1 PS03 Answers

Aryan Goyal
Student number: 18306046

First, I import the data into R

```
incumbent <- read.csv("https://raw.githubusercontent.com/
ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")
```

2 Question 1

2.1 Answer 1a)

I run a linear regression using the lm function where the outcome variable is voteshare and the explanatory variable is difflog

```
q1 <- lm(voteshare~difflog, data=incumbent)
summary(q1)
```

Table 1: Bivariate Regression between difflog(X) and vote share (Y)

	Dependent variable:
	Vote share
Constant	0.579***
	(0.002)
Difflog	0.042***
	(0.001)
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

Our null hypothesis is that there is no association between difference in campaign spending (X) and the incumbent's vote share (Y), $\beta_1 = 0$. β_1 represents the slope or the regression coefficient of the explanatory variable

The alternative hypothesis is that there is an association between difference in campaign spending (X) and the incumbent's vote share (Y), $\beta_1 \neq 0$.

As shown in Table 1, β_1 is significant at the 0.01 alpha level, hence we can reject the null hypothesis and we find support for the alternative hypothesis that there is an association between difference in campaign spending and incumbent's vote share, $\beta_1 \neq 0$. Moreover, we can say that for a 1 unit increase in difference in campaign spending (X), there is 0.042 scale points increase in incumbent's vote share(Y).
I am not interpreting the β_0 value as it does not make sense in this context. If the difference in campaign spending (X) = 0, the incumbent's vote share (Y) = 0.579 scale points which does not provide us with any relevant understanding.

2.2 Answer 1b)

I make a scatterplot with both variables along with a regression line. I also add the correlation value to understand the strength and direction of the linear relationship.

```
    plot(incumbent$difflog, incumbent$voteshare, main='Relationship between incumbent
    voteshare and difference in campaign spending',
    xlab='Difference in campaign spending',ylab='Vote Share')
abline(lm(voteshare~difflog, data=incumbent),col='red',lty='dashed')
cor(incumbent$voteshare,incumbent$difflog)
text(-2, 0.9,sprintf("Correlation=%s",
round(cor(incumbent$difflog,incumbent$voteshare),4)))
```

On the basis of the regression line and correlation value in Figure 1, we can see a positive strong correlation (0.6061) between difference in campaign spending and the vote share.

2.3 Answer 1c)

I save the residuals of the model in a separate object in R

```
    q1residuals <- q1$residuals ##Accessing the residuals using the "$" input
    head(q1residuals)
```

2.4 Answer 1d)

The prediction equation for a simple linear regression can be represented as:
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * X$
where:
$\hat{Y}$ is the predicted value of the outcome variable based on the regression equation
X is the explanatory variable
$\hat{\beta}_0$ is the estimated y-intercept or the constant term
$\hat{\beta}_1$ is the estimated coefficient of the explanatory variable indicating the slope of the line

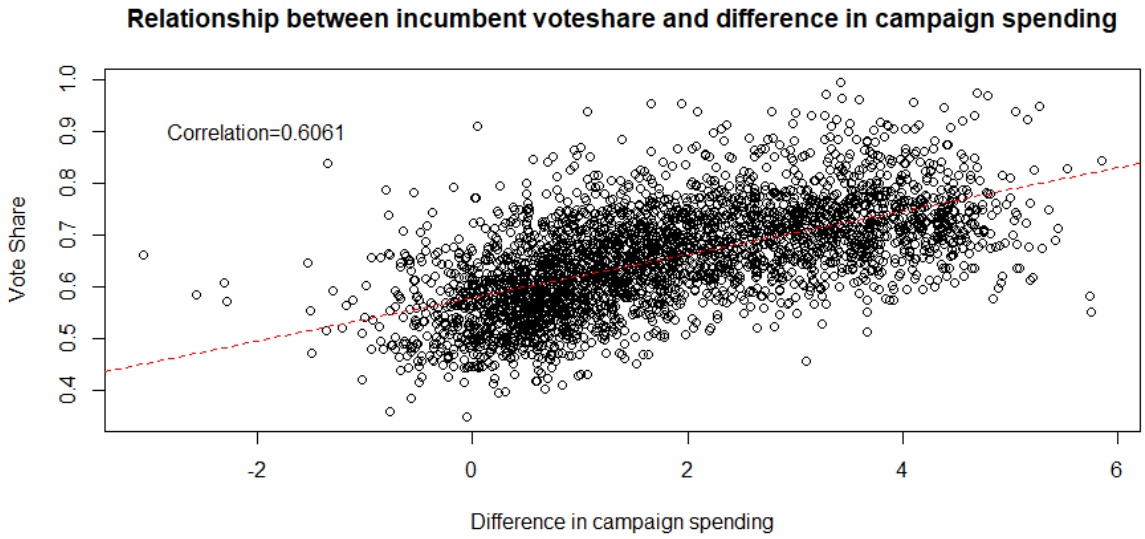

Figure 1: Relationship between Incumbent Vote Share and difference in Campaign Spending

