Problem Set 3

Tianxin Zhang/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.

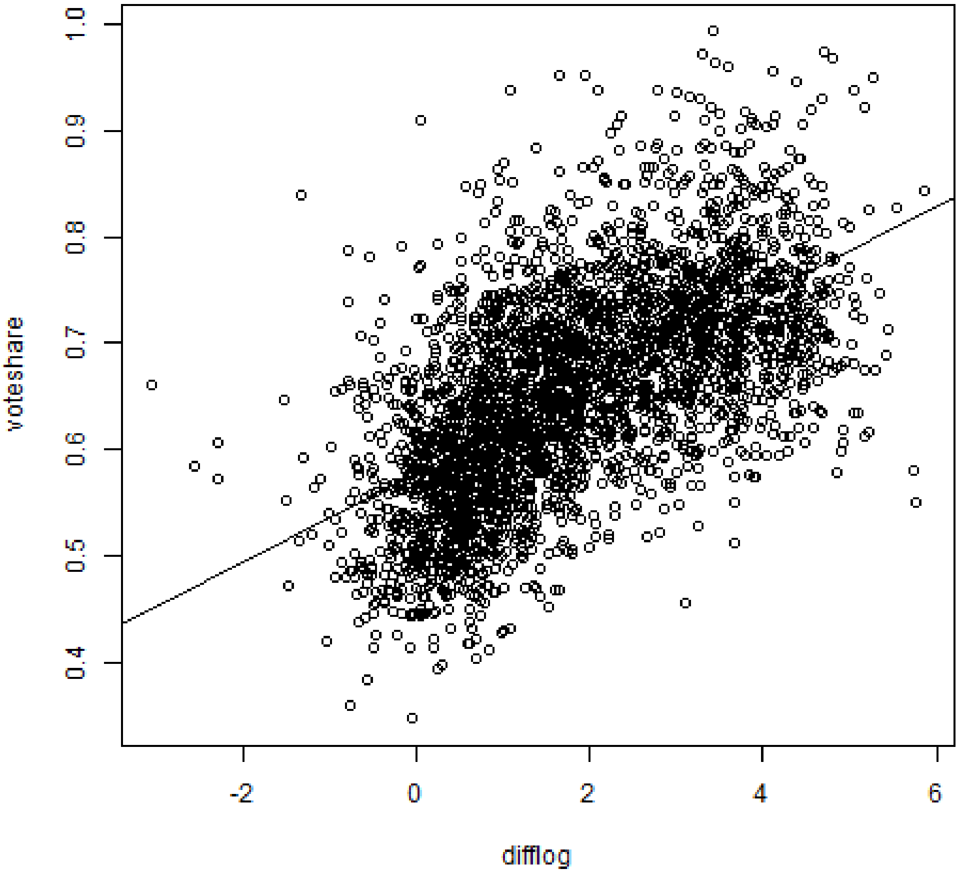
```r
DAT <- read.csv("C:/Users/Caesar/Documents/GitHub/StatsI_Fall2022/datasets/incumbents_subset.csv")

install.packages("stargazer")
library(stargazer)


## Q1:
LR_VS_DL <- lm(voteshare ~ difflog, data = DAT)
summary(LR_VS_DL)

## p-value for the coeeficient of difflog is 2.2e-16, smaller than 0.001, we can
# reject the null hypothesis that there is no association between voteshare and
# difflog statistically significant at the 99.9% level.

stargazer(LR_VS_DL, title="Regression Results: Vote Share ~ Difflog")
```

Table 1: Regression Results: Vote Share Difflog

	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.

```r
png("voteshare ~ difflog.png")
plot(voteshare ~ difflog, data = DAT)
abline(LR_VS_DL)
dev.off()
```



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

```
RS_LR_VS_DL <- residuals(LR_VS_DL)
print(RS_LR_VS_DL)
```

4. Write the prediction equation.
   $\hat{y} = 0.579 + 0.042x$

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`.

