# 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`.

Table 1:

| | *Outcome 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 |

```
# Question 1
# Part 1 - Load the Data
> library(readr)
> incumbents_subset <- read_csv("~/Desktop/GitHub/StatsI_Fall2024/
    datasets/incumbents_subset.csv")
# View the Data
sum(incumbents_subset)

# QUESTION 1

# Part 1: Run a regression where the outcome variable is vote share and
    the explanatory variable
# is difflog.
# Regression analysis:
# Fit the Model
model <- lm(voteshare ~ difflog, data = incumbents_subset)
# View the summary of the model
summary(model)

install.packages("stargazer")
library(stargazer)
stargazer(model, type = "text", title = "Regression Results", digits = 3)
stargazer(model)

# Load ggplot2 package
```

```
install.packages("ggplot2")
library(ggplot2)
```

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

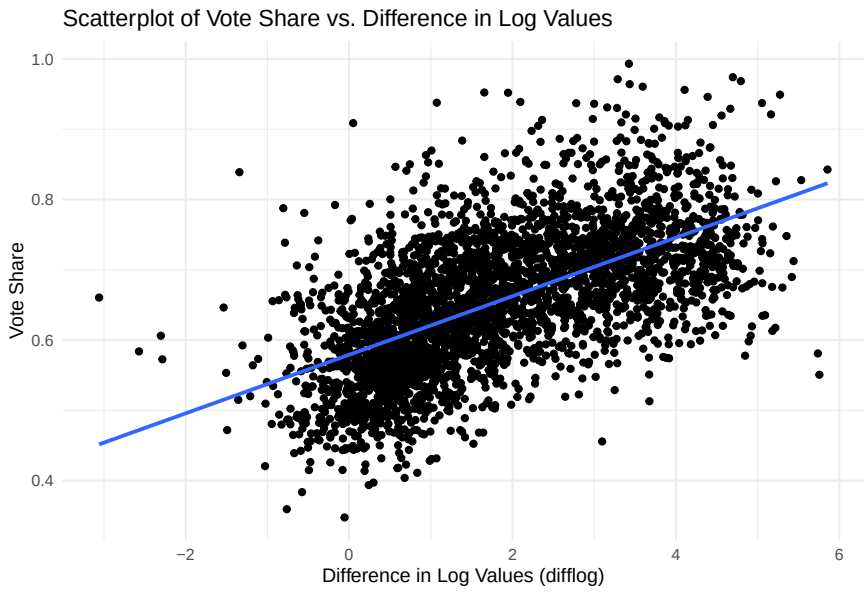


Figure 1: Scatterplot of Vote Share vs Difference in Log Values

```
#Part 2. Make a scatterplot of the two variables and add the regression
pdf("ScatterPlot1")
print(
  ggplot(incumbents_subset, aes(x = difflog, y = voteshare)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Difference in Log Values (difflog)",
         y = "Vote Share",
         title = "Scatterplot of Vote Share vs. Difference in Log Values"
    ) +
    theme_minimal()
)
dev.off()
```

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

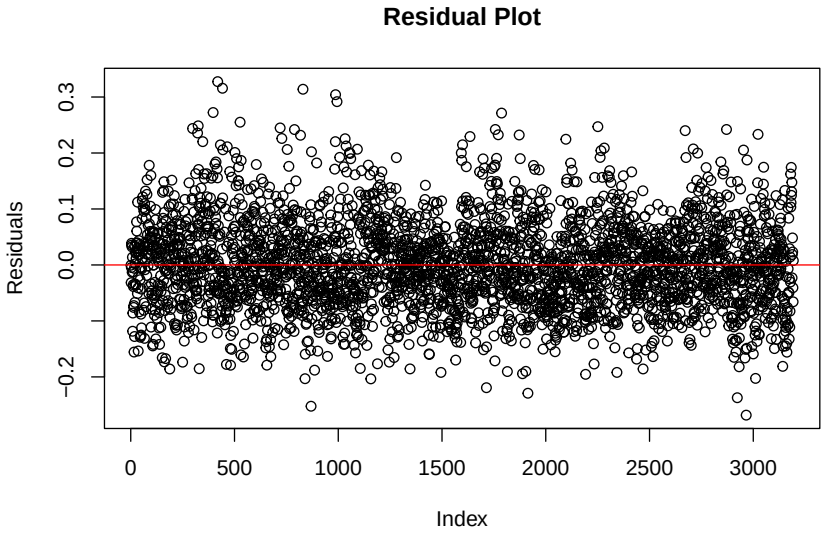


Figure 2: Residual Plot Visualisation

```r
#Save the residuals in separate object
residuals_object <- resid(model)
print(residuals_object)
summary(residuals_object)
#I thought it might be useful for later questions to also place them in a
    plot.
plot(residuals_object, main = "Residual Plot", ylab = "Residuals", xlab =
    "Index")
abline(h = 0, col = "red")
```

