

Problem Set 3

Applied Stats/Quant Methods 1

Due: November 20, 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.
- 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`.

```
1 lm_q1 <- lm(voteshare ~ difflog, data = dat)
2 stargazer(lm_q1, title = "The association between difflog and voteshare")
```

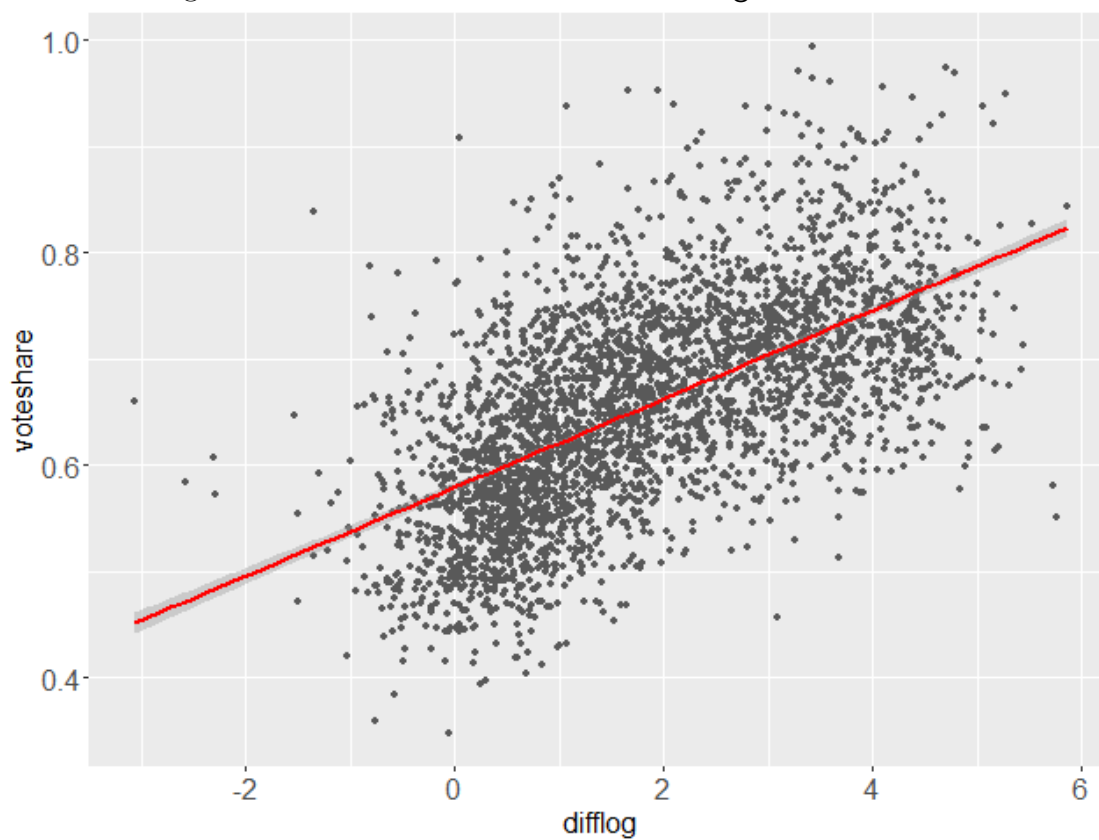
Table 1: The association between `difflog` and `voteshare`

<i>Dependent variable:</i>	
	<code>voteshare</code>
<code>difflog</code>	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)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

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

```
1 png("lm_q1.png", 640, 480)
2 dat %>%
3   ggplot(aes(difflog, voteshare)) +
4   geom_point(color = "gray35") +
5   geom_smooth(method = "lm", color = "red") +
6   theme(axis.text = element_text(size = 16), axis.title = element_text(
7     size = 16))
7 dev.off()
```

Figure 1: The association between `difflog` and `voteshare`



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

```
1 lm_q1_res <- lm_q1$residuals
```

4. Write the prediction equation.

$$y = \beta_0 + \beta_1 x$$

$$y = 0.579 + 0.042x$$

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`.

```
1 lm_q2 <- lm(presvote ~ difflog, data=dat)
2 stargazer(lm_q2, title = "The association between difflog and presvote")
```

