

# 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`.

```
1 # read in data
2 inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/incumbents_subset.csv")
3 inc.sub
4
5 # Run a regression where the outcome variable is voteshare and the
   explanatory variable is difflog:
6 # Run the regression
7 modell <- lm(voteshare ~ difflog, data = inc.sub)
8 summary(modell)
```

```

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

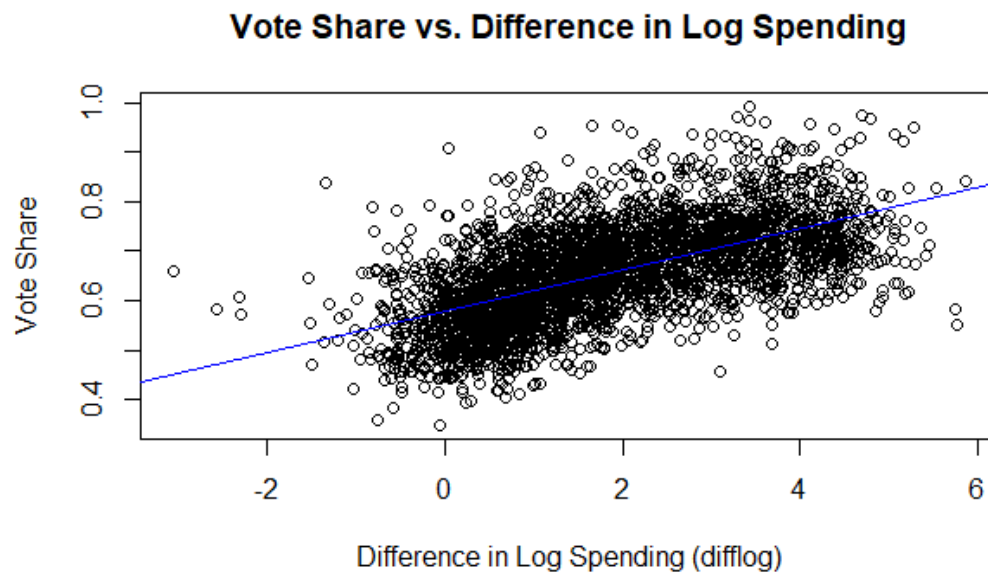
```

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

```

1 plot(inc.sub$difflog, inc.sub$voteshare,
2       main = "Vote Share vs. Difference in Log Spending",
3       xlab = "Difference in Log Spending (difflog)",
4       ylab = "Vote Share")
5 abline(model1, col = "blue")

```



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

```
1 residuals_model1 <- residuals(model1)
2 residuals_model1
```

4. Write the prediction equation.

$$y = \beta_0 + \beta_1 x \quad (1)$$

## 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 model2 <- lm(presvote ~ difflog, data = inc.sub)
2 summary(model2)
```

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 *** |

---

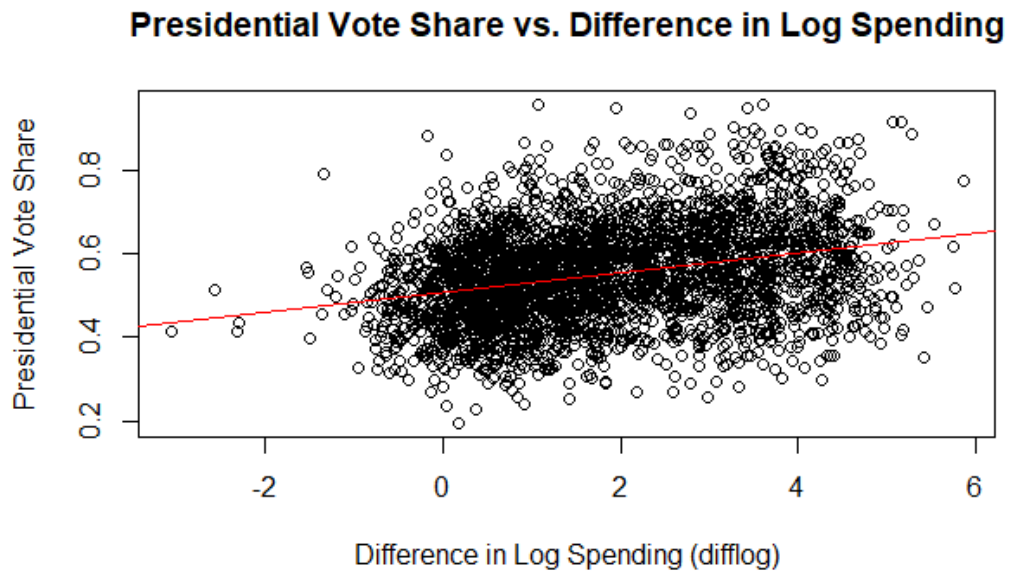
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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.



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

```
1 residuals_model2 <- residuals(model2)
```

4. Write the prediction equation.

$$y = \beta_0 + \beta_1 x \quad (2)$$

## 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 model3 <- lm(voteshare ~ presvote, data = inc.sub)
2 summary(model3)
```

