Problem Set 3

Applied Stats/Quant Methods 1

Due: November 19, 2022

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.

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`.

   At the beginning, I imported the data set, familiarised myself with it, and then extracted relevant data for the Questions and confirmed that the remaining data frame `df` contained the correct data.

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

# Familiarise with CSV data
typeof(inc.sub) # check type
head(inc.sub, 5) # Checking first 5 observations and their values
colnames(inc.sub) # checking column names for which are relevant
dim(inc.sub) # Find number of rows (observations) and columns (variables)
```

```r

# Extract relevant data
df <- inc.sub[, c("difflog", "voteshare", "presvote")]

# Review extracted data, confirm correct
colnames(df) # checking column names
head(df, 5) # Checking first 5 observations and their values
dim(df) # number of rows (observations) and columns (variables)
```

I note that I do not have the code book for the data set, and so do not know the real units for each variable. Therefore, my answers will refer to changes in "unit points" for accuracy.

```r
# Assign model and inspect data through summary
diff_vs_share <- lm(voteshare ~ difflog, data = df)
summary(diff_vs_share)
```

Table 1: Linear Regression Results: Positive Correlation between Vote Share of Incumbent and Difference in Spending

	Dependent 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

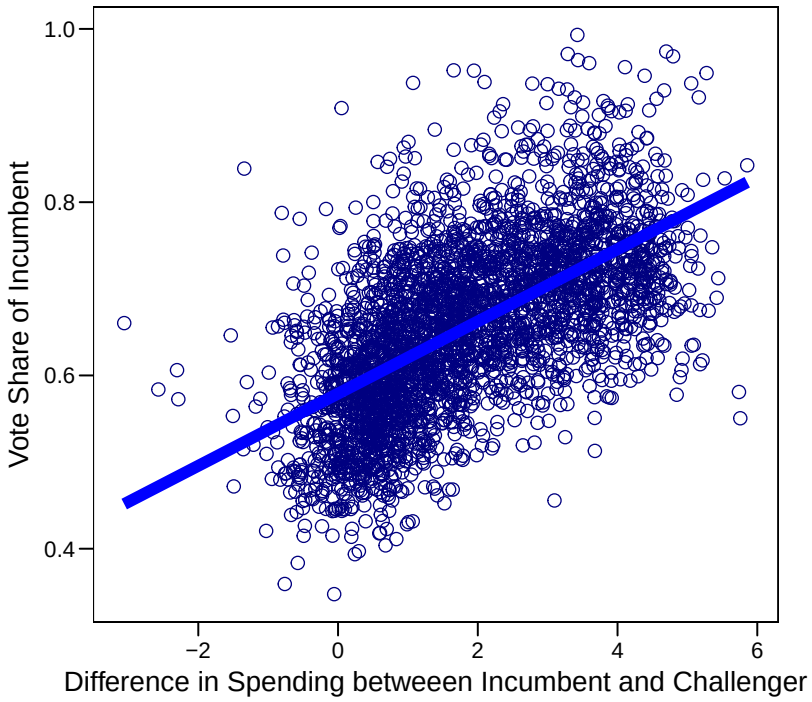
This bi-variate linear regression model indicates that for **every 1-unit increase** in the difference of campaign spending between the incumbent and challenger, on average, the vote share for the incumbent increases by **approx. 0.04 unit points**.

The coefficient has a p-value less than 0.001, which is less than a critical value of 0.05, giving evidence to reject the null hypothesis that the relationship between difference in campaign spending and incumbent vote share is zero - and giving evidence for the alternative hypothesis that the relationship is non-zero. It is not possible to indicate a causal relationship with this method.

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

