

Problem Set 3

Student: Shekhar Kedia (23351315)

Applied Stats/Quant Methods 1

Due: November 19, 2023

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.

R code to read the dataset:

```
1 # read in data
2 inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")
```

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 code to run the regression and create the output:

```
1 model_vote_diff <- lm(voteshare ~ difflog, data = inc.sub)
2 summary(model_vote_diff) #Exploring the output of the model
3
4 #Reporting Regression results
5 stargazer(model_vote_diff,
6           title = "Model 1 Regression Results")
```

Table 1: Model 1 Regression Results

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

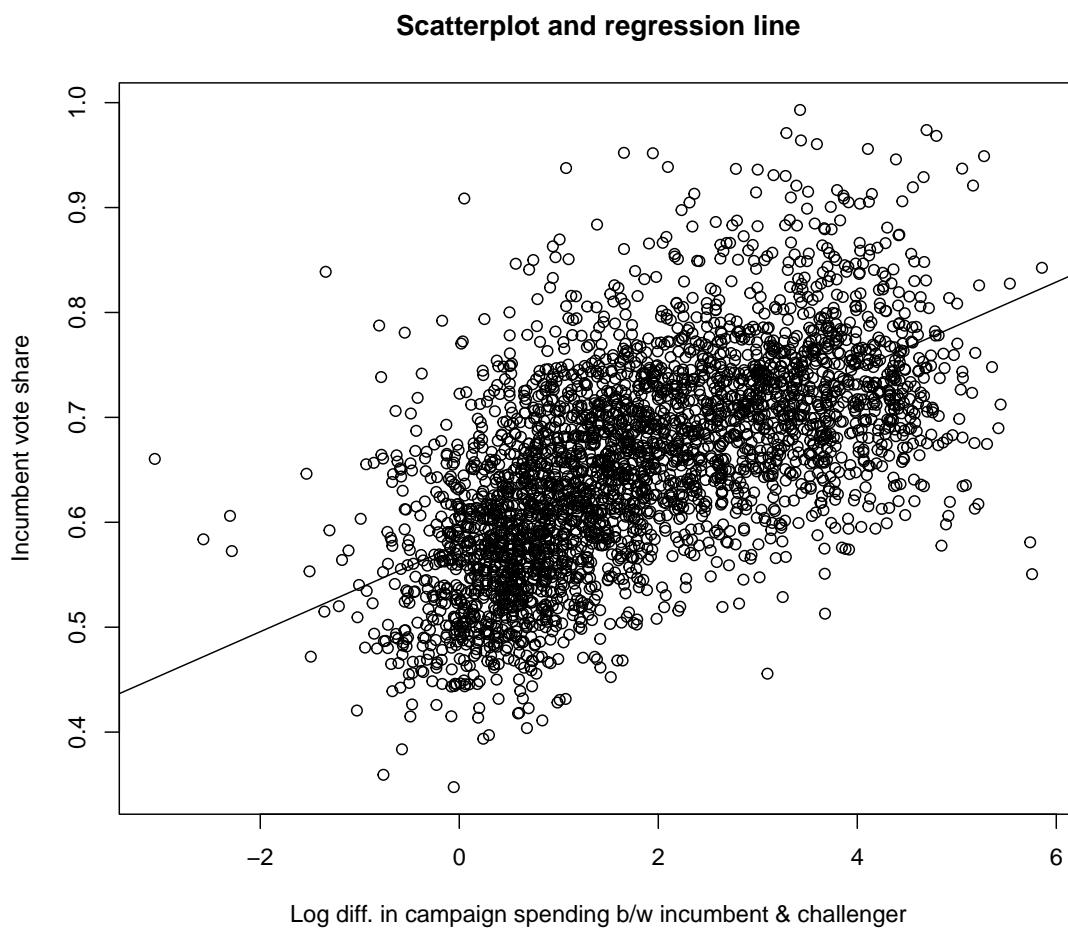
Interpretation:

The above model shows that one unit increase in log difference in campaign spending between incumbent and challenger is associated with, on average, a 0.042 unit increase in incumbent's vote share.