4. Write the prediction equation.

The standard bivariate regression equation is:

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

For this fitted model, the dependent variable ($y$) is `voteshare`, and the independent variable ($x$) is `difflog`. The estimated intercept ($\beta_0$) is 0.579, and the estimated slope coefficient ($\beta_1$) is 0.042. Thus, the fitted model prediction equation is:

$$\text{voteshare} = 0.579 + 0.042 \times \text{difflog} + \epsilon$$

# 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`.

Table 2:

| | *Outcome 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 |

```
#1. 1. Run a regression where the outcome variable is presvote and the
    explanatory variable
#is difflog
model2 <- lm(presvote ~ difflog, data = incumbents_subset)
summary(model2)
stargazer(model2)
#2. 2. Make a scatterplot of the two variables and add the regression
pdf("ScatterPlot2.pdf")
print(
  ggplot(incumbents_subset, aes(x = difflog, y = presvote)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Difference in Log Values (difflog)",
         y = "Presidential Vote",
         title = "Scatterplot of Presidential Vote vs. Difference in Log
    Values") +
    theme_minimal()
)
```

```
dev.off()
```

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

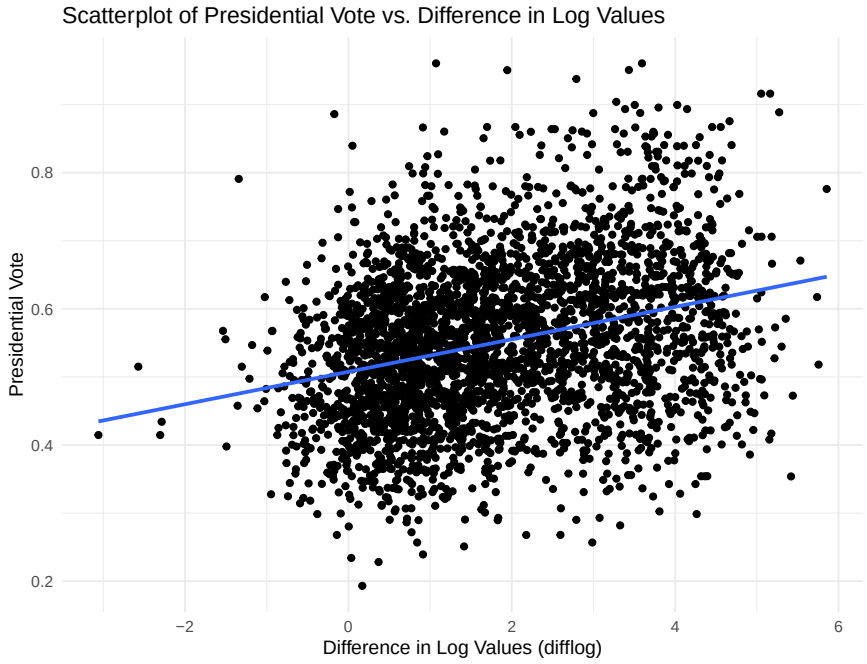


Figure 3: Scatterplot of Presidential Vote vs Difference in Log Values

```
#2. 2. Make a scatterplot of the two variables and add the regression
pdf("ScatterPlot2.pdf")
print(
  ggplot(incumbents_subset, aes(x = difflog, y = presvote)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Difference in Log Values (difflog)",
         y = "Presidential Vote",
         title = "Scatterplot of Presidential Vote vs. Difference in Log
   Values") +
    theme_minimal()
)
dev.off()
#Save the residuals in separate object
```

