Problem Set 3

Applied Stats/Quant Methods 1

Due: November 13, 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`.

```
# read in data
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_2025/main/datasets/incumbents_subset.csv")

View(inc.sub)

#I first create a model which includes both variables and then summarise it.

```

```r
model_q1 <- lm(voteshare ~ difflog, data = inc.sub)
summary(model_q1)
```

Table 1:

	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

The summary of this regression shows that the difference in campaign spending between the incumbent and the challenger (difflog) has, on average, a positive effect of approximately 0.042 on the incumbent's vote share (voteshare). This relationship is statistically significant.

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

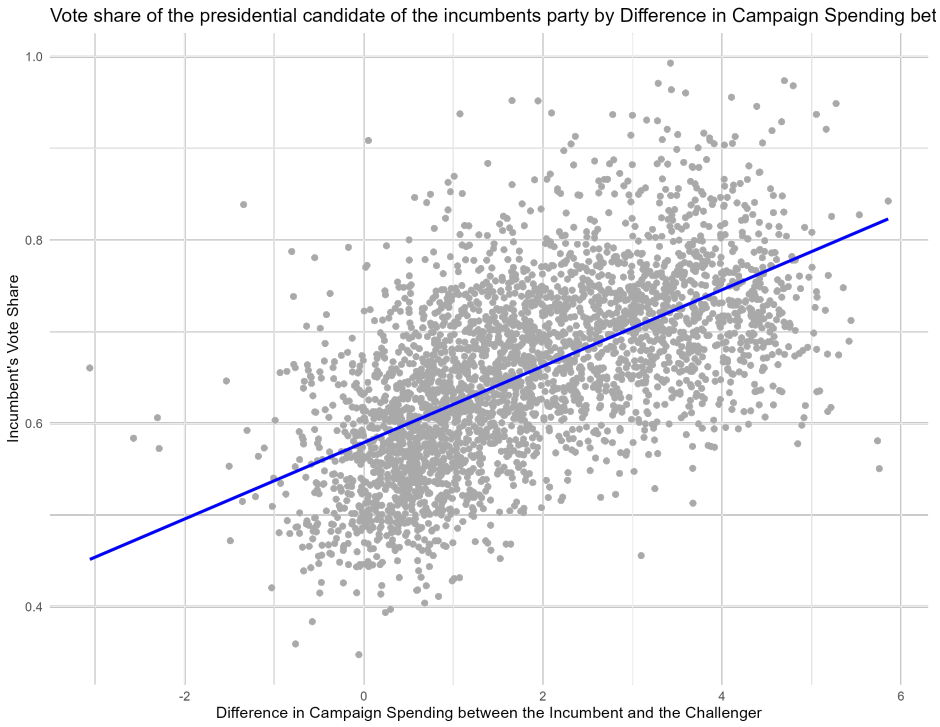
```r
#To create this scatterplot of the two variables, I will load the system
#"ggplot2", which is know for having good visuals.

library(ggplot2)
library(stargazer)

#Then, I will create the scatterplot using, again, "difflog" as the
    explanatory
#variable, and "voteshare" as the outcome variable.

ggplot1 <- ggplot(inc.sub, aes(x = difflog, y = voteshare)) +
  geom_point(colour = "darkgrey") +
  geom_smooth(method = "lm", colour = "blue", se = FALSE) +
```

```
  labs(
    title = "Vote share of the presidential candidate of the incumbents party by Difference in Campaign Spending between the Incumbent and the Challenge",
    x = "Difference in Campaign Spending between the Incumbent and the Challenger",
    y = "Incumbent's Vote Share"
  ) +
  theme_minimal()
```



This scatterplot shows the positive correlation that we had identified with the regression analysis.

