub)
```

   From this, using Stargazer an output table was created (see Table 1) From this regression, there was a significant result between the 'difflog' and 'voteshare' variables, $F(1, 3191) = 1{,}852.791$, $p < .01$, with $R^2 = 0.367$. This means that 36.7% of the variance in the 'Voteshare' variable can be explained by the variance in the 'difflog' variable.

Table 1: Difflog and Voteshare

	Dependent variable:
	voteshare
difflog	0.042***
	(0.001)
Constant	0.579***
	(0.002)
Observations	3,193
R^2	0.367
Adjusted R^2	0.367
Residual Std. Error	0.079 (df = 3191)
F Statistic	1,852.791*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

   A scatterplot was made using the following code:

```
ggplot(aes(x = difflog, # Scatterplot
           y = voteshare),
       data = inc.sub) +
  geom_smooth(method = "lm") +
  geom_point()
```

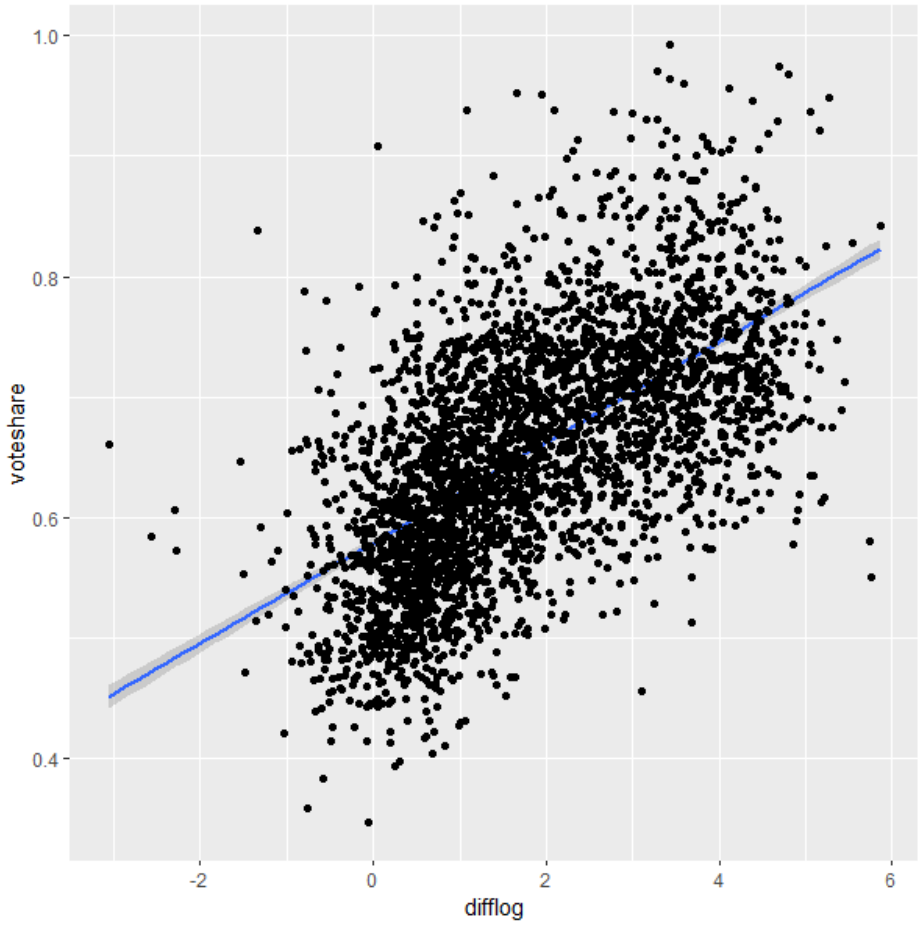
3. Save the residuals of the model in a separate object.

   Residuals were saved using the following line:

```
resid1 <- reg1$residuals
```

4. Write the prediction equation.
   According to the output table, for every increase in one unit in the 'difflog' variable, there is a 0.042 unit increase in 'voteshare'. This is shown via the equation:
   $Y_i = 0.579 + 0.042X_i$

Figure 1: Scatterplot of 'voteshare' and '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`.

   The following code was used to run the regression:

```
reg2 <- lm(presvote ~ difflog, data = inc.sub)
```

   From this, using Stargazer an output table was created (see Table 2) From this regression, there was a significant result between the 'difflog' and 'presvote' variables, $F(1,$

Table 2: Difflog and Presvote

	Dependent variable:
	presvote
difflog	0.024***
	(0.001)
Constant	0.508***
	(0.003)
Observations	3,193
R^2	0.088
Adjusted R^2	0.088
Residual Std. Error	0.110 (df = 3191)
F Statistic	307.715*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

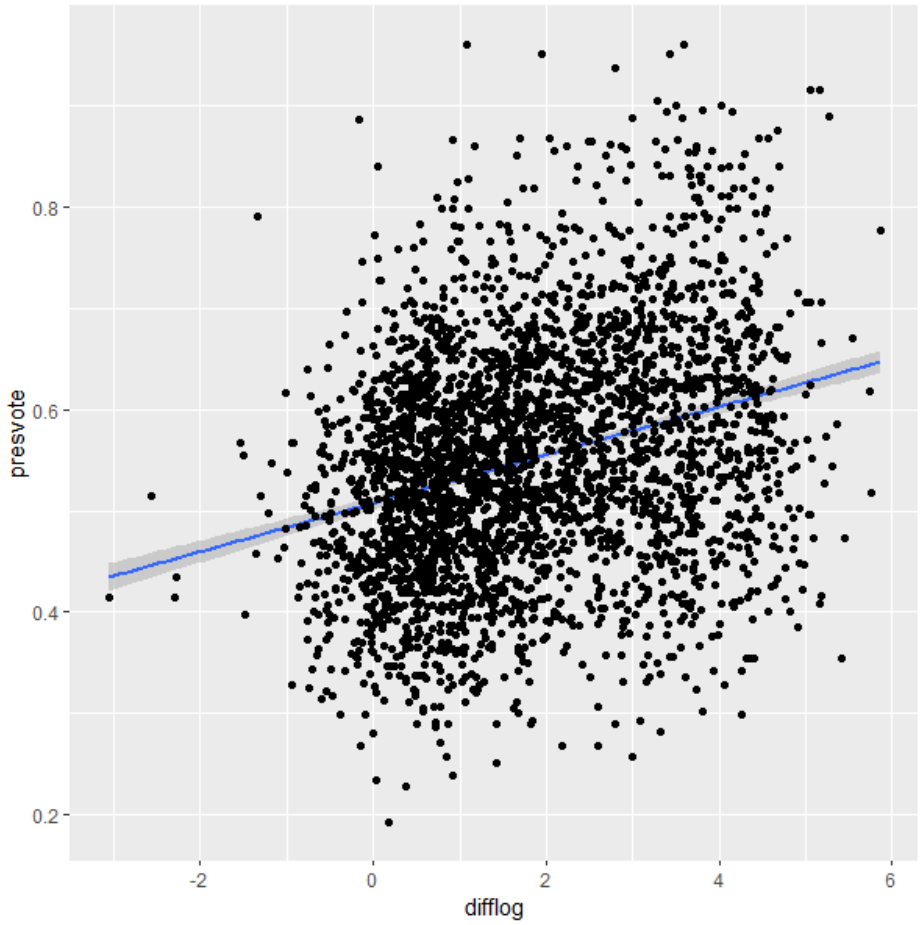
3191) = 307.715, $p < .01$, with $R^2 = 0088$. This means that 8.8% of the variance in the 'Presvote' variable can be explained by the variance in the 'difflog' variable.

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

A scatterplot was made using the following code:

```
ggplot(aes(x = difflog, # Scatterplot
           y = presvote),
       data = inc.sub) +
  geom_smooth(method = "lm") +
  geom_point()
```

Figure 2: Scatterplot of 'presvote' and 'difflog'



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

   Residuals were saved using the following line:

```
resid2 <- reg2$residuals
```

4. Write the prediction equation.
   According to the output table, for every increase in one unit in the 'difflog' variable, there is a 0.024 unit increase in 'voteshare'. This is shown via the equation:
   $Y_i = 0.508 + 0.024X_i$

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`.

   The following code was used to run the regression:

```
reg3 <- lm(voteshare ~ presvote, data = inc.sub)
```

   From this, using Stargazer an output table was created (see Table 3) From this regression, there was a significant result between the 'voteshare' and 'presvote' variables, $F(1, 3191) = 826.950$, $p < .01$, with $R^2 = 0.206$ This means that 20.6% of the variance in the 'Voteshare' variable can be explained by the variance in the 'Presvote' variable.

Table 3: Presvote and Voteshare

	Dependent variable:
	voteshare
presvote	0.388***
	(0.013)
Constant	0.441***
	(0.008)
Observations	3,193
R^2	0.206
Adjusted R^2	0.206
Residual Std. Error	0.088 (df = 3191)
F Statistic	826.950*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

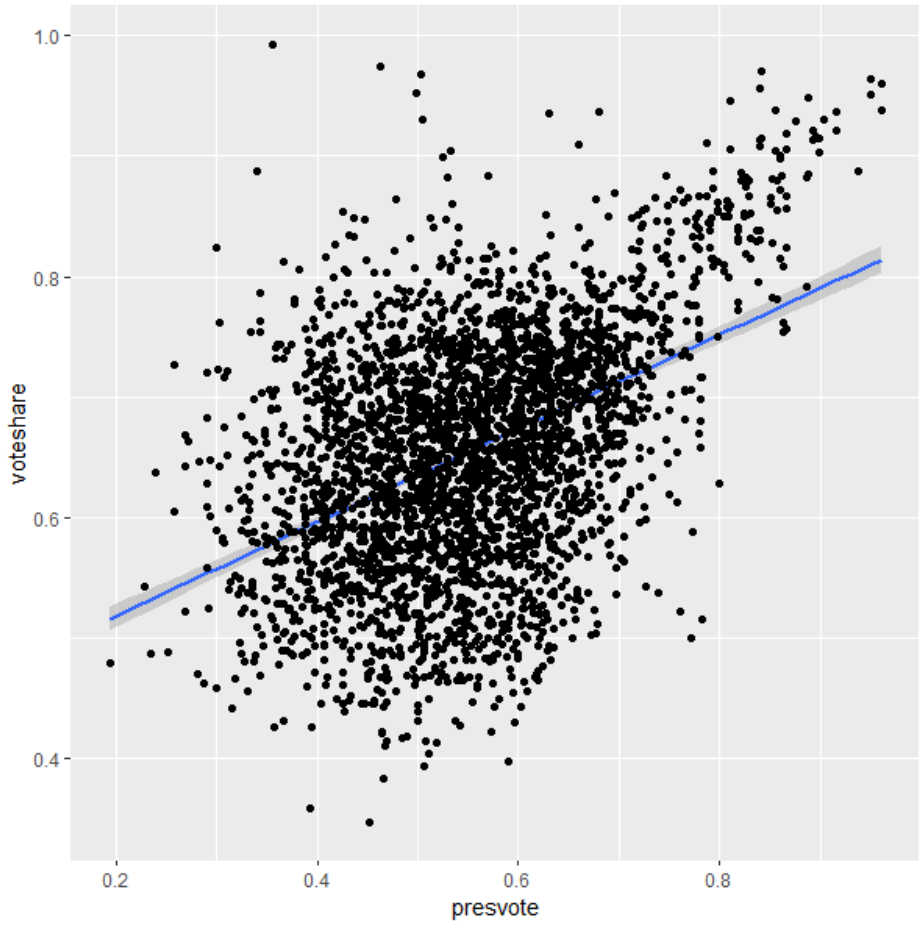
2. Make a scatterplot of the two variables and add the regression line.

   A scatterplot was made using the following code:

```
ggplot(aes(x = presvote, # Scatterplot
```

```
            y = voteshare),
        data = inc.sub) +
  geom_smooth(method = "lm") +
  geom_point()
```

Figure 3: Scatterplot of 'voteshare' and 'presvote'



3. Write the prediction equation.
   According to the output table, for every increase in one unit in the 'presvote' variable, there is a 0.388 unit increase in 'voteshare'. This is shown via the equation:
   $Y_i = 0.441 + 0.388X_i$

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.

   The following code was used to run the regression:

```
reg4 <- lm(resid1 ~ resid2)
```

   From this, using Stargazer an output table was created (see Table 4) From this regression, there was a significant result between the 'difflog' and 'presvote' variables, $F(1, 3191) = 476.975$, $p < .01$, with $R^2 = 0.13$. This means that 13% of the variance in the 'resid1' variable can be explained by the variance in the 'resid2' variable.

Table 4: Residuals of Reg1 and Reg2

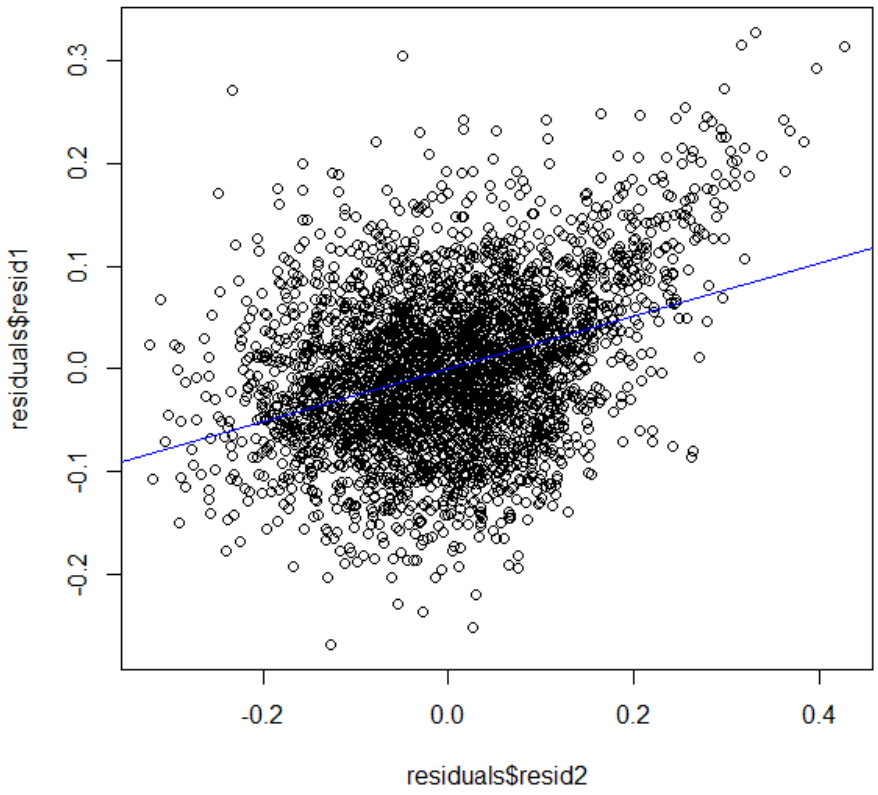
	Dependent variable:
	resid1
resid2	0.257***
	(0.012)
Constant	−0.000
	(0.001)
Observations	3,193
R^2	0.130
Adjusted R^2	0.130
Residual Std. Error	0.073 (df = 3191)
F Statistic	476.975*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

   A scatterplot was made using the following code:

```
plot(residuals$resid2, residuals$resid1)
abline(lm(resid1 ~ resid2), col = "blue")
```

Figure 4: Scatterplot of Residuals from Q1 and Q2



3. Write the prediction equation.
   According to the output table, for every increase in one unit in the 'resid1' variable, there is a 0.257 unit increase in 'resid2'. This is shown via the equation:
   $Y_i = 0.257X_i$

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`.

   The following code was used to run the regression:

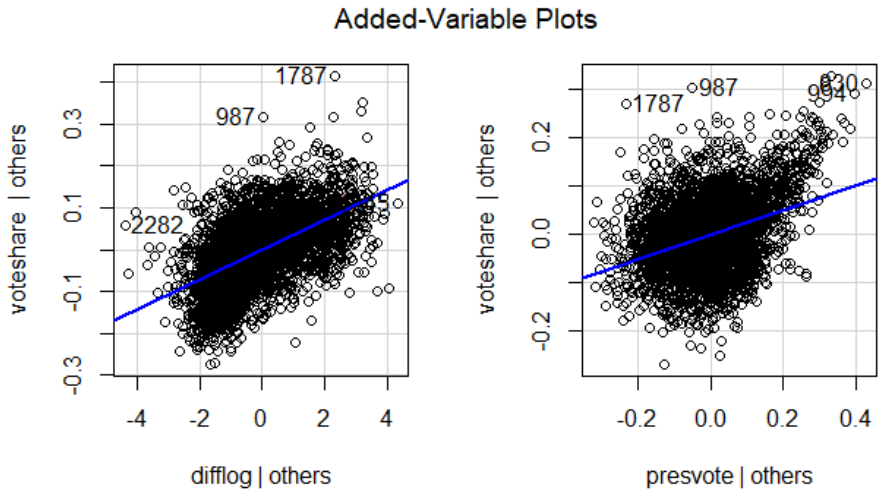
```
reg5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
```

   From this, using Stargazer an output table was created (see Table 5) From this regression, there was a significant result between the 'difflog' and 'presvote' variables, $F(1, 3190) = 1302.947$, $p < .01$, with the adjusted $R^2 = 0.449$. This means that 44.9% of the variance in the 'Voteshare' variable can be explained by the variance in the 'difflog' and 'Presvote' variables.

Table 5: Voteshare and Difflog + Presvote

	Dependent variable:
	voteshare
difflog	0.036***
	(0.001)
presvote	0.257***
	(0.012)
Constant	0.449***
	(0.006)
Observations	3,193
R^2	0.450
Adjusted R^2	0.449
Residual Std. Error	0.073 (df = 3190)
F Statistic	1,302.947*** (df = 2; 3190)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 5: Added-variable plots'



2. Write the prediction equation.

   According to the output table, for every increase in one unit in the 'voteshare' variable, there is a 0.036 unit increase in 'difflog', and a 0.257 unit increase in the 'presvote' variable. This is shown via the equation:
   $Y_i = 0.449 + 0.036X_{i1} + 0.257X_{i2}$

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' coefficient has an identical output to the 'resid2' coefficient in the regression carried out in question 4. This is because the presvote coefficient is derived from the variance in 'voteshare' that isn't explained by 'difflog'. Essentially, the coefficient is identical because it is controlling for the variance explained by the 'difflog' variable.