The error term, ϵ is inherent in the statistical model, but is not explicitly included in the prediction equation. The error term accounts for the "noise" (variability) in the data that our model does not capture

Therefore, the prediction equation for our simple linear regression is:

$$voteshare = 0.579 + 0.042 * difflog$$

3 Question 2

3.1 Answer 2a)

I run a linear regression using the lm function where the outcome variable is presvote and the explanatory variable is difflog

```
q2 <- lm(presvote~difflog, data=incumbent)
summary(q2)
```

Our null hypothesis is that there is no association between difference in campaign spending (X) and the vote share of the presidential candidate of the incumbent party (Y), $\beta_1 = 0$

The alternative hypothesis is that there is an association between difference in campaign spending (X) and the vote share of the presidential candidate of the incumbent party (Y), $\beta_1 \neq 0$

Table 2: Bivariate regression between difflog(X) and presvote(Y)

	Dependent variable:
	Presvote
Constant	0.508^{***}
	(0.003)
Difflog	0.024^{***}
	(0.001)
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

As shown in Table 2, β_1 is significant at the 0.01 alpha level, hence we can reject the null hypothesis and we find support for the alternative hypothesis that there is an association between difference in campaign spending and vote share of the presidential candidate of the incumbent party, $\beta_1 \neq 0$. Moreover, we can say that for a 1 unit increase in difference in campaign spending (X), there is 0.024 scale points increase in the vote share of the presidential candidate of the incumbent party(Y).
Similar to Question 1, the β_0 is not interpreted as it does not provide us with relevant insight in this context.

3.2 Answer 2b)

I make a scatterplot with both variables along with a regression line and the correlation value:

```
    plot(incumbent$difflog,incumbent$presvote,
     main='Relationship between difference in campaign spending
     and vote share of the presidential candidate of the incumbent party',
     xlab='Difference in campaign spending',ylab='Vote share of presidential candidate')
abline(q2,col='red',lty='dashed')
cor(incumbent$presvote,incumbent$difflog)
text(-2, 0.9,sprintf("Correlation=%s", round(cor(incumbent$presvote,incumbent$difflog),4)))
```

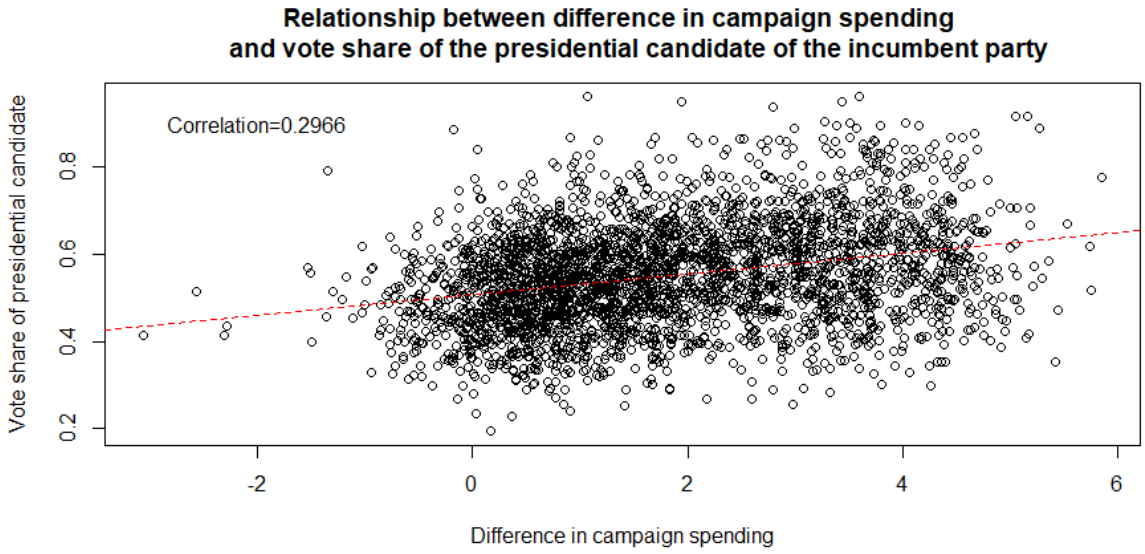

Figure 2: Relationship between difference in campaign spending and vote share of the presidential candidate of the incumbent party