```
pdf("plot_Q1.pdf")
# Set dataframe and axes for scatter plot
ggplot(df, aes(x = difflog, y = voteshare)) +
  # Set point symbol and size
  geom_point(pch = 1, size = 3, color = "navy") +
  # Set regression line details
  geom_smooth(data = diff_vs_share, method = "lm", se = FALSE, color = "
    blue", size = 3) +
  # Add axes labels, exclude main title
  labs(
    # Will add main = "Scatterplot and Regression Line: Positive
    Correlation between Vote Share of Incumbent and Difference in Spending
    " in LaTeX
    x = "Difference in Spending betweeen Incumbent and Challenger",
    y = "Vote Share of Incumbent") +
  # Set theme "par" and axis text formatting
  theme_par() +
  theme(
    axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15))
# Ensure graphics instructions end
dev.off()
```

Figure 1: Scatterplot and Regression Line: Positive Correlation between Vote Share of Incumbent and Difference in Spending.



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

```
# Assign residuals
residuals_diff_vs_share <- diff_vs_share$residuals
# Inspect output
head(residuals_diff_vs_share, 10)
typeof(residuals_diff_vs_share)
length(residuals_diff_vs_share) == dim(inc.sub)[1]
```

4. Write the prediction equation.

   Below is a generic formal prediction equation for a bi-variate linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

   The components are:

   (a) Y: The (predicted) value for the outcome variable

   (b) β_0: The y-intercept (i.e. when the explanatory variable equals zero).

   (c) β_1: The coefficient or slope. This represents the **average change** in Y with a one-unit change in the corresponding explanatory variable.

   (d) X_i: The value for the (input) explanatory variable.

   (e) ε: The **theoretical** (stochastic) error term, which represents the unobserved/unaccounted for factors which influence Y yet are not included in the model.

   Note a key assumption for linear regression models is that ε has a *constant and independent average effect* on one or several variables within the model which cancels itself out when given a large sample N (see notation below). Given this theoretical assumption, it is typically excluded from written prediction equations. This is why a prediction equation is also called the "least-square equation".

$$\varepsilon_i \overset{iid}{\sim} N(O, \sigma^2)$$

   Also note that the theoretical error ε is distinct from the observed residual e_i for each data point. Residuals are discussed in greater detail in Question 4.

   Below is the prediction equation for the linear regression where the outcome variable is `voteshare` and the explanatory variable is `difflog`.

$$\begin{aligned}&\text{Predicted vote share for the incumbent} = \\ &\beta_{\text{vote share for the incumbent when spending difference is zero}} \\ &+ (\beta_{\text{coefficient for 1-unit increase in spending difference}} \times X_{\text{spending difference}})\end{aligned}$$

Below is the prediction equation with the relevant values (rounded to two decimal places):

$$\text{Predicted vote share for the incumbent} = 0.58 + (0.04 * X_{\text{spending difference}})$$

This equation can be used to find the predicted (estimated) vote share for the incumbent, corresponding to the input value for the spending difference between the incumbent and challenger.

Caution should be maintained for inputting values of spending difference which are **greater than 5.86 and lower than -3.06**. These are the maximum and minimum observed data points for spending difference, and so the model cannot be assumed to be valid outside of these value thresholds.

