Problem Set 3

Applied Stats/Quant Methods 1

Due: November 11, 2024

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 11, 2024. 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.

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`.

   Using the function lm() in R, I name my regression q_1_regression and use the dataset I named inc.sub.

```
> q_1_regression <- lm(formula = voteshare ~ difflog, data = inc.sub)
> summary(q_1_regression)

Call:
lm(formula = voteshare ~ difflog, data = inc.sub)
```

```
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   <2e-16 ***
difflog     0.041666   0.000968   43.04   <2e-16 ***
---
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: < 2.2e-16
```

I am going to repeat this general process for Question 2 and 3.

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

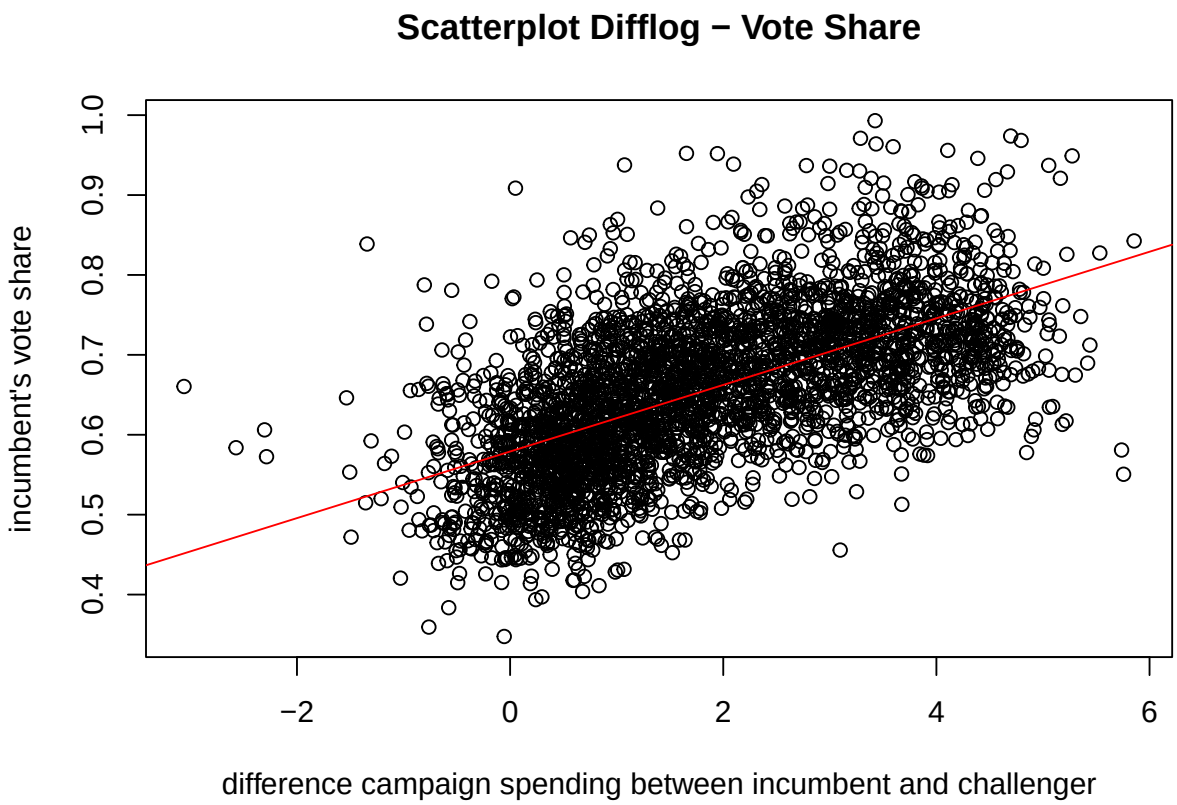
   First I create a scatterplot using

   ```
   > plot(inc.sub$difflog, inc.sub$voteshare)
   ```

   Then (so the line is visible on top of all datapoints) I create a red line on top of the graph using

   ```
   abline(lm(voteshare ~ difflog, data = inc.sub), col="red")
   ```

   This gives me the following graph:



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

   I named the object to save my residuals in q_1_residuals and created it in the following way. I am going to repeat this pattern in Question 2 as well.

```
> q_1_residuals <- q_1_regression$residuals
```

4. Write the prediction equation.

   Using the numbers calculated with

```
> q_1_regression <- lm(formula = voteshare ~ difflog, data = inc.sub)
```

   I get
   Y = 0.57903+0.04167x
   Or in our specific case:

voteshare = 0.57903 + 0.04167*difflog
Incumbent's Vote Share = 0.57903 + 0.04167 * difference in campaign spending between incumbent and challenger

Question 2

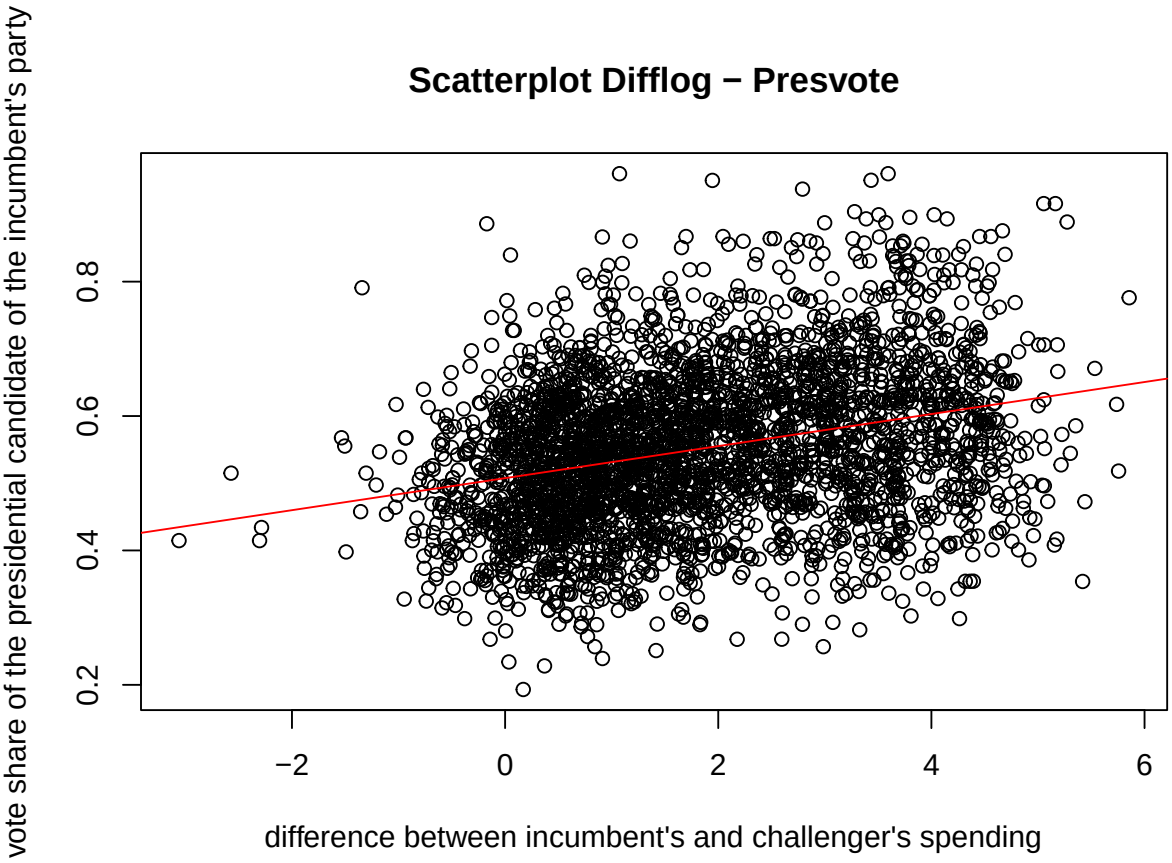
We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

1. Run a regression where the outcome variable is `presvote` and the explanatory variable is `difflog`.

```
> q_2_regression<-lm(formula = presvote ~ difflog, data = inc.sub)
> summary(q_2_regression)

Call:
lm(formula = presvote ~ difflog, data = inc.sub)

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   <2e-16 ***
difflog     0.023837   0.001359   17.54   <2e-16 ***
---
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: < 2.2e-16
```