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

R code to create the scatter plot and add the regression line:

```
1 pdf("plot1.pdf", width=8)
2 plot(inc.sub$difflog ,
3      inc.sub$voteshare ,
4      xlab = "Log diff. in campaign spending b/w incumbent & challenger" ,
5      ylab = "Incumbent vote share" ,
6      main = "Scatterplot and regression line") #Scatter plot
7 abline(model_vote_diff) #Adding regression line
8 dev.off()
```



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

R code to save the residuals of the model in a separate object:

```
1 residuals_vote_diff <- model_vote_diff$residuals
```

4. Write the prediction equation.

The general form for bivariate equation is: $\hat{Y} = b_0 + b_1X_1$

Where,

\hat{Y} is the predicted value of the outcome variable,

b_0 is the y-intercept (the value of \hat{Y} when X_1 is 0),

b_1 is the slope of the regression line, representing the change in \hat{Y} for a one-unit change in X_1 ,

X_1 is the explanatory variable

In the context of the model, the predicted equation is:

$$voteshare = 0.579 + 0.042 * difflog$$

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**.

R code to run the regression and create the output:

```
1 model_pres_diff <- lm(presvote ~ difflog, data = inc.sub)
2 summary(model_pres_diff) #Exploring the output of the model
3
4 #Reporting Regression results
5 stargazer(model_pres_diff,
6           title = "Model 2 Regression Results")
```

Table 2: Model 2 Regression Results

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

Interpretation:

The above model shows that one unit increase in log difference in campaign spending between incumbent and challenger is associated with, on average, a 0.024 unit increase in vote share of the presidential candidate of the incumbent's party.

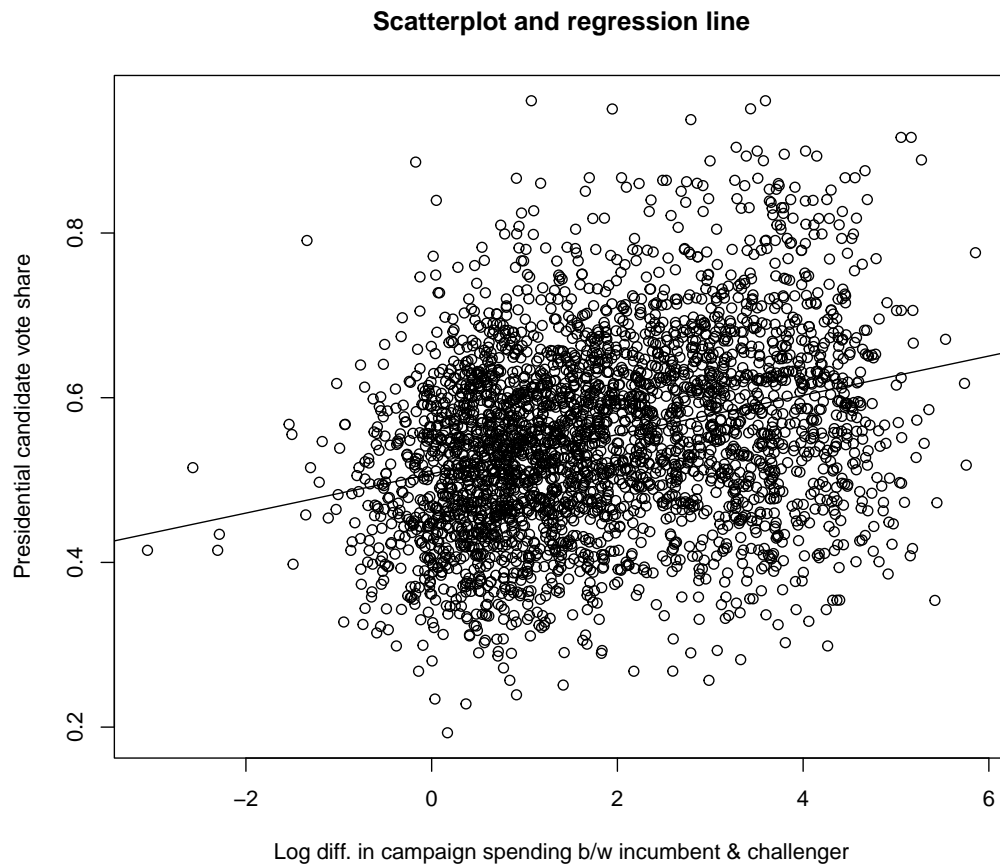
2. Make a scatterplot of the two variables and add the regression line.