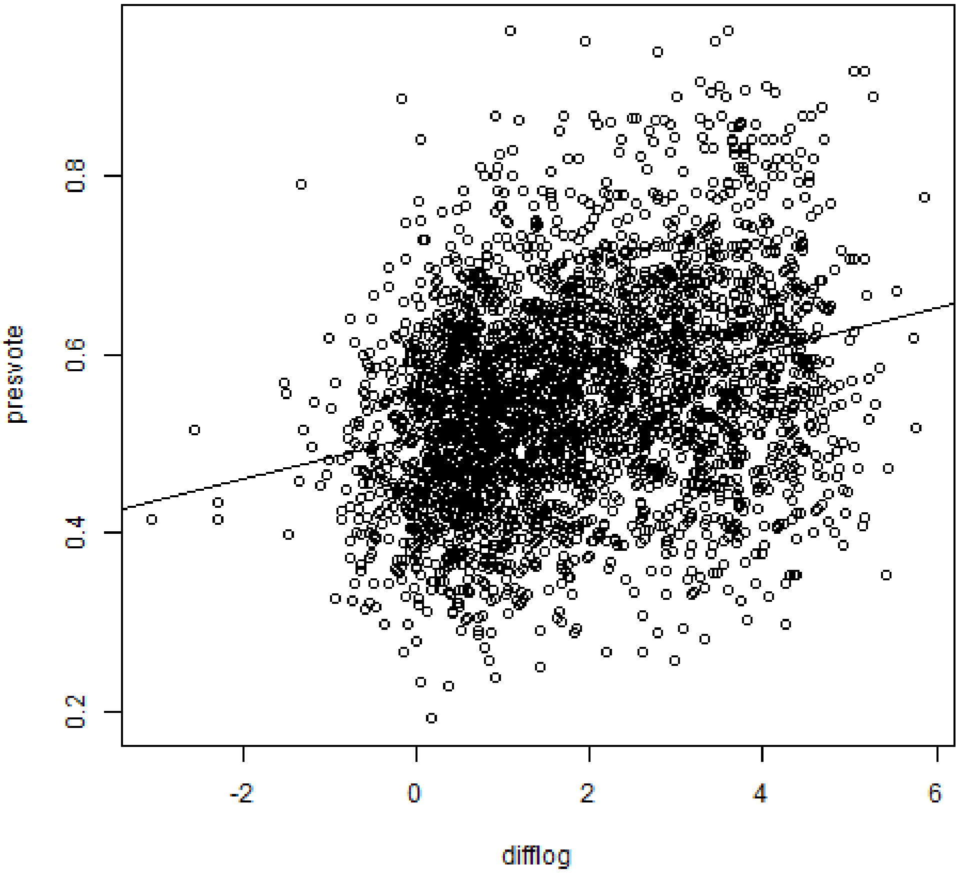
```r
## Q2:
LR_PV_DL <- lm(presvote ~ difflog, data = DAT)
stargazer(LR_PV_DL, title="Regression Results: Presvote ~ Difflog")

## p-value for the coefficient of difflog is 2.2e-16, smaller than 0.001, we can
# reject the null hypothesis that there is no association between presvote and
# difflog statistically significant at the 99.9% level.
```

Table 2: Regression Results: Presvote Difflog

	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

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

```r
png(”presvote ~ difflog.png”)
plot(presvote ~ difflog, data = DAT)
abline(LR_PV_DL)
dev.off()
```



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

```r
RS_LR_PV_DL <- residuals(LR_PV_DL)
print(RS_LR_PV_DL)
```

4. Write the prediction equation.
   $\hat{y} = 0.508 + 0.024x$

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`.

