Problem Set 3

Hanyu Li(Student ID:25346841)

November 11, 2025

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 Thursday November 13, 2025. 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`.
   After checking the type and range of the outcome and explanatory variable(both are quantitative variables with no missing values), we can build a liner regression model to predict their relationship:

$$\text{voteshare}_i = \theta_0 + \theta_1\,\text{difflog}_i + \varepsilon_{1i}$$

$$H_0 : \theta_1 = 0 \qquad H_a : \theta_1 \neq 0$$

```r
#Q1.1  Run a regression between voteshare and difflog
# Check data information and missed value
summary(inc.sub)
sum(is.na(inc.sub[, c("voteshare", "difflog")]))
# Build regression model
m1 <- lm(voteshare ~ difflog, data = inc.sub)
```

The regression outcome is shown as below:

```
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
```

We can briefly conclude that:
The p-value in model 1 is lower than 0.01, indicating that the null hypothesis can be rejected under a significance level of 1%.
On average, incumbents gain about 4.17% in vote share for each one-unit increase in the log spending difference.
The model's $R^2 = 0.367$ means that 36.7% of the variation in vote share is explained by differences in campaign spending.

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

```r
#Q1.2 Plot the two variables and add the regression line
# Use ggplot2
pdf("Q1_scatter_plot.pdf")
ggplot(inc.sub, aes(x = difflog, y = voteshare)) +
  geom_point(size = 0.8, alpha = 0.8, color = "skyblue") +
  geom_smooth(method = "lm", se = FALSE , linewidth = 0.5 ) +
  labs(x = "spending difference between incumbent and challenger", y = "incumbent's vote share") +
  theme_bw()
```

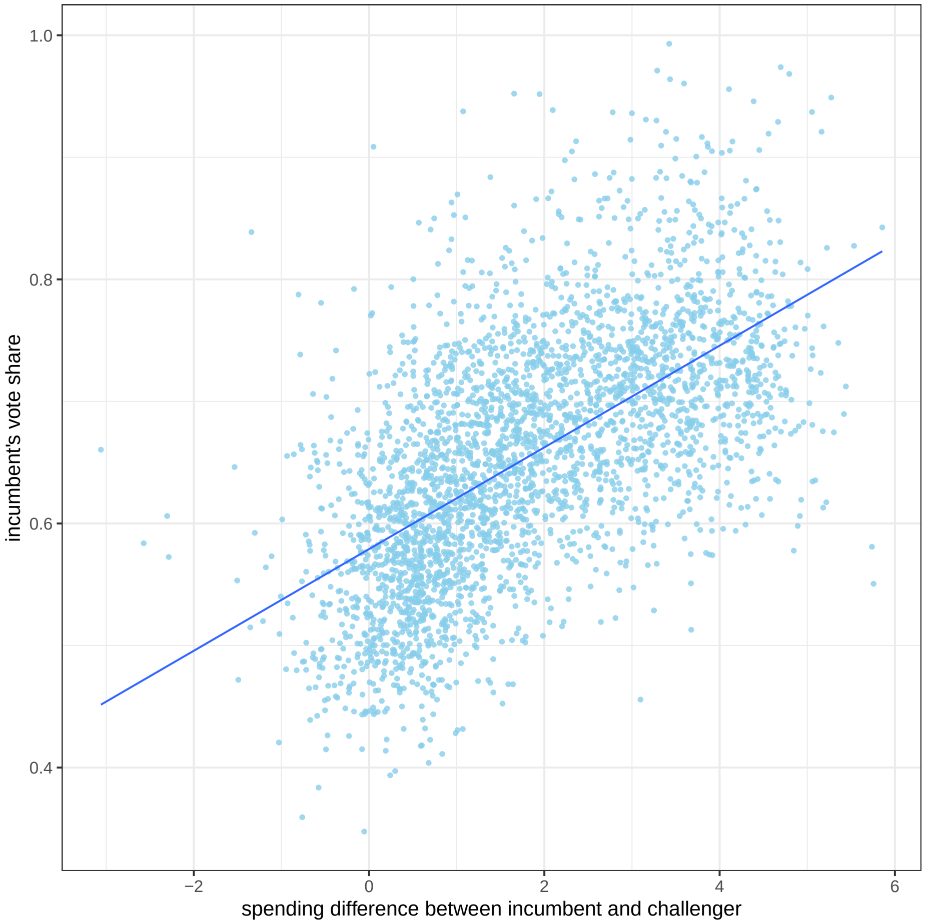


Figure 1: Relationship between Vote Share and Log Spending Difference

3. Save the residuals of the model in a separate object.
   Assign residuals to the object *res1*,it includes the distance from the predicted incumbents' vote share to the observed ones for each data point.

```

#Q1.3 Save the residuals of the model in a separate object
res1 <- resid(m1)
```

4. Write the prediction equation.

$$\widehat{\text{voteshare}} = 0.5790 + 0.0417 \times \text{difflog}$$

In this equation, the coefficient means:on average, one-unit increase in the log spending difference is associated with a 4.17% rise in the incumbent's vote share.