```
max(df$difflog)
min(df$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
# Assign model and inspect through summary
diff_vs_presvote <- lm(presvote ~ difflog, data = df)
summary(diff_vs_presvote)
```

Table 2: Linear Regression Results: Positive Correlation between Vote Share of Presidential Candidate of Incumbent's Party and Difference in Spending

	Dependent 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

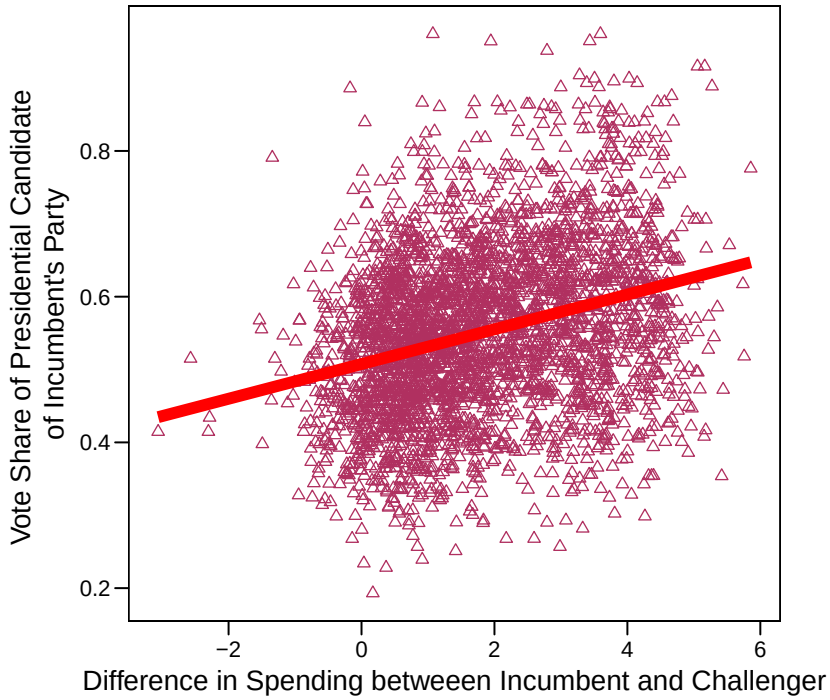
This bi-variate linear regression model indicates that for **every 1-unit increase** in the difference of campaign spending between the incumbent and challenger, on average, the vote share for the Presidential Candidate of the Incumbent's Party increases by **approx. 0.02 unit points**.

The coefficient has a p-value less than 0.001, which is less than a critical value of 0.05, giving evidence to reject the null hypothesis that the relationship between difference in campaign spending and vote share of the Incumbent Party's Presidential Candidate is zero - and giving evidence for the alternative hypothesis that the relationship is non-zero. It is not possible to indicate a causal relationship with this method.

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

```
pdf("plot_Q2.pdf")
# Set dataframe and axes for scatter plot
ggplot(df, aes(x = difflog, y = presvote)) +
  # Set point symbol and size
  geom_point(pch = 2, size = 2, color = "maroon") +
  # Set regression line details
  geom_smooth(data = diff_vs_presvote, method = "lm", se = FALSE, color =
    "red", size = 3) +
  # Add axes labels, exclude main title
  labs(
    # Will add main = "Scatterplot and Regression Line: Positive
    Correlation between Vote Share of Presidential Candidate of Incumbent'
    s Party and Difference in Spending" in LaTeX
    x = "Difference in Spending betweeen Incumbent and Challenger",
    y = "Vote Share of Presidential Candidate \nof Incumbent's Party") +
  # Set theme "par" and axis text formatting
  theme_par() +
  theme(
    axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15))
# Ensure graphics instructions end
dev.off()
```

Figure 2: Scatterplot and Regression Line: Positive Correlation between Vote Share of Presidential Candidate of Incumbent's Party and Difference in Spending.



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

```
# Assign residuals
residuals_diff_vs_presvote <- diff_vs_presvote$residuals
# Check output
head(residuals_diff_vs_presvote, 10)
typeof(residuals_diff_vs_presvote)
length(residuals_diff_vs_presvote) == dim(inc.sub)[1]
```

4. Write the prediction equation.

   Below is the prediction equation for the linear regression where the outcome variable is `presvote` and the explanatory variable is `difflog`. See Question 1 for generalised discussion.

$$\begin{aligned}&\text{Predicted vote share for the incumbent party's presidential candidate} = \\ &\beta_{\text{vote share for the incumbent party's presidential candidate when spending difference is zero}} \\ &+ \beta_{\text{coefficient for 1-unit increase in spending difference}} X_{\text{spending difference}}\end{aligned}$$

   Below is the prediction equation with the relevant values (rounded to two decimal places):

$$\begin{aligned}&\text{Predicted vote share for the incumbent party's presidential candidate} = \\ &0.51 + (0.02 * X_{\text{spending difference}})\end{aligned}$$

   This equation can be used to find the predicted (estimated) vote share for the incumbent party's presidential candidate, corresponding to the input value for the spending difference between the incumbent and challenger.

   Caution should be maintained for inputting values of spending difference which are **greater than 5.86 and lower than -3.06**. These are the maximum and minimum observed data points for spending difference, and so the model cannot be assumed to be valid outside of these value thresholds. See Question 1, Part 4 for code checking min/max value.

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`.