R code to create the scatter plot and add the regression line:

```

1 pdf("plot2.pdf", width=8)
2 plot(inc.sub$difflog,
3       inc.sub$presvote,
4       xlab = "Log diff. in campaign spending b/w incumbent & challenger",
5       ylab = "Presidential candidate vote share",
6       main = "Scatterplot and regression line") #Scatter plot
7 abline(model_pres_diff) #Adding regression line
8 dev.off()

```



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

R code to save the residuals of the model in a separate object:

```
1 residuals_pres_diff <- model_pres_diff$residuals
```

4. Write the prediction equation.

The general form for bivariate equation is: $\hat{Y} = b_0 + b_1X_1$

Where,

\hat{Y} is the predicted value of the outcome variable,

b_0 is the y-intercept (the value of \hat{Y} when X_1 is 0),

b_1 is the slope of the regression line, representing the change in \hat{Y} for a one-unit change in X_1 ,

X_1 is the explanatory variable

In the context of the model, the predicted equation is:

presvote = 0.508 + 0.024 * *difflog*

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 code to run the regression and create the output:

```
1 model_vote_pres <- lm(voteshare ~ presvote, data = inc.sub)
2 summary(model_vote_pres) #Exploring the output of the model
3
4 #Reporting Regression results
5 stargazer(model_vote_pres,
6           title = "Model 3 Regression Results")
```

Table 3: Model 3 Regression Results

	<i>Dependent variable:</i>
	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

Interpretation:

The above model shows that one unit increase in the vote share of the presidential candidate of the incumbent's party is associated with, on average, a 0.388 unit increase in incumbent's vote share.

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

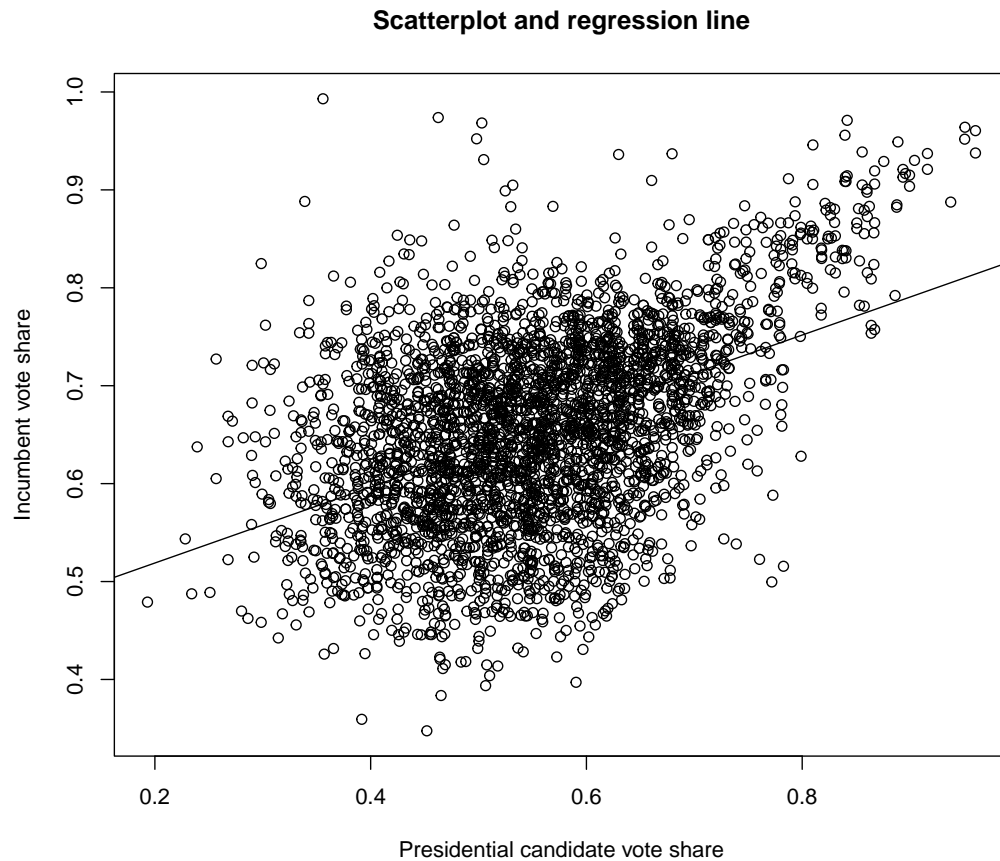
R code to create the scatter plot and add the regression line:

```
1 pdf("plot3.pdf", width=8)
2 plot(inc.sub$presvote,
```

```

3   inc.sub$voteshare,
4   xlab = "Presidential candidate vote share",
5   ylab = "Incumbent vote share",
6   main = "Scatterplot and regression line") #Scatter plot
7 abline(model_vote_pres) #Adding regression line
8 dev.off()