Table 2: The association between `difflog` and `presvote`

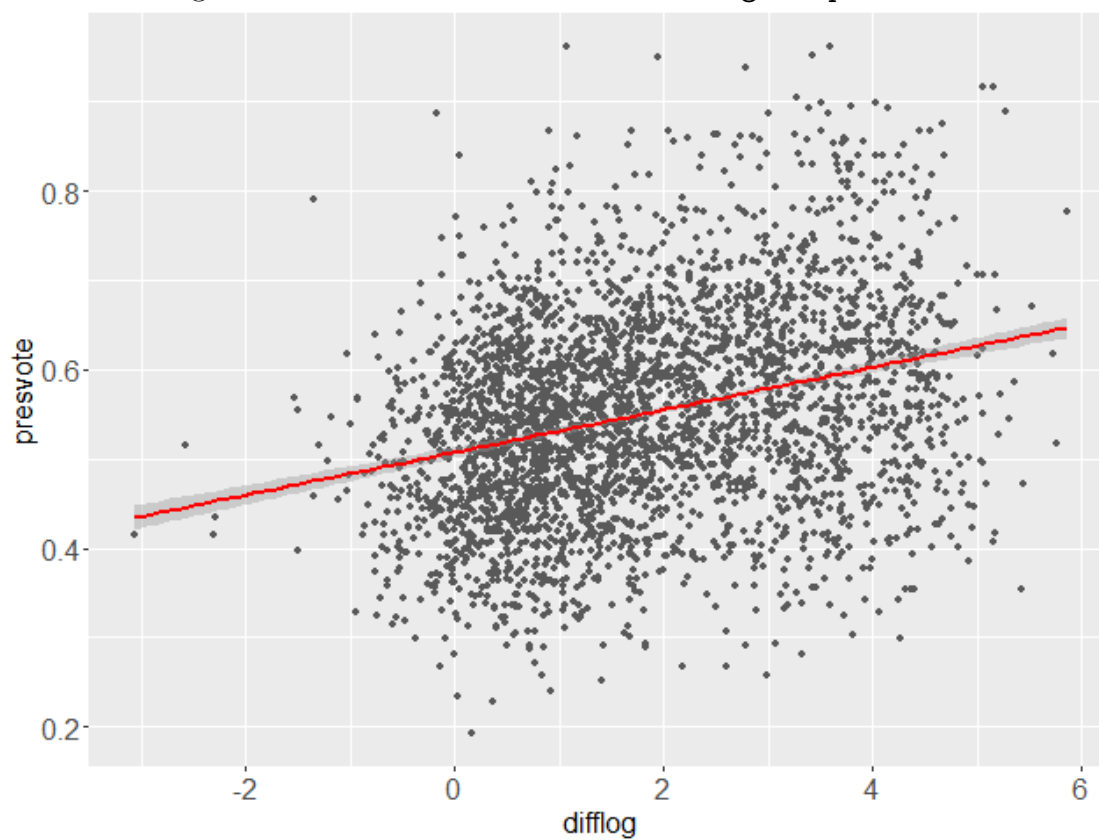
<i>Dependent variable:</i>	
	<code>presvote</code>
<code>difflog</code>	0.024*** (0.001)
Constant	0.508*** (0.003)
Observations	3,193
R ²	0.088
Adjusted R ²	0.088
Residual Std. Error	0.110 (df = 3191)
F Statistic	307.715*** (df = 1; 3191)

Note: *p<0.1; **p<0.05; ***p<0.01

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

```
1 png("lm_q2.png", 640, 480)
2 dat %>%
3   ggplot(aes(difflog, presvote)) +
4   geom_point(color = "gray35") +
5   geom_smooth(method = "lm", color = "red") +
6   theme(axis.text = element_text(size = 16), axis.title = element_text(
7     size = 16))
7 dev.off()
```

Figure 2: The association between difflog and presvote



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

```
1 lm_q2_res <- lm_q2$residuals
```

4. Write the prediction equation.

$$y = \beta_0 + \beta_1 x$$

$$y = 0.508 + 0.024x$$

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`.

```
1 lm_q3 <- lm(voteshare ~ presvote, data=dat)
2 stargazer(lm_q3, title = "The association between presvote and voteshare"
  )
```