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

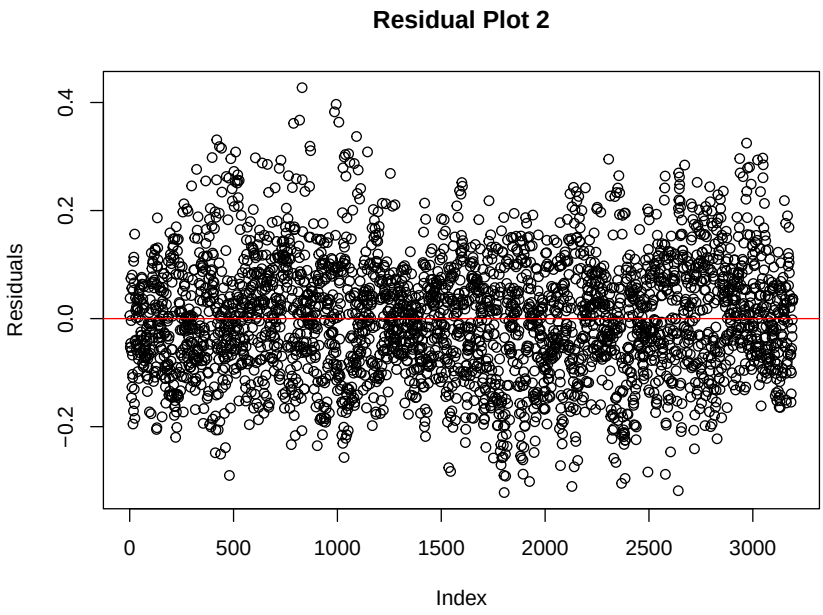


Figure 4: Residual Plot Visualisation

```r
#Save the residuals in separate object
residuals_object2 <- resid(model2)
print(residuals_object)
summary(residuals_object)
# As previously, I thought it might be useful to view them in a plot
plot(residuals_object2, main = "Residual Plot 2", ylab = "Residuals",
    xlab = "Index")
abline(h = 0, col = "red")
```

4. Write the prediction equation.

   The prediction equation for this model is:

$$\text{presvote} = 0.507583 + 0.023837 \times \text{difflog} + \epsilon$$

# 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`.

Table 3:

| | *Outcome 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 |

```
# Question 3

#1. Run a regression where the outcome variable is voteshare and the
    explanatory variable
#is presvote
model3 <- lm(voteshare ~ presvote, data = incumbents_subset)
summary(model3)
stargazer(model3)
```

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

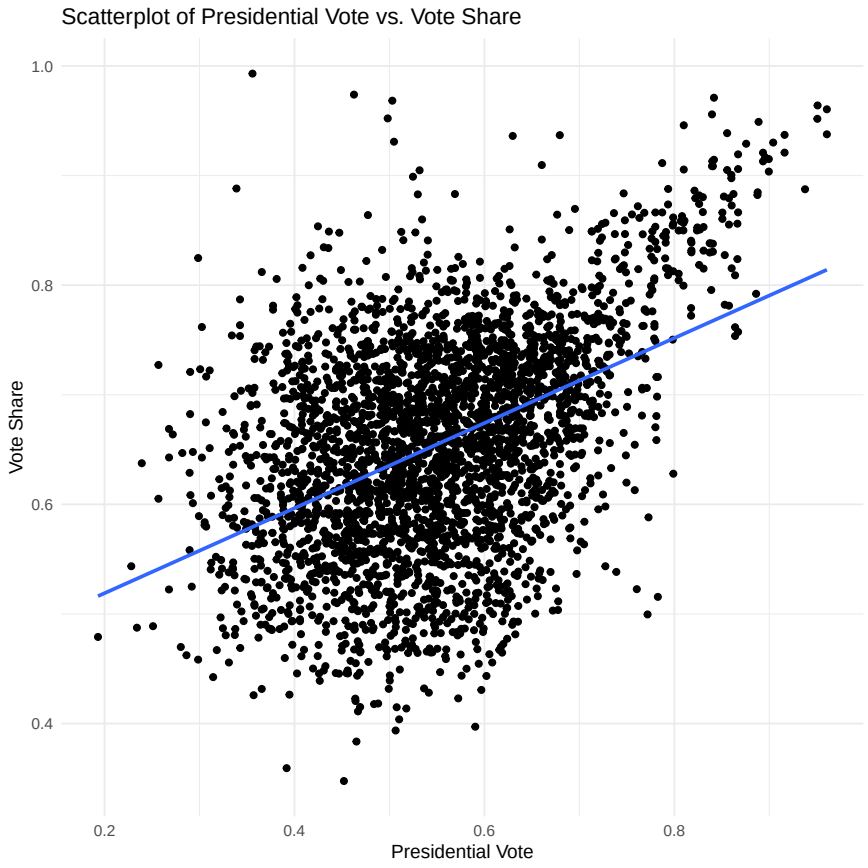


Figure 5: Scatterplot of Vote Share vs Presidential Votes

```
#2. Make a scatterplot of the two variables and add the regression
pdf("ScatterPlot3.pdf")
print(
  ggplot(incumbents_subset, aes(x = presvote, y = voteshare)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Presidential Vote",
         y = "Vote Share",
         title = "Scatterplot of Presidential Vote vs. Vote Share") +
    theme_minimal()
)
dev.off()
```