```
# Assign model and inspect through summary
presvote_vs_share <- lm(voteshare ~ presvote, data = df)
summary(presvote_vs_share)
```

Table 3: Linear Regression Results: Positive Correlation between Vote Share of Presidential Candidate of Incumbent's Party and Vote Share of Incumbent

	Dependent 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

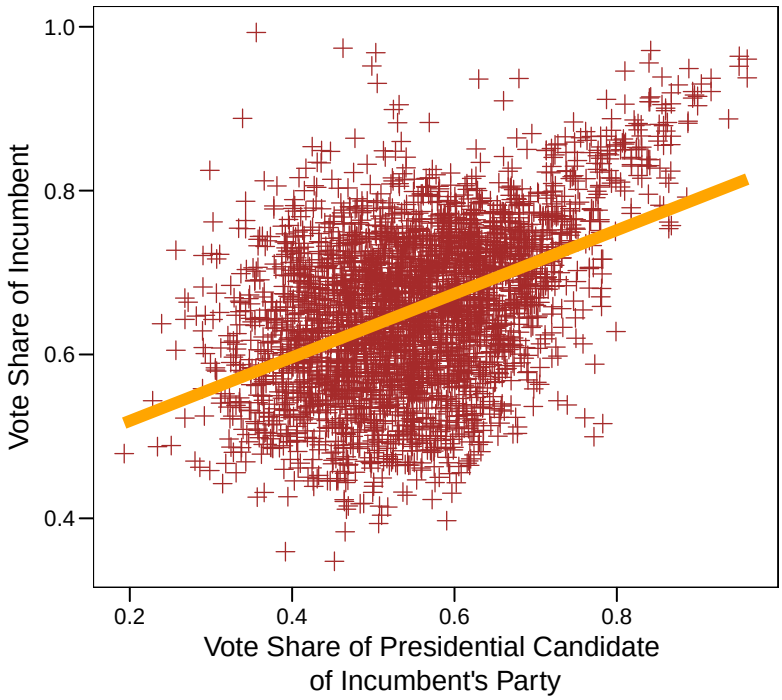
This bi-variate linear regression model indicates that for **every 1-unit increase** in the Vote Share of the Presidential Candidate of the Incumbent's Party, on average, the Vote Share for the Incumbent increases by **approx. 0.39 unit points**.

The coefficient has a p-value less than 0.001, which is less than a critical value of 0.05, giving evidence to reject the null hypothesis that the relationship between the vote share of the Incumbent Party's Presidential Candidate and the vote share of the Incumbent is zero - and giving evidence for the alternative hypothesis that the relationship is non-zero. It is not possible to indicate a causal relationship with this method.

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