The intercept represents the predicted vote share of 57.9% for an incumbent when the incumbent and the challenger both spend equally.

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`.
   Follow the same step in Q1,we assume that:

$$\text{presvote}_i = \gamma_0 + \gamma_1 \,\text{difflog}_i + \varepsilon_{2i}$$

$$H_0 : \gamma_1 = 0 \qquad H_a : \gamma_1 \neq 0$$

```

#Q2.1 Run a regression between presvote and difflog
sum(is.na(inc.sub[, c("presvote")]))
m2 <- lm(presvote ~ difflog, data = inc.sub)
```

The regression outcome is shown as below:

```
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
```

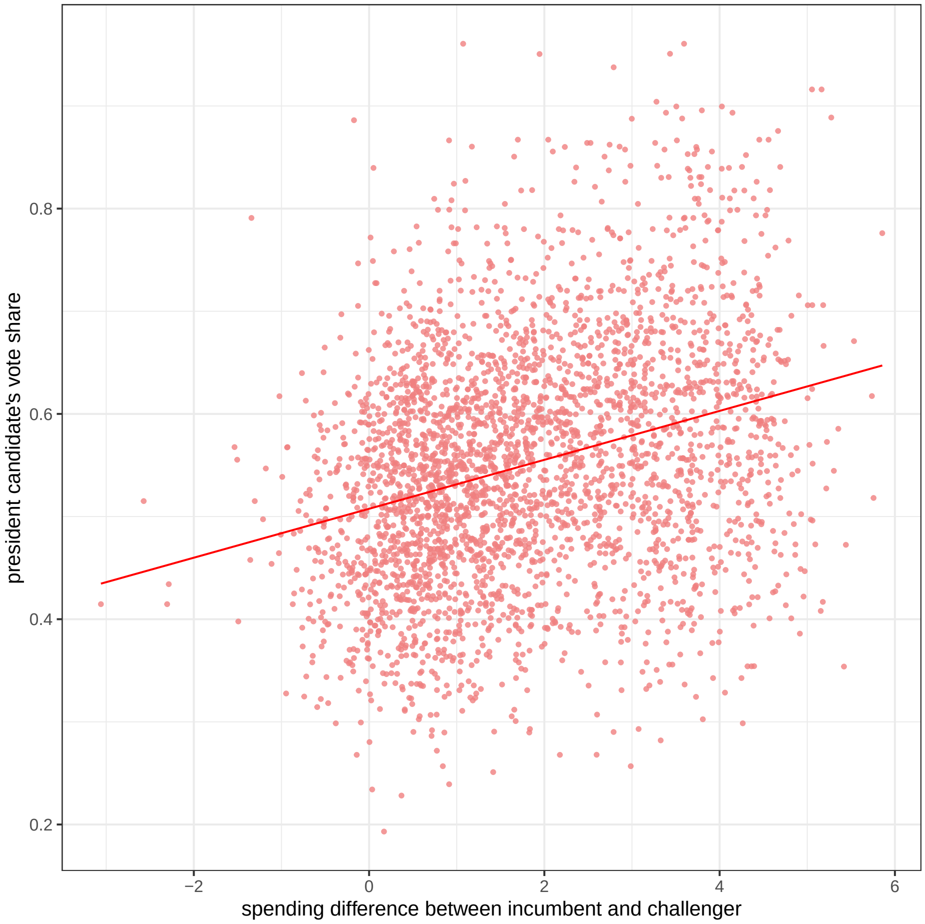
We can also conclude that:
Model 2's p-value is also lower than 0.01, so the null hypothesis $H_0 : \gamma_1 = 0$ can be rejected under a significance level of 1%.
On average, vote share of the president candidate from the incumbent raise by 2.38% for each one-unit increase in the log spending difference.
$R^2 = 0.08795$ in model 2 means that only around 8.8% of the variation in outcome variable is explained by differences in campaign spending, apparently smaller than model 1's.

