Problem Set 3

Applied Stats/Quant Methods 1

Name: Darragh McGee (18319331)

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`.

Linear Regression Assumptions

- **Linear Relationship:** There is a linear relationship between the outcome and explanatory variables.
- **Independence of Errors:** The errors (residuals) are independent of each other.
- **Normality of errors:** For any given value of the explanatory variable, the errors (residuals) are assumed to follow a normal distribution.
- **Constant variance (Homoscedasticity)**: The variance of the errors is constant across all values of the explanatory variable.
- **No Perfect Multi-Collinearity:** The explanatory variables should not be perfectly correlated with each other.

Read in Data:

```
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2024/main/datasets/incumbents_subset.csv")
```

Regression Model (1) for `voteshare` and `difflog`:

```
model_1 <- lm(voteshare ~ difflog, data = inc.sub)
summary(model_1)
```

Table 1: Summary of Model 1 Regression Results

Variable	**Estimate**	**Std. Error**	**Significance**
Intercept	0.579031	0.002251	***
presvote	0.041666	0.000968	***
Model Summary			
Residual Std. Error	0.07867		
Multiple R-squared	0.3673		
Adjusted R-squared	0.3671		
Number of Observations	3192		

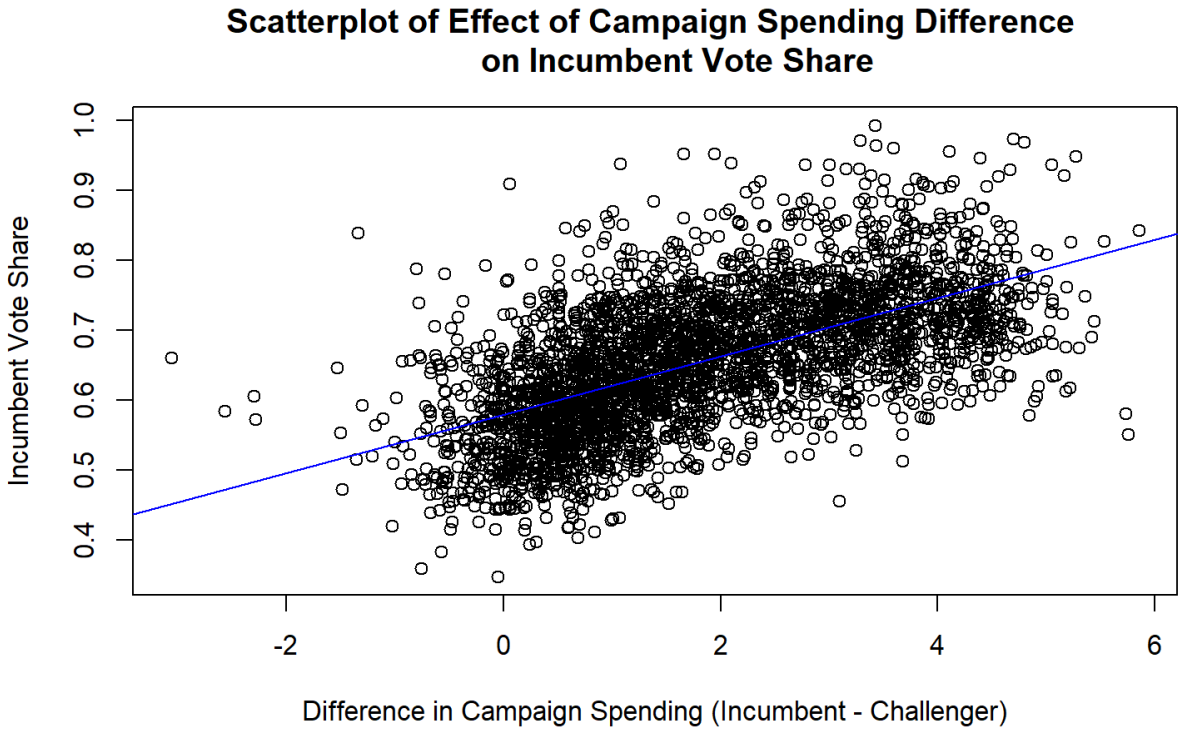
Significance Codes:
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- The intercept of 0.579031 indicates that when the logarithmic difference in campaign spending between the incumbent and challenger (`difflog`) is zero, the incumbent's predicted vote share (`voteshare`) is 57.9 percent.
- There is a positive, statistically significant relationship between `difflog` and `voteshare`, with a p-value of less than 0.001 (indicated by three stars in the regression table).
- Specifically, a one-unit increase in `difflog` is associated with an average increase of 0.04 (or 4 percentage points) in `voteshare`, holding other factors constant.