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

```
> plot(inc.sub$difflog, inc.sub$presvote,
+      xlab = "difference between incumbent's and challenger's spending",
+      ylab = "vote share of the presidential candidate of the incumbent's party",
+      main = "Scatterplot Difflog - Presvote")
> abline(lm(presvote ~ difflog, data = inc.sub), col="red")
```



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

```
q_2_residuals <- lm(formula = presvote ~ difflog, data = inc.sub)$residuals
```

4. Write the prediction equation. Using the numbers calculated with

```
q_2_regression<-lm(formula = presvote ~ difflog, data = inc.sub)
```

I get
Y=0.50758+0.02384*X
Or in our specific case:
presvote = 0.50758 + 0.02384*difflog
vote share of the presidential candidate of the incumbent's party = 0.50758 + 0.02384 * difference in campaign spending between incumbent and challenger

Question 3

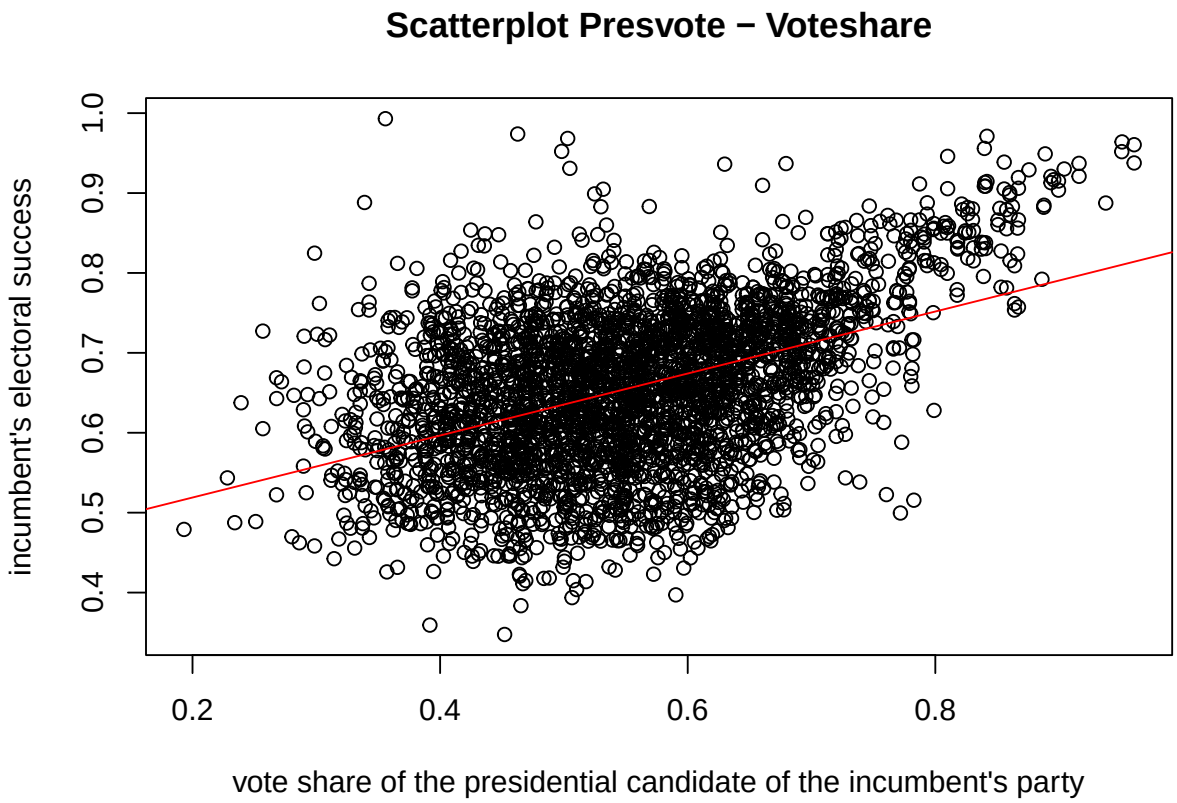
We are interested in knowing how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success.