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

```r

#Q2.2 Plot the two variables and add the regression line
pdf("Q2_scatter_plot.pdf")
ggplot(inc.sub, aes(x = difflog , y = presvote)) +
  geom_point(size = 0.8, alpha = 0.8, color = "lightcoral") +
  geom_smooth(method = "lm", se = FALSE ,linewidth = 0.5 , color = "red") +
  labs(x = "spending difference between incumbent and challenger", y = "president candidate's vote share") +
  theme_bw()
```

Figure 2: Relationship between President Candidate's Vote Share and Log Spending Difference



3. Save the residuals of the model in a separate object.
   Assign residuals in model 2 to the object *res2*,it shows the distance from the predicted president candidate's vote share to the observed values for each data point.

```r

#Q2.3 Save the residuals of the model in a separate object
res2 <- resid(m2)
```

4. Write the prediction equation.

$$\widehat{\text{presvote}} = 0.5076 + 0.0238 \times \text{difflog}$$

In this equation, the coefficient means that, on average, a one-unit increase in the log of spending difference is associated with a 2.38% increase in the presidential candidate's vote share.
The intercept represents the predicted vote share of 50.76% for the president candidate when the incumbent and the challenger both spend equally on campaign.

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`.

   Same as the above,we assume that:

   $$\text{voteshare}_i = \delta_0 + \delta_1 \, \text{presvote}_i + \varepsilon_{3i}$$

   $$H_0 : \delta_1 = 0 \qquad H_a : \delta_1 \neq 0$$

   ```

   #Q3.1 Run aregression where the outcome variable is voteshare and the explanatory variable is presvote.
   m3 <- lm(voteshare ~ presvote, data = inc.sub)
   ```

   The regression outcome is shown as below:

   ```
   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
   ```

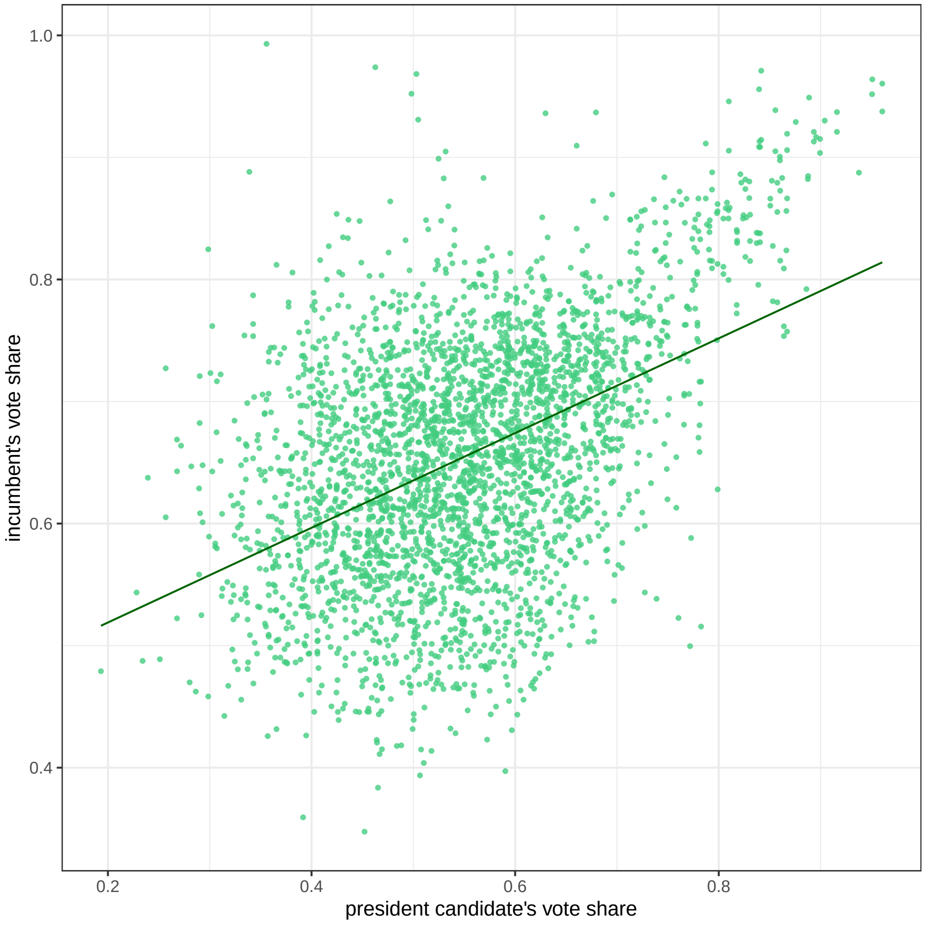
   In model 3 we can infer that:
   Model 3's p-value is lower than 0.01, so the null hypothesis $H_0 : \delta_1 = 0$ can be rejected under a significance level of 1%.
   On average, the incumbent's vote share raise by 38.80% for each one-unit increase in vote share of the presidential candidate.
   $R^2 = 0.2058$ in model 3 means that 20.58% of the variation in the incumbent's vote share is explained by variation in that of the presidential candidate.