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

Code to Plot the Relationship:

```
plot(voteshare ~ difflog, data = inc.sub,
     xlab = "Difference in Campaign Spending (Incumbent - Challenger)",
     ylab = "Incumbent Vote Share",
     main = "Scatterplot of Effect of Campaign Spending Difference
     on Incumbent Vote Share")
abline(model_1, col = "blue", lwd = 1)
```



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

Save Residuals as a Separate Object:

```
residuals.model.1 <- residuals(model_1)
```

4. Write the prediction equation.

   Below is the Typical Linear Regression Prediction Equation Structure:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{1}$$

   where:

- $\hat{y}$ is the predicted value of the outcome variable,
- $\hat{\beta}_0$ is the estimated intercept,
- $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ are the estimated coefficients for each explanatory variable,
- $x_1, x_2, \ldots, x_p$ are the explanatory variables.

The first model is a bivariate model meaning there is a single outcome and explanatory variable. `voteshare` is the outcome variable and `difflog` is the explanatory variable.

$$\text{voteshare} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{difflog} \tag{2}$$

Inputting the Values from Model 1 Regression Output:

$$\text{voteshare} = 0.5709 + 0.4167 \cdot \text{difflog} \tag{3}$$

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`.

Regression Model (2) for `presvote` and `difflog`:

```
model_2 <- lm(presvote ~ difflog, data = inc.sub)
summary(model_2)
```

Table 2: Summary of Model 2 Regression Results

Variable	**Estimate**	**Std. Error**	**Significance**
Intercept	0.507583	0.003161	***
difflog	0.023837	0.001359	***
Model Summary			
Residual Std. Error	0.07867	(df = 3191)	
Multiple R-squared	0.3673		
Adjusted R-squared	0.3671		
Number of Observations	3192		

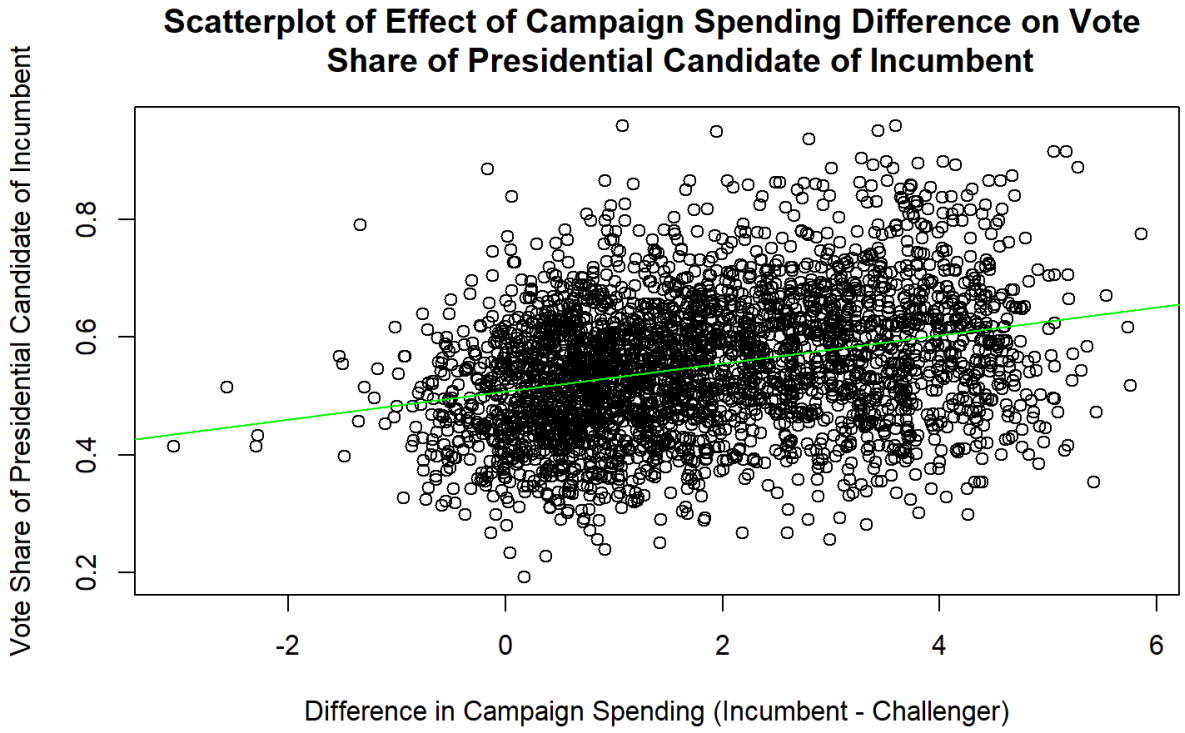
Significance Codes:
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- The intercept of 0.5075 indicates that when the logarithmic difference in campaign spending between the incumbent and challenger (`difflog`) is zero, the predicted vote share of the presidential candidate of the incumbent (`presvote`) is 50.75 percent.
- There is a positive, statistically significant relationship between `difflog` and `presvote`, with a p-value of less than 0.001 (indicated by three stars in the regression table).
- Specifically, a one-unit increase in `difflog` is associated with an average increase of 0.0238 (or 2.38 percentage points) in the vote share of the presidential candidate of the incumbent.