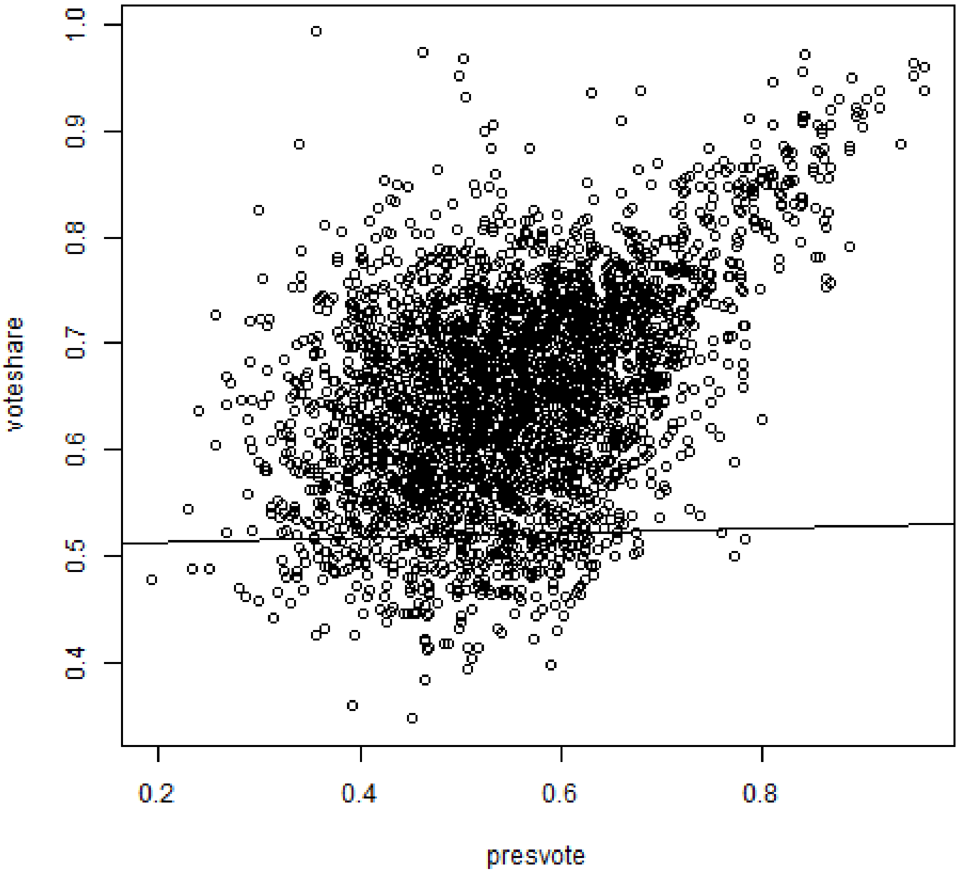
```r
## Q3:
LR_VS_PV <- lm(voteshare ~ presvote, data = DAT)
stargazer(LR_VS_PV, title="Regression Results: voteshare ~ Presvote")

## p-value of coefficient for presvote is 2e-16, smaller than 0.001, we can
# reject the null hypothesis that there is no association statistically
# significant between voteshare and presvote at the 99.9% level.
```

Table 3: Regression Results: voteshare Presvote

	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.

```
png(”voteshare ~ presvote.png”)
plot(voteshare ~ presvote, data = DAT)
abline(LR_PV_DL)
dev.off()
```



3. Write the prediction equation.
   $\hat{y} = 0.441 + 0.388x$

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.