1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`.

```
> q_3_regression<-lm(formula = voteshare ~ presvote, data = inc.sub)
> summary(q_3_regression)

Call:
lm(formula = voteshare ~ presvote, data = inc.sub)

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   <2e-16 ***
presvote    0.388018   0.013493   28.76   <2e-16 ***
---
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: < 2.2e-16
```

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

```
> plot(inc.sub$presvote, inc.sub$voteshare,
+      xlab = "vote share of the presidential candidate of the incumbent's party"
+      ylab = "incumbent's electoral success",
+      main = "Scatterplot Presvote - Voteshare")
> abline(lm(voteshare ~ presvote, data = inc.sub), col="red")
```



3. Write the prediction equation.

   Using the numbers calculated with

   ```
   > q_3_regression<-lm(formula = voteshare ~ presvote, data = inc.sub)
   ```

   I get
   Y=0.4413+0.3880X
   Or in our specific case:
   voteshare = 0.4413 + 0.3880*presvote
   incumbent's electoral success = 0.4413 + 0.3880 * vote share of the presidential candidate of the incumbent's party

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.

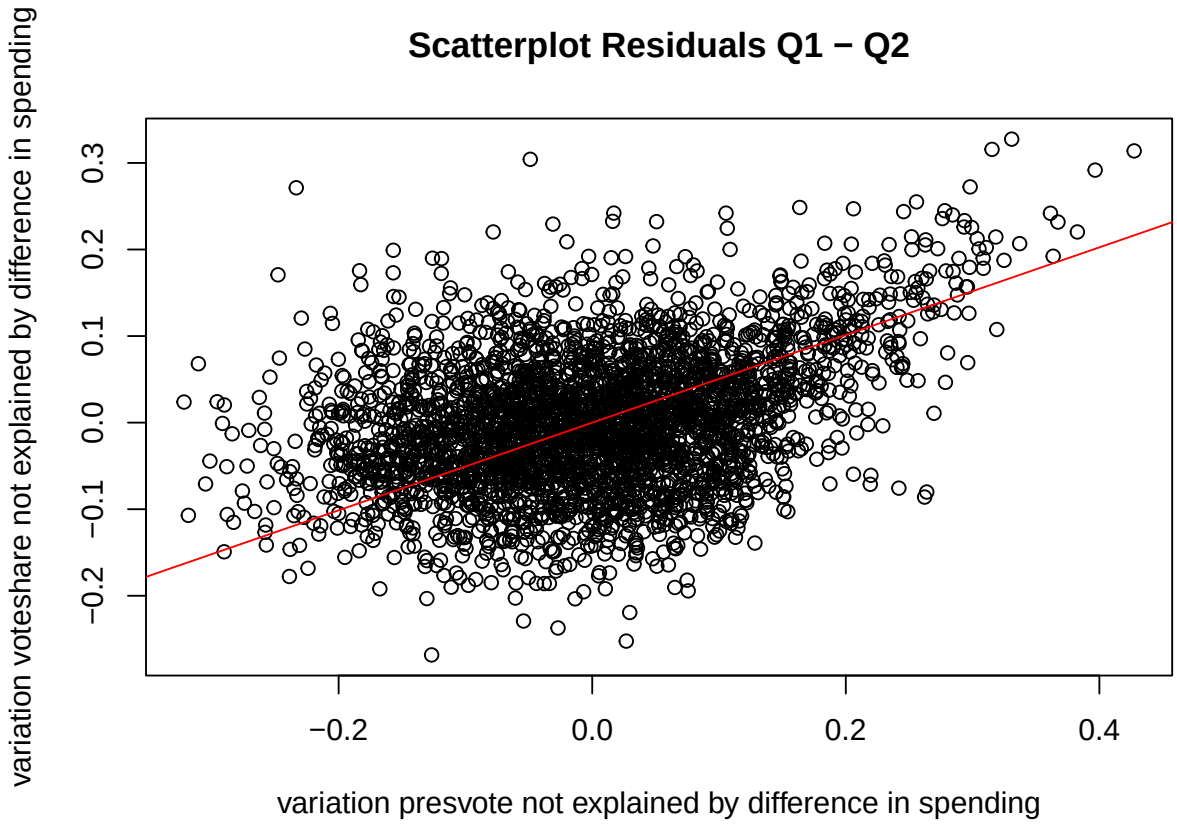
1. Run a regression where the outcome variable is the residuals from Question 1 and the explanatory variable is the residuals from Question 2.
   q_1_residuals and q_2_residuals are the saved residuals from Question 1 and Question 2 respectively. Using these, I just run the regression in the same way as before, only this time I am not specifying the dataset as the variables that I created myself are not tied directly to the dataset inc.sub.

```
> q_4_regression<-lm(formula = q_1_residuals ~ q_2_residuals)
> summary(q_4_regression)

Call:
lm(formula = q_1_residuals ~ q_2_residuals)

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)   -1.942e-18  1.299e-03    0.00        1
q_2_residuals  2.569e-01  1.176e-02   21.84   <2e-16 ***
---
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: < 2.2e-16
```