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

```

#Q3.2 Plot the two variables and add the regression line
pdf("Q3_scatter_plot.pdf")
ggplot(inc.sub, aes(x = presvote, y = voteshare)) +
  geom_point(size = 0.8, alpha = 0.8, color = "seagreen3") +
  geom_smooth(method = "lm", se = FALSE ,linewidth = 0.5 , color = "darkgreen") +
  labs(x = "president candidate's vote share", y = "incumbent's vote share") +
  theme_bw()
```

Figure 3: Relationship between Incumbent's and President Candidate's Vote Share



3. Write the prediction equation.

$$\widehat{\text{voteshare}} = 0.4413 + 0.3880 \times \text{presvote}$$

In this equation, the coefficient means:on average, one-unit increase in the vote share of president candidate is related to a 38.80% rise in that of incumbent party.
The intercept represents a predicted vote share of 44.13% for the incumbent party when the president candidate's vote share equals zero.

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.
   we assume that:

$$\text{res1}_i = \mu_0 + \mu_1 \, \text{res2}_i + \varepsilon_{4i}$$

$$H_0 : \mu_1 = 0 \qquad H_a : \mu_1 \neq 0$$

```

#Q4.1 Run a regression between res1 and res2
df <- data.frame('m1_residuals'= res1, 'm2_residuals' = res2)
m_control <- lm(res1 ~ res2 , data = df)
summary(m_control)
# show real number of res2's coefficient(0.2569) and intercept
    (-0.000000000000000005934)
```

The regression outcome is shown as below:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.934e-18  1.299e-03    0.00        1
res2         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
```

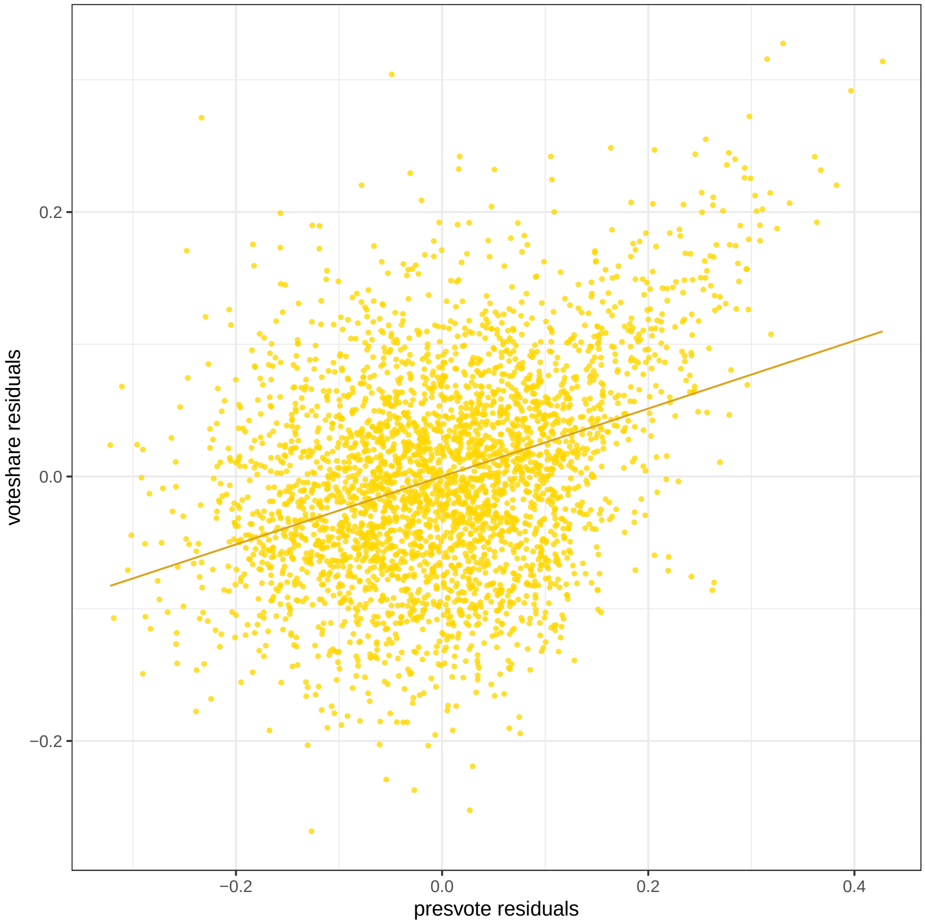
In this regression, both res1 and res2 represent the unexplained variation of the incumbent's and the president's vote shares after controlling spending differences. The coefficient on res2 (0.257,$p < 0.001$) indicates that, holding spending difference constant, we can reject null hypothesis $H_0 : \mu_1 = 0 = 0$ under the 1% significance level and get the unconfounded effect of the president's residual vote share on the incumbent's vote share.

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