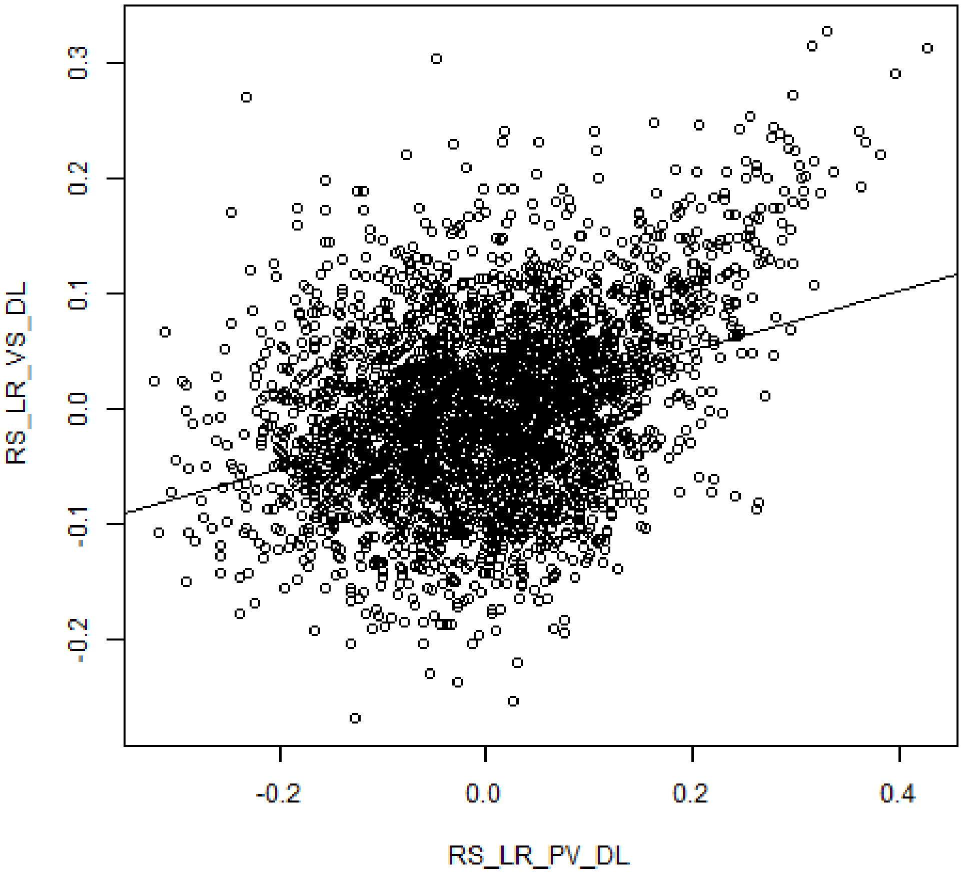
```r
## Q4:

LR_RS_PV_DL_VS_DL <- lm(RS_LR_VS_DL ~ RS_LR_PV_DL, data = DAT)
stargazer(LR_RS_PV_DL_VS_DL, title="Regression Results: voteshare ~ Presvote")

# p-value for the coefficient of residuals from regression model of Q2 is
# 2e-16 < 0.001, we can reject the null hypothesis that there is no association
# statistically significant between residuals from regression model of Q2 and
# regression model of Q1 at the 99.9% level.
```

Table 4: Regression Results: voteshare Presvote

	Dependent variable:
	RS_LR_VS_DL
RS_LR_PV_DL	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.

```
png("Residual (voteshare ~ difflog) ~ Redisual (presvote ~ difflog).png")
plot(RS_LR_VS_DL ~ RS_LR_PV_DL, data = DAT)
abline(LR_RS_PV_DL_VS_DL)
dev.off()
```



3. Write the prediction equation.
   $\hat{y} = 0.257x$

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`.

```r
## Q5:
LR_VS_DL_PV <- lm(voteshare ~ difflog + presvote, data = DAT)
stargazer(LR_VS_DL_PV, title = "Regression Results: voteshare ~ difflog + presvote")

# p-value of the coefficient for difflog is 2e-16, smaller than 0.001, we can
# reject the null hypothesis that there is no association statistically
# significant between vote share and difflog at the 99.9% level.

# p-value of the coefficient for presvote is 2e-16, smaller than 0.001, we can
# reject the null  hypothesis that there is no association statistically
# significant between vote share and presvote at the 99.9% level.
```

Table 5: Regression Results: voteshare 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

2. Write the prediction equation.
   $\hat{y} = 0.449 + 0.036x_1 + 0.257x_2$

3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case?
   Residuals of models from Q4 and Q5 are the same, which equals to 0.073. In the Regression Model of Q4, the residuals from the regression model (Voteshare difflog), is statistically associated with the residuals from the regression model (presvote difflog), which means RSS of Q4 refers to the unexplained variations by variables voteshare, difflog and presvote. The residuals of Q5 also refers to the unexplained variations by variables voteshare, difflog and presvote. So the residuals of Q4 and Q5 have the same value.