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

Code to Plot the Relationship:

```
plot(presvote ~ difflog, data = inc.sub,
     ylab = "Vote Share of Presidential Candidate of Incumbent",
     xlab = "Difference in Campaign Spending (Incumbent - Challenger)",
     main = "Scatterplot of Effect of Campaign Spending Difference on Vote
     Share of Presidential Candidate of Incumbent")
abline(model_2, col = "green", lwd = 1)
```



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

Save Residuals as a Separate Object:

```
residuals.model.2 <- residuals(model_2)
```

4. Write the prediction equation.

The second model is a bivariate model, meaning there is a single outcome and explanatory variable. `presvote` is the outcome variable, and `difflog` is the explanatory variable.

$$\text{presvote} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{difflog} \tag{4}$$

Inputting the values from the Model 2 Regression Output:

$$\text{presvote} = 0.5076 + 0.0238 \cdot \text{difflog} \tag{5}$$

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`.

   Regression Model (3) for `voteshare` and `presvote`:

```r
model_3 <- lm(voteshare ~ presvote, data = inc.sub)
summary(model_3)
```

Table 3: Summary of Model 3 Regression Results

Variable	**Estimate**	**Std. Error**	**Significance**
Intercept	0.441330	0.007599	***
presvote	0.388018	0.013493	***
Model Summary			
Residual Std. Error	0.08815	(df = 3191)	
Multiple R-squared	0.2058		
Adjusted R-squared	0.2056		
Number of Observations	3192		

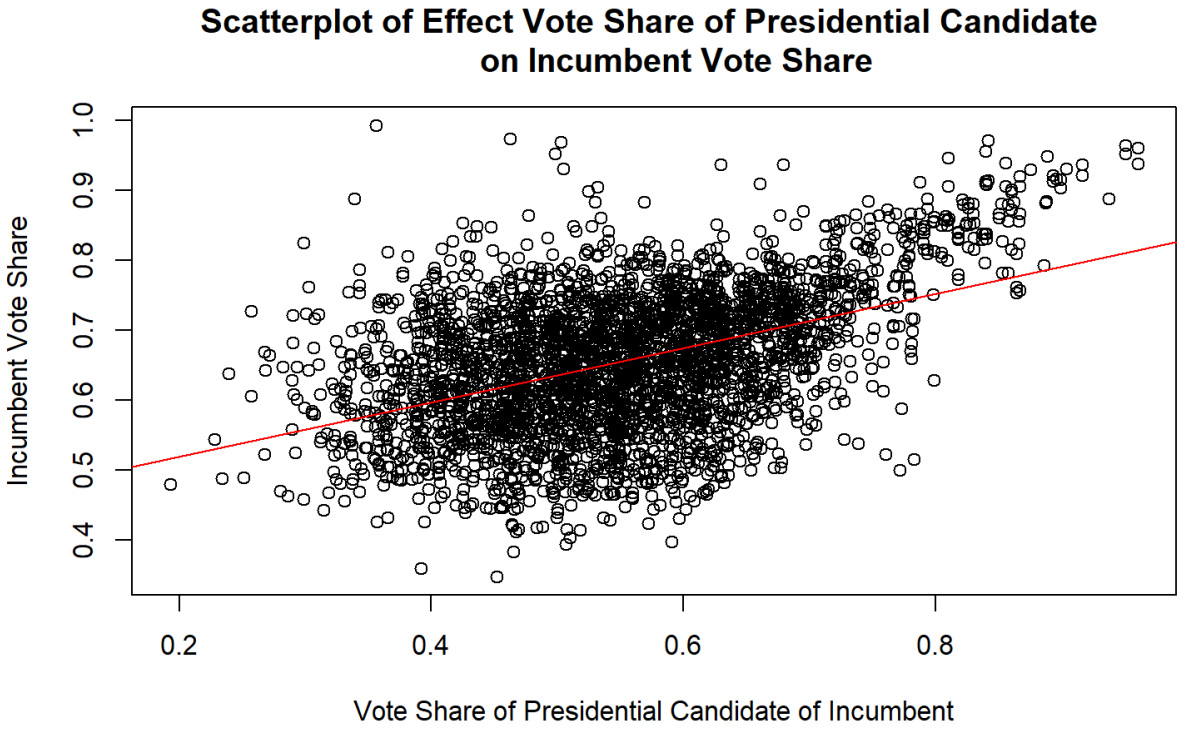
Significance Codes:
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- The intercept of 0.4413 indicates that when the vote share of the incumbent's presidential candidate (`presvote`) is zero, the predicted vote share of the incumbent (`voteshare`) is 44.13 percent.
- There is a positive, statistically significant relationship between `presvote` and `voteshare`, with a p-value of less than 0.001 (indicated by three stars in the regression table).
- Specifically, a one-unit increase in `presvote` is associated with an average increase of 0.388 (or 38.8 percentage points) in `voteshare`.