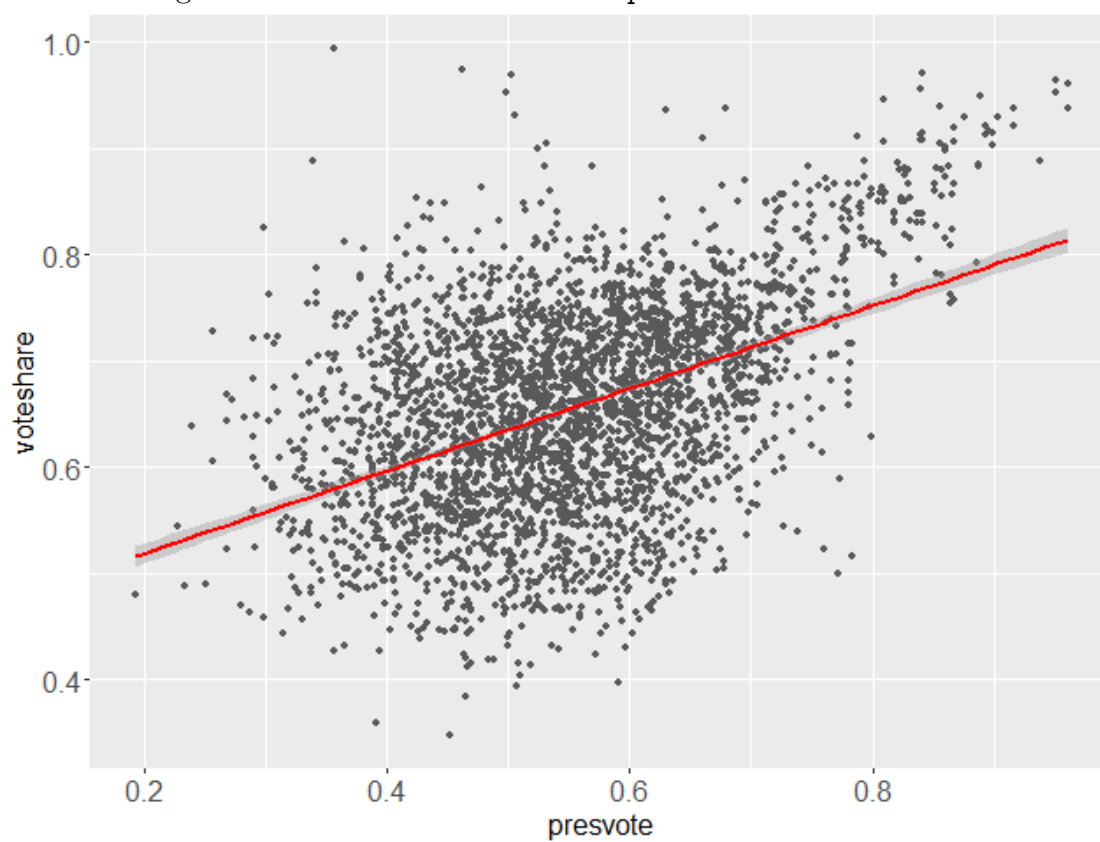
Table 3: The association between `presvote` and `voteshare`

	<i>Dependent variable:</i>
	<code>voteshare</code>
<code>presvote</code>	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)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

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

```
1 png("lm_q3.png", 640, 480)
2 dat %>%
3   ggplot(aes(presvote, voteshare)) +
4   geom_point(color = "gray35") +
5   geom_smooth(method = "lm", color = "red") +
6   theme(axis.text = element_text(size = 16), axis.title = element_text(
7     size = 16))
7 dev.off()
```

Figure 3: The association between `presvote` and `voteshare`



3. Write the prediction equation.

$$y = \beta_0 + \beta_1 x$$
$$y = 0.441 + 0.388x$$

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.

```
1 lm_q4 <- lm(lm_q1_res ~ lm_q2_res)
2 stargazer(lm_q4, title = "The association between the residuals from
3   Question 1 and Question 2",
4   column.labels = "Residuals from Question 1",
5   covariate.labels = "Residuals from Question 2")
```

Table 4: The association between the residuals from Question 1 and Question 2

	<i>Dependent variable:</i>
	lm_q1_res Residuals from Question 1
Residuals from Question 2	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)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

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

```
1 png("lm_q4.png", 640, 480)
2 dat %>%
3   ggplot(aes(lm_q2_res, lm_q1_res)) +
4   geom_point(color = "gray35") +
5   geom_smooth(method = "lm", color = "red") +
6   xlab("Residuals from Question 2") +
7   ylab("Residuals from Question 1") +
```