```r
pdf("plot_Q3.pdf")
# Set dataframe and axes for scatter plot
ggplot(df, aes(x = presvote, y = voteshare)) +
  # Set point symbol and size
  geom_point(pch = 3, size = 3, color = "brown") +
  # Set regression line details
  geom_smooth(data = presvote_vs_share, method = "lm", se = FALSE, color
    = "orange", size = 3) +
  # Add axes labels, exclude main title
  labs(
    # Will add main = "Scatterplot and Regression Line: Positive
    Correlation between Vote Share of Presidential Candidate of Incumbent'
    s Party and Vote Share of Incumbent" in LaTeX
     x = "Vote Share of Presidential Candidate \nof Incumbent's Party",
     y = "Vote Share of Incumbent") +
  # Set theme "par" and axis text formatting
  theme_par() +
  theme(
    axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15))
# Ensure graphics instructions end
dev.off()
```

Figure 3: Scatterplot and Regression Line:
Positive Correlation between Vote Share of Presidential Candidate of Incumbent's Party
and Vote Share of Incumbent.



3. Write the prediction equation.

   Below is the prediction equation for the linear regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`. See Question 1 for generalised discussion.

$$\begin{aligned}&\text{Predicted vote share for the incumbent} = \\ &\beta_{\text{vote share for the incumbent when the incumbent party's presidential candidate has zero vote share}} \\ &+ \beta_{\text{coefficient for 1-unit increase in vote share for the incumbent party's presidential candidate}} \\ &\quad X_{\text{vote share for the incumbent party's presidential candidate}}\end{aligned}$$

   Below is the prediction equation with the relevant values (rounded to two decimal places):

$$\begin{aligned}&\text{Predicted vote share for the incumbent} = \\ &0.44 + (0.39 * X_{\text{vote share for the incumbent party's presidential candidate}})\end{aligned}$$

   This equation can be used to find the predicted (estimated) vote share for the incumbent, corresponding to the input value for the vote share for the incumbent party's presidential candidate.

   Caution should be maintained for inputting values of vote share for the incumbent party's presidential candidate which are **greater than 0.96 and lower than 0.19**. These are the maximum and minimum observed data points for the given presidential candidate's vote share, and so the model cannot be assumed to be valid outside of these value thresholds.

```
max(df$presvote)
min(df$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.

```
# Assign model and inspect through summary
not_diff_share_vs_presvote <- lm(residuals_diff_vs_share ~ residuals_diff
    _vs_presvote)
summary(not_diff_share_vs_presvote)
```

Table 4: Linear Regression Results: Positive Correlation
between Variation in Vote Share of Presidential Candidate of Incumbent's Party
and Variation in Vote Share of Incumbent
NOT Correlated with Variation in Difference in Spending

	Dependent variable:
	residuals_diff_vs_share
residuals_diff_vs_presvote	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

As noted in Question 1, the theoretical error ε is distinct from the model's residuals e_i for each observed data point. Each 'residual' is the difference between the observed value and predicted value from the regression model for the outcome variable:

$$e_i = y - \hat{y}$$

Residuals can be used to calculate the total variation in the outcome variable NOT correlated with variation in the explanatory variable(s). This is called the residual sum

of squares ($SS_{Residuals}$) or Sum of Squares for Errors (SSE). This can be calculated as follows:

$$SSE = \sum_{i=1}^{n}](y_i - \hat{y}_i)^2$$

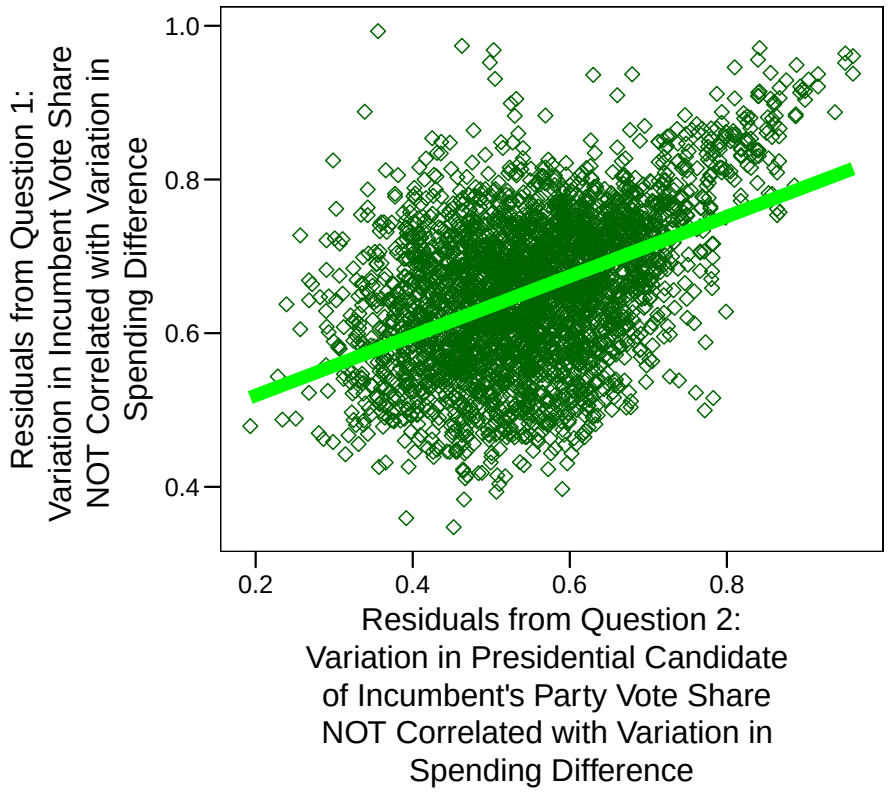
This bi-variate linear regression model indicates that there is a positive correlation (a.k.a. a positive linear relationship) between the "noise" in both regression models (which are comparable as they share a common data point i.e. the explanatory variable). In particular, for **every 1-unit increase** in the residuals from Question 1's model, on average, the residuals from Question 2's model increases by **approx. 0.257 unit points**.

The coefficient has a p-value less than 0.001, which is less than a critical value of 0.05, giving evidence to reject the null hypothesis that the relationship between the residuals of the two models is zero - and giving evidence for the alternative hypothesis that the relationship is non-zero. This indicates some **systematic dependence between the "noise" of the two models**, or in other words, that variation in the vote share of the Incumbent is **NOT independent** from variation in the vote share of Presidential Candidates of the Incumbent's Party. It is not possible to indicate a causal relationship with this method.

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

```
pdf("plot_Q4.pdf")
# Set dataframe and axes for scatter plot
ggplot(df, aes(x = presvote, y = voteshare)) +
  # Set point symbol and size
  geom_point(pch = 5, size = 2, color = "darkgreen") +
  # Set regression line details
  geom_smooth(data = presvote_vs_share, method = "lm", se = FALSE, color
    = "green", size = 3) +
  # Add axes labels, exclude main title
  labs(
    # Will add main title in LaTeX
    x = "Residuals from Question 2:\nVariation in Presidential Candidate
    \nof Incumbent's Party Vote Share \nNOT Correlated with Variation in \
    nSpending Difference",
    y = "Residuals from Question 1:\nVariation in Incumbent Vote Share \
    nNOT Correlated with Variation in \nSpending Difference") +
  # Set theme "par" and axis text formatting
  theme_par() +
  theme(
    axis.title.x = element_text(size = 15),
    axis.title.y = element_text(size = 15))
# Ensure graphics instructions end
dev.off()
```

Figure 4: Scatterplot and Regression Line:
Positive Correlation between the Variation in Vote Share of Incumbent and the Variation in Vote Share of Presidential Candidate of Incumbent's Party which are NOT Correlated with Variation in Spending Differences.



3. Write the prediction equation.

   Below is the prediction equation for the linear regression where the outcome variable is `residuals_diff_vs_share` from Question 1 and the explanatory variable is `residuals_diff_vs_presvote` from Question 2.

   Below is the prediction equation with the relevant values (rounded to two decimal places):

$$\text{Predicted "noise" in Q1 Model} = 0 + (0.26 * X_{\text{"noise" value from Q2 Model}})$$

   One way interpret this equation is that it indicates that larger positive residuals in one model are associated, on average, with positive residuals in the other model - or that the degree of noise in the two models is, on average, systematically dependent.