```r

#Q4.1 Plot the regression between res1 and res2
pdf("Q4_scatter_plot.pdf")
ggplot(df, aes(x = res2, y = res1)) +
  geom_point(size = 0.8, alpha = 0.8, color = "gold") +
  geom_smooth(method = "lm", se = FALSE ,linewidth = 0.5 , color = "goldenrod") +
  labs(x = "presvote residuals", y = "voteshare residuals") +
```

Figure 4: Relationship between Residuals of Incumbent and Presidential Vote Shares



3. Write the prediction equation.

$$\widehat{\text{res1}} = -5.934e - 18 + 0.2569 \times \text{res2}$$

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`.
   For the multivariate regression, we assume that:

$$\text{voteshare}_i = \alpha_0 + \beta_1 \,\text{difflog}_i + \beta_2 \,\text{presvote}_i + \varepsilon_{5i}$$

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_a : \text{At least one of } \beta_1, \beta_2 \text{ is not equal to zero.}$$

```

#Q5.1 Build a multivariate regression model
m5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
```

The regression outcome is shown as below:

```
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
```

Table 1: Multivariate Regression of Incumbent Vote Share

	Dependent variable:
	Incumbent Vote Share
Log Spending Difference	0.0355***
	(0.0009)
President Candidate's Vote Share	0.2569***
	(0.0118)
Constant	0.4486***
	(0.0063)
Observations	3,193
R^2	0.4496
Adjusted R^2	0.4493
Residual Std. Error	0.0734 (df = 3190)
F Statistic	1,302.9470*** (df = 2; 3190)
Note:	*p<0.1; **p<0.05; ***p<0.01

From the regression results, the overall F-statistic ($F = 1303$, $p < 0.001$) shows that the two predictors jointly explain the variation in the incumbent's vote share. Therefore, we have evidence to reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ at the 1% significance level.

Each independent variable also contributes to explaining the dependent variable(both their coefficients' $p < 0.001$):
Specifically, $\beta_1 = 0.0355$ indicates that, holding the presidential vote share constant, on average, a one-unit increase in the log spending difference is associated with a 3.55% increase in the incumbent's vote share, while $\beta_2 = 0.2569$ suggests that,holding the log spending difference constant, a one-unit increase in the president candidate's vote share corresponds to a 25.69% rise in the incumbent's vote share on average.

The intercept of 0.4486 represents the predicted vote share of 44.9% when both explanatory variables are equal to zero.

Compared with Model 3 (which includes only *presvote*, $R^2 = 0.2058$), Model 5's $R^2 = 0.4496$ indicates that about 45% of the variation in the incumbent's vote share is jointly explained by differences in campaign spending and president candidate's vote share.

2. Write the prediction equation.

$$\widehat{\text{voteshare}} = 0.4486 + 0.0355\,\text{difflog} + 0.2569\,\text{presvote}$$

3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case?
By comparing the prediction equations of Models 4 and 5, we find that the coefficient of *res2* (the part of *presvote* unexplained by campaign spending differences) equals the coefficient of *presvote* in Model 5.

$$\text{Model 4:}\quad \widehat{\text{res1}}_i = -5.93 \times 10^{-18} + 0.2569\,\text{res2}_i$$

$$\text{Model 5:}\quad \widehat{\text{voteshare}}_i = 0.4486 + 0.0355\,\text{difflog}_i + 0.2569\,\text{presvote}_i$$

This is because, as shown in the models above, campaign spending difference affects both the presidential vote share and the incumbent's vote share, and therefore functions as a confounding variable in the relationship between them.

Model 4 uses the residuals of the incumbent's vote share and the presidential vote share that cannot be explained by campaign spending difference, effectively examining the independent effect of the presidential vote share on the incumbent's vote share after removing the influence of the confounder.

Model 5 uses a multiple regression that holds campaign spending difference constant when analyzing the effect of the presidential vote share on the incumbent's vote share.

Both models, in essence, isolate the independent effect of the presidential vote share on the incumbent's vote share by controlling for the confounding effect of campaign spending difference.