```



3. Write the prediction equation. The general form for bivariate equation is: $\hat{Y} = b_0 + b_1 X_1$

Where,

\hat{Y} is the predicted value of the outcome variable,

b_0 is the y-intercept (the value of \hat{Y} when X_1 is 0),

b_1 is the slope of the regression line, representing the change in \hat{Y} for a one-unit change in X_1 ,

X_1 is the explanatory variable

In the context of the model, the predicted equation is:

$$voteshare = 0.441 + 0.388 * presvote$$

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 code to run the regression and create the output:

```
1 model_res <- lm(residuals_vote_diff ~ residuals_pres_diff)
2 summary(model_res)
3
4 #Reporting Regression results
5 stargazer(model_res,
6           title = "Model 4 Regression Results")
```

Table 4: Model 4 Regression Results

	<i>Dependent variable:</i>
	residuals_vote_diff
residuals_pres_diff	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

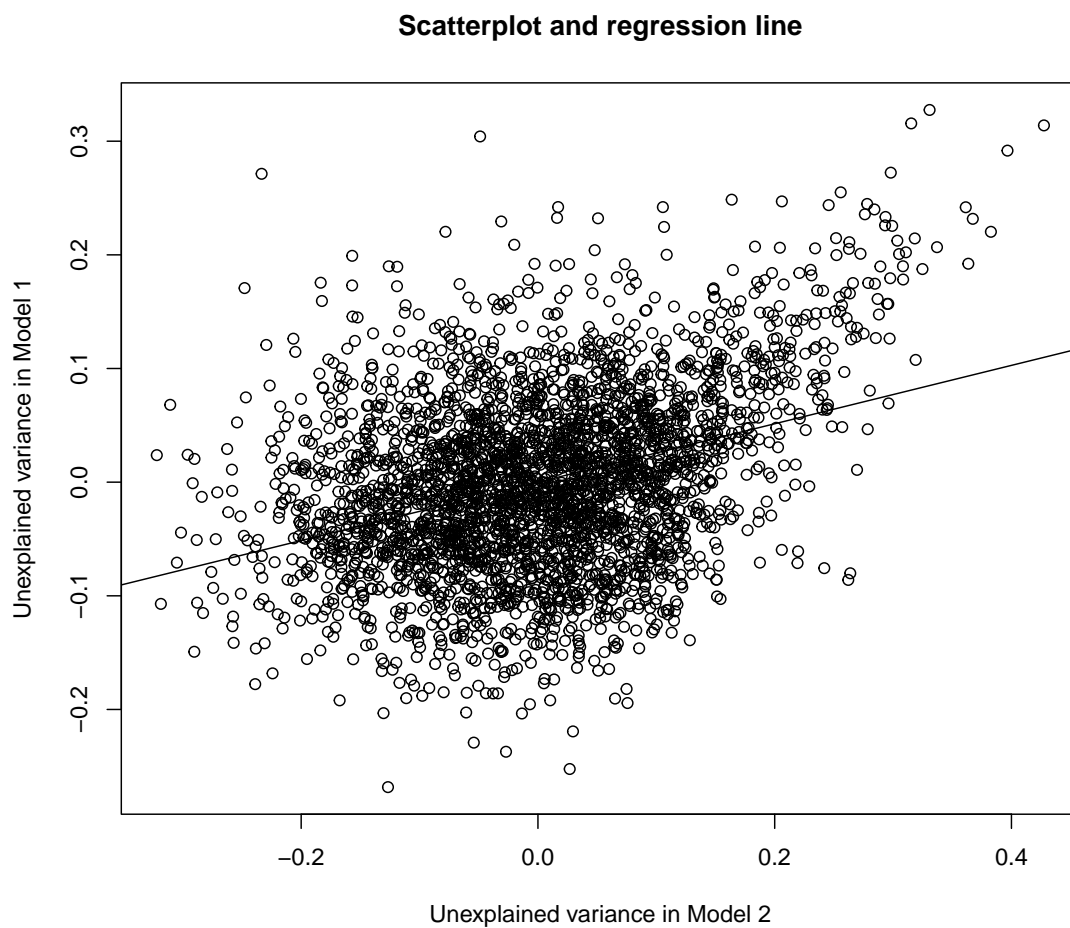
Interpretation:

The above model shows that one unit increase in residuals from Question 2 (i.e, unexplained variation in **presvote** by the log difference in spending between incumbent and challenger) is associated with, on average, a 0.257 unit increase in residuals from Question 1 (i.e, unexplained variation in **voteshare** by the log difference in spending between incumbent and challenger).

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

R code to create the scatter plot and add the regression line:

```
1 pdf("plot4.pdf", width=8)
2 plot(residuals_pres_diff,
3      residuals_vote_diff,
4      xlab = "Unexplained variance in Model 2",
5      ylab = "Unexplained variance in Model 1",
6      main = "Scatterplot and regression line") #Scatter plot
7 abline(model_res) #Adding regression line
8 dev.off()
```



3. Write the prediction equation.

The general form for bivariate equation is: $\hat{Y} = b_0 + b_1X_1$

Where,

\hat{Y} is the predicted value of the outcome variable,

b_0 is the y-intercept (the value of \hat{Y} when X_1 is 0),

b_1 is the slope of the regression line, representing the change in \hat{Y} for a one-unit change in X_1 ,

X_1 is the explanatory variable

In the context of the model, the predicted equation is:

$$residuals_vote_diff = -5.934e^{-18} + 0.257 * residuals_pres_diff$$

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 code to run the regression and create the output:

```
1 model_vote_diff_pres <- lm(voteshare ~ difflog + presvote, data = inc.sub
  )
2 summary(model_vote_diff_pres)
3
4 #Reporting Regression results
5 stargazer(model_vote_diff_pres,
6           title = "Model 5 Regression Results")
```

Table 5: Model 5 Regression Results

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

Interpretation:

The above model shows that holding the president's popularity constant, one unit increase in log difference in spending between incumbent and challenger is associated with, on average, a 0.036 unit increase in incumbent's vote share.

Furthermore, the model shows that holding the log difference in spending between incumbent and challenger constant, one unit increase in president's popularity is associated with, on average, a 0.257 unit increase in incumbent's vote share.

2. Write the prediction equation.

The general form for multivariate equation is: $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$

Where,

\hat{Y} is the predicted value of the outcome variable,

b_0 is the y-intercept (the value of \hat{Y} when X_1, X_2, \dots, X_n are 0),

b_1, b_2, \dots, b_n are the slopes of the regression model, representing the change in \hat{Y} for a one-unit change in X_1, X_2, \dots, X_n respectively keeping all other values constant,

X_1, X_2, \dots, X_n are the explanatory variables

In the context of the model, the predicted equation is:

$$voteshare = 0.448 + 0.0355 * difflog + 0.256 * presvote$$

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 coefficient of **presvote** in the model from Question 5 is identical to the coefficient of residuals from Question 2 (i.e, unexplained variation in **presvote** by the log difference in spending between incumbent and challenger) in the model from Question 4, i.e., 0.257.

Looking at the regression model for Question 4, we understand how much the residuals of Question 2 contribute to explaining the variance in the residuals of Question 1. Now, the regression model for Question 5 includes both **presvote** and log difference in spending between incumbent and challenger as predictor for incumbent's vote share. Here, the coefficient of **presvote** in Question 5 represents the partial impact of **presvote** on incumbent's vote share, accounting for the relationship between **presvote** and the residuals of Question 1 which is similar to Question 4.