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
2 # Assign model and inspect data through summary
3 diff_and_presvote_vs_share <- lm(voteshare ~ difflog + presvote, data =
    df)
4 summary(diff_and_presvote_vs_share)
```

Table 5: Multi-Variate Linear Regression Results

	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

This multi-variate linear regression model indicates that for **every 1-unit increase** in the difference of campaign spending between the incumbent and challenger, on average, the vote share for the incumbent increases by **approx. 0.04 unit points**, controlling for the vote share of the President Candidate of the Incumbent's Party - in other words, 0.04 is the average partial effect of every "unit" of the difference of campaign spending between the incumbent and the challenger on the incumbent's vote share.

Similarly, the average partial effect of every "unit" of the vote share of the Presidential Candidate of the Incumbent's Party on the vote share of the Incumbent is 0.26 (i.e. holding the difference in spending constant).

Each of these partial effects indicate a p-value less than 0.001, which is less than a critical value of 0.05, giving evidence to reject the null hypothesises that the (partial) relationship between the vote share of the Incumbent and both the difference in campaign spending and the vote share of the Incumbent Party's Presidential Candidate is zero - and giving evidence for the alternative hypothesises that each of the relationships is non-zero. It is not possible to indicate a causal relationship with this method.

Further, I note that the R^2 value for the multi-variate regression model is 0.45, which is greater than all prior models where the vote share of the incumbent was the outcome variable (Q1 = 0.37; Q3 = 0.21). This indicates that including both explanatory variables informs the model to more accurately estimate the observed data of vote share of the Incumbent official compared to either explanatory variable individually.

Lastly, the F-Statistic has a p-value less than a critical value of 0.05, giving evidence to reject the null hypothesis that these is no relationship between either explanatory variables (vote share of Presidential Candidate of the Incumbent's Party; difference in spending between incumbent and challenger) and the outcome variable (vote share of the incumbent). This also gives evidence to fail to reject the alternative hypothesis that there is a (partial) relationship between at least one explanatory variable and vote share of the Incumbent.

2. Write the prediction equation.

   Below is a generic formal prediction equation for a multi-variate linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_i$$

   Notice how it is an extension of the bi-variate linear regression The components are:

   (a) Y: The outcome variable

   (b) β_0: The y-intercept (i.e. when the explanatory variable equals zero). When the model includes more than one explanatory variable, β_0 is the y-intercept with all explanans equal to zero.

   (c) $\beta_1, \beta_2, \ldots, \beta_n$: The coefficients. These represent the **average change** in Y with a one-unit change in the corresponding explanatory variable(s). When the model includes more than one explanatory variable, the coefficients represent the **partial effect** of each explanan, controlling for all other explanatory variables.

   (d) $X_1, X_2, \ldots, X_n$: The explanatory variable(s).

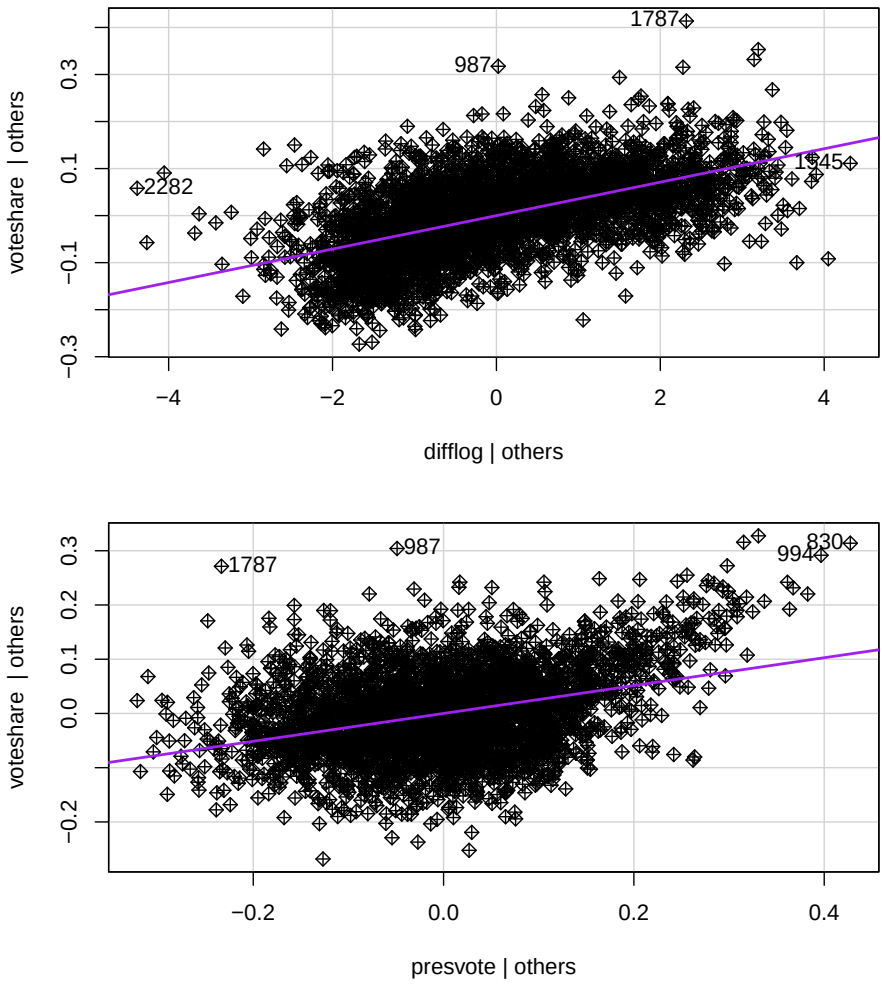
   (e) ε_i: The error (stochastic) term, which represents the unobserved/accounted factors which influence Y yet are not included within the model. Note a key assumption for linear regression models is that ε_i has a constant average effect which cancels out when given a large sample N. This is explained further in previous question.

Inputting the relevant labels and the values for the slopes (coefficients):

$$\begin{aligned}\text{Predicted vote share for the incumbent} = \\ & 0.45 \\ & + (0.04 * X_{\text{Difference in Campaign Spending}}) \\ & + (0.26 * X_{\text{Vote Share for Presidential Candidate}})\end{aligned}$$

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

Figure 5: Added Variable Plots: Partial Effects of Difference in Campaign Spending and of Vote Share of Presidential Candidate upon Vote Share of the Incumbent.