On the basis of the regression line and the correlation value, we can say that there is a weak positive correlation (0.2966) between difference in campaign spending and vote share of presidential candidate of the incumbent party

3.3 Answer 2c)

I save the residuals of the model in a separate object in R:

```
q2residuals <- q2$residuals
head(q2residuals)
```

3.4 Answer 2d)

The prediction equation for our simple linear regression is:

$PresVote = 0.508 + 0.024 * diff log$

4 Question 3

4.1 Answer 3a)

I run a linear regression using the lm function where the outcome variable is vote share and the explanatory variable is presvote

```
q3 <- lm(voteshare~presvote, data=incumbent)
summary(q3)
```

Our null hypothesis is that there is no association between the vote share of the presidential candidate of the incumbent party (X) and the incumbent's vote share (Y), $\beta_1 = 0$

The alternative hypothesis is that there is an association between the vote share of the presidential candidate of the incumbent party (X) and the incumbent's vote share (Y), $\beta_1 \neq 0$

Table 3: Bivariate regression between presvote(X) and voteshare(Y)

	Dependent variable:
	voteshare
Constant	0.441^{***}
	(0.008)
presvote	0.388^{***}
	(0.013)
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

As shown in Table 3, β_1 is significant at the 0.01 alpha level, hence we can reject the null hypothesis and we find support for the alternative hypothesis that there is an association between the vote share of the presidential candidate of the incumbent party (X) and the incumbent's vote share (Y), $\beta_1 \neq 0$. Moreover, we can say that for a 1 unit increase in the vote share of the presidential candidate of

the incumbent party (X), there is a 0.388 scale points increase in the incumbent's vote share (Y).
Similar to Question 1/2, the β_0 is not interpreted as it does not provide us with relevant insight in this context.

4.2 Answer 3b)

I make a scatterplot with both variables along with a regression line and the correlation value:

```
    plot(incumbent$presvote,incumbent$voteshare,
      main="Relationship between vote share
      of the presidential candidate
     and the incumbent's electoral success",
     xlab='Vote Share of presidential candidate',ylab='Vote Share')
abline(q3,col='red',lty='dashed')
cor(incumbent$voteshare,incumbent$presvote)
text(0.28, 0.92,sprintf("Correlation=%s",
round(cor(incumbent$voteshare,incumbent$presvote),4)))
```

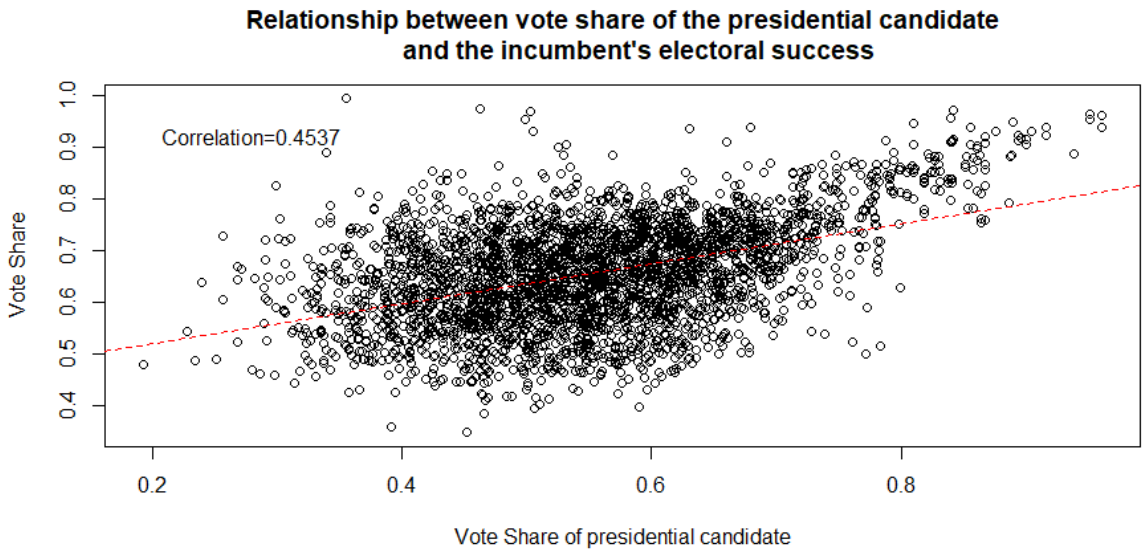

Figure 3: Relationship between vote share of the presidential candidate and the incumbent's electoral success