3. Write the prediction equation.

   The prediction equation is as follows:

$$\text{voteshare} = 0.441330 + 0.388018 \times \text{presvote} + \epsilon$$

# 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.

Table 4:

| | *Dependent variable:* |
|---|---|
| | residuals_object |
| residuals_object2 | 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 |

```
# Question 4
# Run a regression where the outcome variable is the residuals from
    Question 1
#and the explanatory variable is the residuals from Question 2.

model4 <- lm(residuals_object ~ residuals_object2, data = incumbents_
    subset)
View(model4)
summary(model4)
stargazer(model4)
```

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

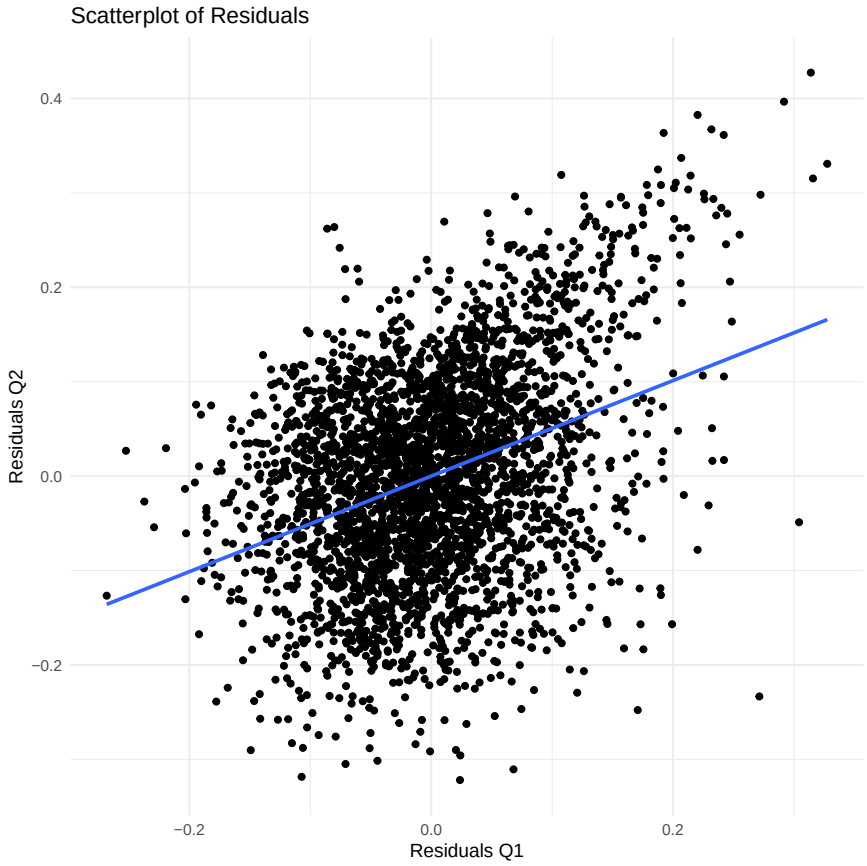


Figure 6: Scatterplot of Residuals

```
##2. Make a scatterplot of the two variables and add the regression line
pdf("ScatterPlot4.pdf")
print(
  ggplot(incumbents_subset, aes(x = residuals_object, y = residuals_object2))
   +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Residuals Q1",
         y = "Residuals Q2",
         title = "Scatterplot of Residuals") +
    theme_minimal()
)
dev.off()
```

3. Write the prediction equation.

$$\text{residuals\_object} = -1.942 \times 10^{-18} + 0.2569 \times \text{residuals\_object2} + \epsilon$$

# 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`.

Table 5:

| | *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.

$$\text{voteshare} = 0.4486442 + 0.0355431 \times \text{difflog} + 0.2568770 \times \text{presvote} + \epsilon$$

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 regression of the residuals in Table 4 reveals a positive and statistically significant coefficient (0.257), this coefficient is identical to the presvote coefficient in Table 5 (again, 0.257). This indicates that the unexplained variation in voteshare is positively associated with the unexplained variation in presvote. The fact that the coefficient for presvote is consistent in Table 5 suggests a strong relationship between the residuals of voteshare and presvote, which is not affected when we control for differences in campaign spending

```r
# Question 5
# Run the regression model
model5 <- lm(voteshare ~ difflog + presvote, data = incumbents_subset)
summary(model5)
stargazer(model5)
```