```
# Creating Added Variable Plots
pdf("plot_Q5.pdf", height = 8)
avPlots(diff_and_presvote_vs_share,
        main = "", # Will add "main = " title in LaTeX
        layout = c(2,1),
        pch = 9,
        col.lines = "purple")
dev.off()
```

In Table 4, the value of the coefficient for residuals of Q2 Model (predictor = Difference in Spending, outcome = vote share of the Presidential Candidate) on residuals for Q1 Model (predictor = Difference in Spending, outcome = vote share of the Incumbent) is **identical to the partial effect** of the coefficient for vote share of the Presidential Candidate on vote share of the Incumbent in Table 5.

This is because they are measuring the same phenomenon - the y-intercept in Table 4 is effectively 0, indicating that the regression effectively carves out all of the "noise" which the variation in vote share of the Presidential Candidate contributes to the variation in vote share of the Incumbent. This is the **partial effect** which is presented in Table 5. This is also demonstrated by how the "Residuals" values for both regressions (the minimum residual value, maximum residual value, median value, etc.) are identical (see R outputs) - these indicate the two regressions are measuring the same data points.

The "Residual Standard Error" in both Table 4 and Table 5 are also identical (with the multi-variate regression including one less degree of freedom as it necessarily uses an additional (*second*)value point in its calculation). This additional degree of freedom is also demonstrated in the degrees of freedom value and number of observations used in calculating the F-statistic for the multi-variable regression (Table 5).