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

```
> plot(q_2_residuals, q_1_residuals,
+      xlab = "variation presvote not explained by difference in spending",
+      ylab = "variation voteshare not explained by difference in spending",
```

```
+      main = "Scatterplot Residuals Q1 - Q2")
> abline(lm(formula = q_2_residuals ~ q_1_residuals, data = inc.sub), col="red")
```



3. Write the prediction equation. Using the numbers calculated with

```
> q_4_regression<-lm(formula = q_1_residuals ~ q_2_residuals)
```

I get:
Y=-1.942e-18+2.569e-01 * X
Or in our specific case:
Q_1_residuals = 1.942e-18 + 2.569e-01 * Q_2_residuals
amount of variation in voteshare not explained by the difference in spending between incumben and challenger = 1.942e-18 + 2.569e-01 * amount of variation in presvote not explained by the difference in spending between incumbent and challenger in the district
The intercept is very close to 0 (almost negligible).

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`.

   I essentially run the same regression lm() in R as before, only this time I connect the second explanatory variable to the first with a "+"

```
> q_5_regression<-lm(voteshare ~ difflog + presvote, data = inc.sub)
> summary(q_5_regression)

Call:
lm(formula = voteshare ~ difflog + presvote, data = inc.sub)

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   <2e-16 ***
difflog     0.0355431  0.0009455   37.59   <2e-16 ***
presvote    0.2568770  0.0117637   21.84   <2e-16 ***
---
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: < 2.2e-16
```

2. Write the prediction equation.

   I get:
   Y=0.44864+0.03554 * X_1+0.25688*X_2
   Or in our specific case:
   voteshare = 0.44864 + 0.03554 * difflog+0.25688 * presvote
   incumbent's vote share = 0.44864 + 0.03554 * difference in spending between incumbent and challenger+0.25688 * president's popularity

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 for q_2_residuals in q_4_regression (0.2569) is identical to the coefficient for presvote in q_5_regression (0.2569).

   This is because q_4_regression is regressing the residuals of voteshare difflog on the residuals of presvote difflog. This isolates the effect of presvote on voteshare while controlling for difflog.

   In q_5_regression, both difflog and presvote are directly included in the model as explanatory variables, so the effect of presvote on voteshare in the presence of difflog is similarly isolated.

   This isolation also explains why the standard error and t-value for presvote in q_5_regression match the standard error and t-value for q_2_residuals in q_4_regression. Both sets of statistics reflect the same underlying relationship between voteshare and presvote while controlling for difflog.

   The residual summaries are also identical across models because both produce residuals based on the relationship between voteshare and the combination of difflog and presvote.

   Applying this explanation to the real meanings of the variables, both models isolate the influence of the incumbent's party's presidential success on the incumbent's vote share, while controlling for the difference in campaign spending between incumbent and challenger.

   Tldr: coefficient, standard error and t-value for presvote and the residual summary statistics are identical in both outputs because, in both models, the effect of presvote on voteshare is isolated while controlling for difflog, either by regressing the residuals (q_4_regression) or by including both variables in a multivariate model (q_5_regression).