On the basis of the regression line and correlation value in Figure 3, we can say that there is a moderate positive correlation (0.4537) between vote share of presidential candidate and the vote share of the incumbent party.

4.3 Answer 3c)

The prediction equation for our simple linear regression is:

$VoteShare = 0.441 + 0.388 * presvote$

5 Question 4

5.1 Answer 4a)

I run a linear regression using the lm function where the outcome variable is the residuals from question 1 and the explanatory variable is the residuals from question 2

```
q4 <- lm(q1$residuals~q2$residuals)
summary(q4)
```

Our null hypothesis is that there is no association between the residuals from Q2 (X) and the Residuals from Q1 (Y), $\beta_1 = 0$

The alternative hypothesis is that there is an association between the residuals of Q2 (X) and the residuals of Q1 (Y), $\beta_1 \neq 0$

Table 4: Bivariate regression between Q2-Residuals (X) and Q1-Residuals(Y)

	Dependent variable:
	Q1-Residuals
Constant	−5.934e-18
	(0.001)
Q2-Residuals	0.257***
	(0.012)
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 shown in Table 4, β_1 is significant at the 0.01 alpha level, hence we can reject the null hypothesis and we find support for the alternative hypothesis that there is an association between the residuals from Question 2 (X) and the residuals from Question 1 (Y), $\beta_1 \neq 0$. Moreover, we can say that for a 1 unit increase in the residuals from Q2 (X), there is a 0.257 increase in the residuals from Q1 (Y).

5.2 Answer 4b)

I make a scatterplot with both variables along with a regression line and the correlation value:

```
plot(q2$residuals,q1$residuals,main='Relationship between
    Residuals from Q1 and Residuals from Q2',
     xlab='Q2-Residuals',ylab='Q1-Residuals')
abline(q4,col='red',lty='dashed')
cor(q1$residuals,q2$residuals)
text(-0.22, 0.22,sprintf("Correlation=%s", round(cor(q1$residuals,q2$residuals),4)))
```