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

Code to Plot the Relationship:

```
png("Figure_3_1.png", width = 1500, height = 950, res = 200)
plot(voteshare ~ presvote, data = inc.sub,
```



3. Write the prediction equation.

The third model is a bivariate model, meaning there is a single outcome and explanatory variable. `voteshare` is the outcome variable, and `presvote` is the explanatory variable.

$$\text{voteshare} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{presvote} \tag{6}$$

Inputting the values from the Model 3 Regression Output:

$$\text{voteshare} = 0.4413 + 0.3880 \cdot \text{presvote} \tag{7}$$

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

Regression Model (4) for `residuals.model.1` and `residuals.model.2`:

```
model_4 <- lm(residuals.model.1 ~ residuals.model.2)
summary(model_4)
```

Table 4: Summary of Model 4 Regression Results

Variable	**Estimate**	**Std. Error**	**Significance**
Intercept	-5.934e-18	0.001299	n.s.
residuals.model.2	0.2569	0.01176	***
Model Summary			
Residual Std. Error	0.07338	(df = 3191)	
Multiple R-squared	0.1300		
Adjusted R-squared	0.1298		
Number of Observations	3192		

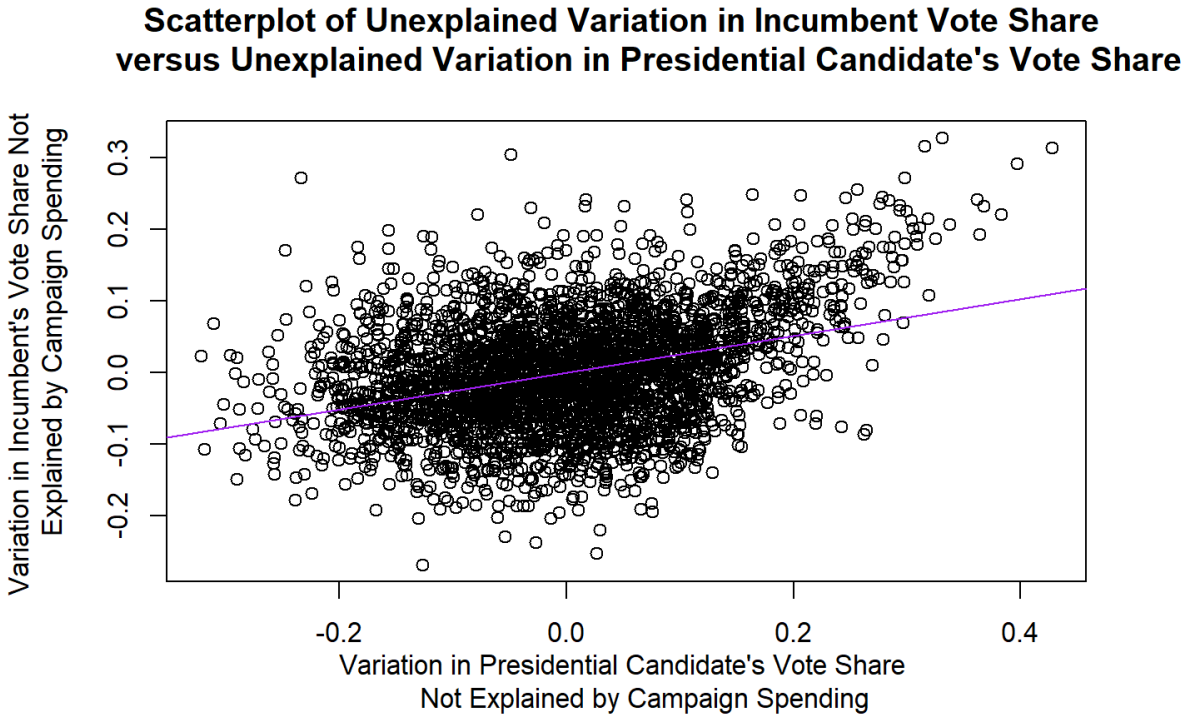
Significance Codes:
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- There is a positive, statistically significant relationship between the residuals from model 1 and model 2, with a p-value of less than 0.001 (indicated by three stars in the regression table).
- Specifically, a one-unit increase in the residual (error) from model 2 is associated with an average increase of 0.2569 in the residual (error) from model 1.