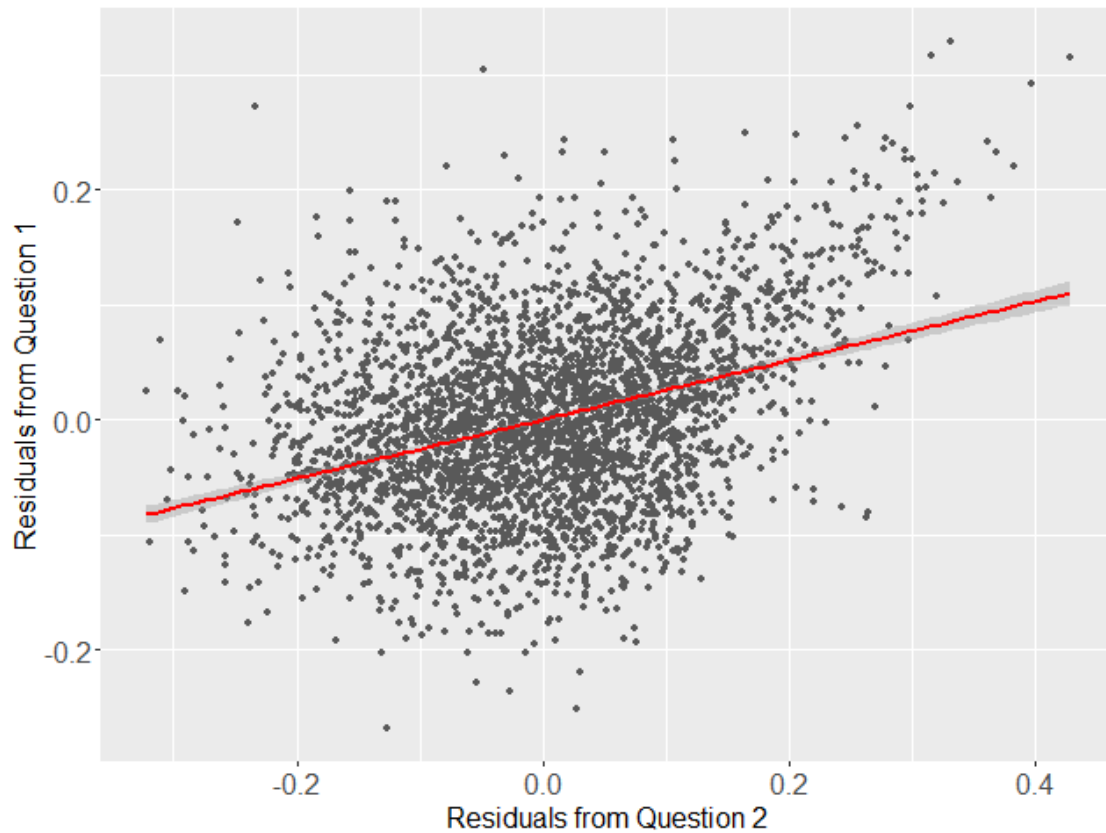


```

8 theme(axis.text = element_text(size = 16), axis.title = element_text(
  size = 16))
9 dev.off()

```

Figure 4: The association between the residuals from Question 1 and Question 2



3. Write the prediction equation.

$$y = \beta_0 + \beta_1 x$$

$$y = 0 + 0.257x$$

$$y = 0.257x$$

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.

```
1 lm_q5 <- lm(voteshare ~ difflog + presvote, data = dat)
2 stargazer(lm_q1, lm_q3, lm_q5, title="The association among difflog,
  presvote, and voteshare")
```

Table 5: The association among difflog, presvote, and voteshare

	<i>Dependent variable:</i>		
	voteshare		
	(1)	(2)	(3)
difflog	0.042*** (0.001)		0.036*** (0.001)
presvote		0.388*** (0.013)	0.257*** (0.012)
Constant	0.579*** (0.002)	0.441*** (0.008)	0.449*** (0.006)
Observations	3,193	3,193	3,193
R ²	0.367	0.206	0.450
Adjusted R ²	0.367	0.206	0.449
Residual Std. Error	0.079 (df = 3191)	0.088 (df = 3191)	0.073 (df = 3190)
F Statistic	1,852.791*** (df = 1; 3191)	826.950*** (df = 1; 3191)	1,302.947*** (df = 2; 3190)

Note:

*p<0.1; **p<0.05; ***p<0.01

2. Write the prediction equation.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$y = 0.449 + 0.036x_1 + 0.257x_2$$

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 slope (β_1) in the bivariate regression model of the residuals ($y = 0.257x$) is the same slope (β_2) of the variable `presvote` (x_2) in the multivariate regression model ($y = 0.449 + 0.036x_1 + 0.257x_2$). They are both 0.257.

To understand why that is the case, let's remember that the residuals regressed in question 4 represent the variations in `voteshare` and `presvote` which are not explained by `difflog`. Therefore, the bivariate regression model of the residuals captures how much of the unexplained variation in `voteshare` is associated to the unexplained variation in `presvote`, after having already considered the impact of `difflog`. In other words, the model of the residuals captures the effect of `presvote` on `voteshare` when we control for `difflog` — which is exactly what β_2 in the multivariate regression model represents.

For that reason, the slopes in both models are expected to be the same. After all, they are measuring the same thing: the variation in `voteshare` explained by `presvote` while controlling for `difflog`.