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

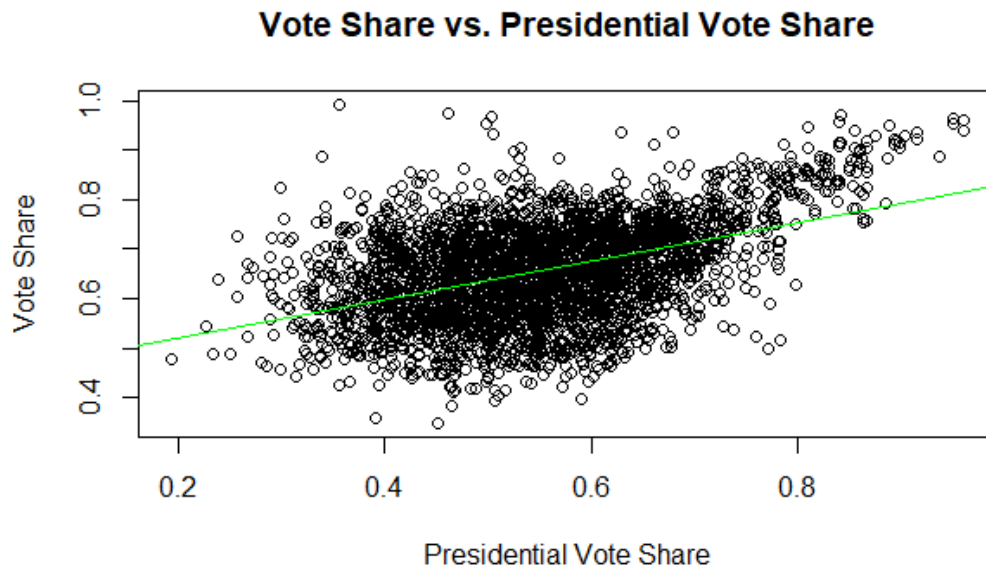
Residual standard error: 0.08815 on 3191 degrees of freedom

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.

```
1 plot(inc.sub$presvote, inc.sub$voteshare,
2       main = "Vote Share vs. Presidential Vote Share",
3       xlab = "Presidential Vote Share",
4       ylab = "Vote Share")
5 abline(model3, col = "green")
```



3. Write the prediction equation.

$$y = \beta_0 + \beta_1 x \quad (3)$$

## 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 model4 <- lm(residuals_model1 ~ residuals_model2)
2 summary(model4)
```

Call:

```
lm(formula = residuals_model1 ~ residuals_model2)
```

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)      | -4.860e-18 | 1.299e-03  | 0.00    | 1          |
| residuals_model2 | 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.

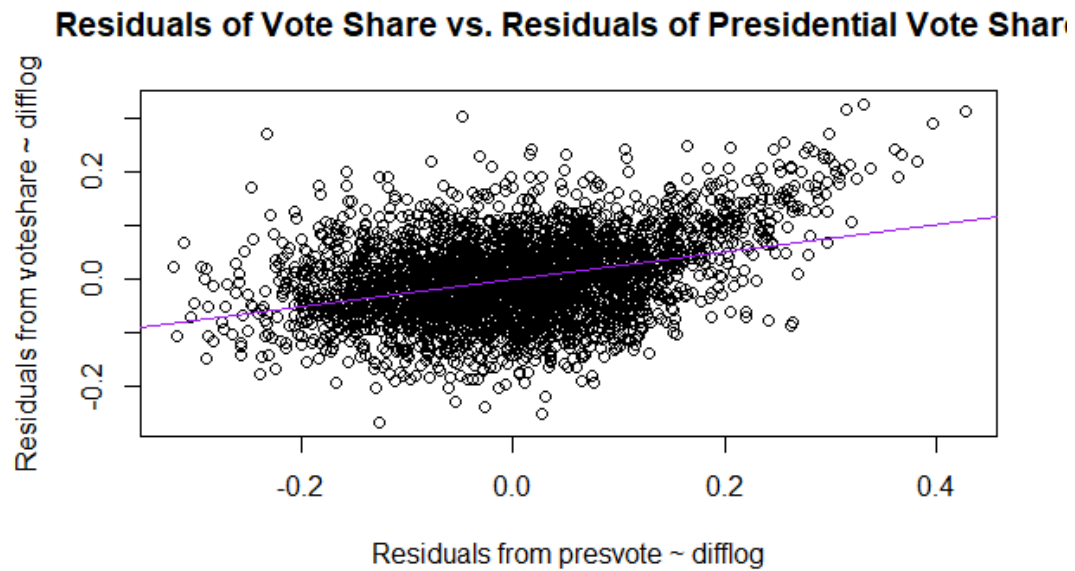
```
1 plot(residuals_model2, residuals_model1,
```



```

2   main = "Residuals of Vote Share vs. Residuals of Presidential Vote
    Share",
3   xlab = "Residuals from presvote ~ difflog",
4   ylab = "Residuals from voteshare ~ difflog")
5   abline(model4, col = "purple")

```



3. Write the prediction equation.

$$y = \beta_0 + \beta_1 x \quad (4)$$

## 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 model5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
2 summary(model5)
```

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.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (5)$$

3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case?

In question 4, the residuals from `voteshare ~ difflog` are regressed on the residuals from `presvote ~ difflog`. The residuals of `presvote` and `vote share` were used in model 4

In question 5, the multiple regression of `voteshare` on both `difflog` and `presvote` achieves a similar goal by adjusting for `difflog` within the regression model itself. The coefficient for `presvote` in this model reflects its unique contribution to predicting `voteshare`, independent of `difflog`.

Both approaches yield similar coefficients for `presvote` because they measure the same partial effect, isolating the relationship between `presvote` and `voteshare` without the influence of `difflog`.