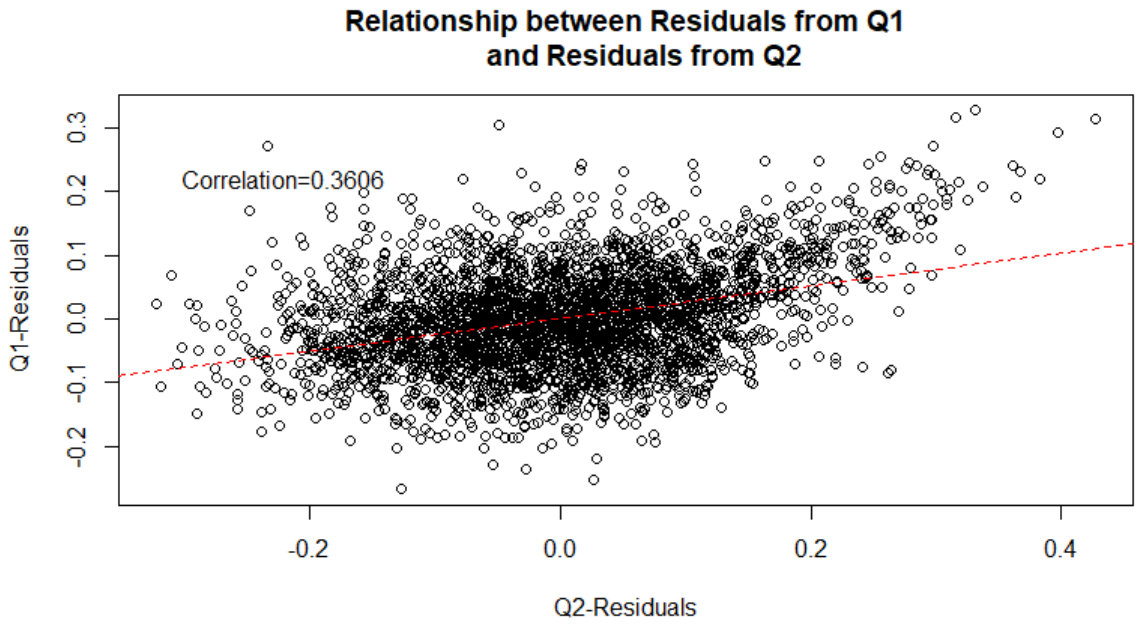

Figure 4: Relationship between Residuals from Q1 and Residuals from Q2

On the basis of the regression line and correlation value in Figure 4, we can say that there is a weak positive correlation (0.3606) between the residuals from Q1 and the residuals from Q2

5.3 Answer 4c)

The prediction equation for our simple linear regression is:

$Q1Residuals = -5.934e-18 + 0.257 * Q2Residuals$

6 Question 5

6.1 Answer 5a

I run a multiple linear regression using the lm function where the outcome variable is the incumbent's vote share and the explanatory variables are difflog and presvote

```
q5 <- lm(voteshare~difflog+presvote, data=incumbent)
summary(q5)
```

For multiple linear regression, our null hypothesis for the F-test is that there is no relationship between the explanatory variables and the outcome variable, holding other variables constant, $\beta_i = 0$

The alternative hypothesis is that there is a relationship between the explanatory variables and the outcome variable, $\beta_i \neq 0$

Table 5: Multiple linear regression model between difflog(X1),presvote(X2) and voteshare(Y)

	Dependent variable:
	voteshare
Constant	0.449***
	(0.006)
difflog	0.036***
	(0.001)
presvote	0.257***
	(0.012)
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

As we see in Table 5, the F statistic of 1,302.947 is significant at the 0.01 level and therefore, we can reject the null hypothesis and have support for the alternate hypothesis that at least one of the $\beta s \neq 0$. This makes sense as both our individual regression coefficients for difflog and presvote are significant at the 0.01 level and hence, the F-test not being significant would be unusual.

We can also interpret both the individual coefficients which are significant at the 0.01 level. First, we can say that for a 1 unit increase in difflog, this

corresponds to a 0.036 scale points increase in vote share, while controlling for the effects of presvote Next, we can say that for a 1 unit increase in presvote, this corresponds to an increase of 0.257 scale points in vote share, while controlling for the effects of difflog.

6.2 Answer 5b)

The prediction equation for a multiple linear regression can be represented as:
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2 + ... + \hat{\beta}_p * X_p$

where:

$\hat{Y}$ is the predicted value of the outcome variable based on the multiple regression equation

$X_1, X_2, .., X_p$ are the explanatory variables

$\hat{\beta}_0$ is the estimated y-intercept or the constant term. This is the value of the outcome variable when all the explanatory variables are equal to zero

$\hat{\beta}_1, \beta_2, \hat{...}, \hat{\beta}_p$ are the estimated coefficients corresponding to each explanatory variable

The error term, ϵ is inherent in the statistical model, but is not explicitly included in the prediction equation.

On the basis of this, the prediction equation for our multiple linear regression model is:

$VoteShare = 0.449 + 0.036 * difflog + 0.257 * presvote$

6.3 Answer 5c)

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

Answer:

The slope/regression coefficient for pres vote (0.257) in Question 5 is identical to the regression coefficient for Q2 residuals (0.257) in Question 4. Even their standard errors are the same.

This is my explanation for this phenomenon. In Question 1, I looked at the relationship between difflog and vote share. The residuals for Question 1 represents the part of vote share that is not linearly related to difflog. Simiarly, in Question 2, I looked at the relationship between difflog and pres vote. In this case, the residuals represent the part of pres vote that is not linearly related to difflog. Therefore, the regression coefficient in Question 4 between the residuals of Question 1 and Question 2 represents the effect of Pres vote on vote share after taking out the effects of difflog from both vote share and presvote. When it comes to multiple linear regression in Question 5, the regression coefficient (0.257) is the partial regression coefficient as it represents the contribution of presvote to the outcome variable of vote share after both variables had been linearly adjusted to the other predictor variable, difflog. From this, we also learn that simple and multiple linear regression coefficients are not the same unless

the explanatory variables are uncorrelated. When dealing with observational data, explanatory variables are rarely uncorrelated. Finally, this also helps in emphasizing the importance of controlling for other explanatory variables through multiple linear regression to better understand the relationship between different variables.