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

Code to Plot the Relationship:

```
plot(residuals.model.1 ~ residuals.model.2, data = inc.sub,
     xlab = "Variation in Presidential Candidate's Vote Share
     Not Explained by Campaign Spending",
     ylab = "Variation in Incumbent's Vote Share Not
     Explained by Campaign Spending",
     main = "Scatterplot of Unexplained Variation in Incumbent Vote Share
     versus Unexplained Variation in Presidential Candidate's Vote Share")
abline(model_4, col = "purple", lwd = 1)
```



3. Write the prediction equation.

The fourth model is a bivariate model, meaning there is a single outcome and explanatory variable. `residuals.model.1` is the outcome variable, and `residuals.model.2` is the explanatory variable.

$$\text{residuals.model.1} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{residuals.model.2} \tag{8}$$

Inputting the values from the Model 4 Regression Output:

$$\text{residuals.model.1} = 0 + 0.2569 \cdot \text{residuals.model.2} \tag{9}$$

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`.

Regression Model (5) for `voteshare`, `difflog`, and `presvote`:

```
model_5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
summary(model_5)
```

Table 5: Summary of Regression Results for Model: voteshare difflog + presvote

Variable	**Estimate**	**Std. Error**	**Significance**
Intercept	0.4486	0.0063	***
difflog	0.0355	0.0009	***
presvote	0.2569	0.0118	***
Model Summary			
Residual Std. Error	0.07339	(df = 3190)	
Multiple R-squared	0.4496		
Adjusted R-squared	0.4493		
Number of Observations	3193		

Significance Codes:
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- The intercept of 0.4486 represents the average predicted vote share of the incumbent when both `difflog` and `presvote` are zero.
- There is a positive, statistically significant relationship between the log-transformed difference in campaign spending (`difflog`) and the incumbent's vote share. Specifically, a one-unit increase in `difflog` is associated with an average increase of 3.55 percentage points in the incumbent's vote share, holding `presvote` constant.
- There is also a positive, statistically significant relationship between the presidential candidate's vote share (`presvote`) and the incumbent's vote share. Specifically, a one-unit increase in `presvote` is associated with an average increase of 25.69 percentage points in the incumbent's vote share, holding `difflog` constant.
- The multivariate regression model explains approximately 44.96% of the variation in `voteshare`, as indicated by the Multiple R-squared of 0.4496, suggesting a moderate fit.

2. Write the prediction equation.

The fourth model is a multivariate model, meaning there is a single outcome variable with multiple explanatory variables. `voteshare` is the outcome variable, while `difflog` and `presvote` are the explanatory variables.

$$\text{voteshare} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{difflog} + \hat{\beta}_2 \cdot \text{presvote} \tag{10}$$

Inputting the values from the Model 4 Regression Output:

$$\text{voteshare} = 0.4486 + 0.0355 \cdot \text{difflog} + 0.2569 \cdot \text{presvote} \tag{11}$$

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

   - The coefficients for `residuals.model.2` in Model 4 and `presvote` in Model 5 are the same because they both measure the same underlying relationship.
   - This relationship reflects the co-variation between the presidential candidate's vote share (`presvote`) and the incumbent's vote share (`voteshare`) that is not explained by the differences in campaign spending.
   - In Model 4, this relationship is measured through the residuals (unexplained variation), while in Model 5, it captures the partial effect of `presvote` directly.