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

```
#I will create a new object using the function "residuals()" and I will name it
#"residuals_model_q1)

residuals_model_q1 <- residuals(model_q1)
mean(residuals_model_q1)
residuals_model_q1
```

4. Write the prediction equation.

   The general formula is Y = a + bX.
   In this case, it is:
   Incumbent vote share = Intercept + slope * difflog
   According to the regression's results:
   Y = 0.579031 + 0.041666 * X
   voteshare = 0.579031 + 0.041666 * 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`.

```
model_q2 <- lm(presvote ~ difflog, data = inc.sub)
summary(model_q2)
```

Table 2:

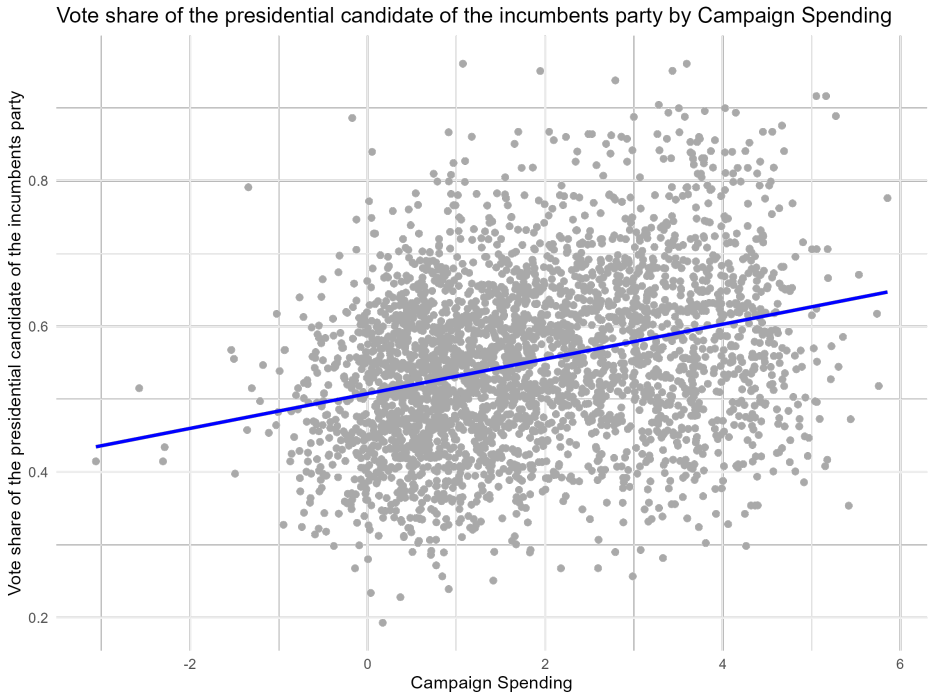
	Dependent variable:
	presvote
difflog	0.024***
	(0.001)
Constant	0.508***
	(0.003)
Observations	3,193
R^2	0.088
Adjusted R^2	0.088
Residual Std. Error	0.110 (df = 3191)
F Statistic	307.715*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

The summary of this regression shows that the difference in campaign spending between the incumbent and the challenger (difflog) has, on average, a positive effect of approximately 0.024 on the vote share of the presidential candidate of the incumbent's party (presvote). This relationship is statistically significant.

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

```r
stargazer(model_q2)

ggplot2 <- ggplot(inc.sub, aes(x = difflog, y = presvote)) +
  geom_point(colour = "darkgrey") +
  geom_smooth(method = "lm", colour = "blue", se = FALSE) +
  labs(
    title = "Vote share of the presidential candidate of the incumbents party by Campaign Spending",
    x = "Campaign Spending",
    y = "Vote share of the presidential candidate of the incumbents party"
  ) +
  theme_minimal()
```



This scatterplot shows the positive correlation that we had identified with the regression analysis.

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

```r
residuals_model_q2 <- residuals(model_q2)
mean(residuals_model_q2)
residuals_model_q2
```

4. Write the prediction equation.

   The general formula is Y = a + bX.
   In this case, it is:
   Incumbent vote share = Intercept + slope * Difference in spending
   According to the regression's results:
   Y = 0.507583 + 0.023837 * X
   presvote = 0.507583 + 0.023837 * difflog

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`.

```
model_q3 <- lm(voteshare ~ presvote, data = inc.sub)
summary(model_q3)
```

Table 3:

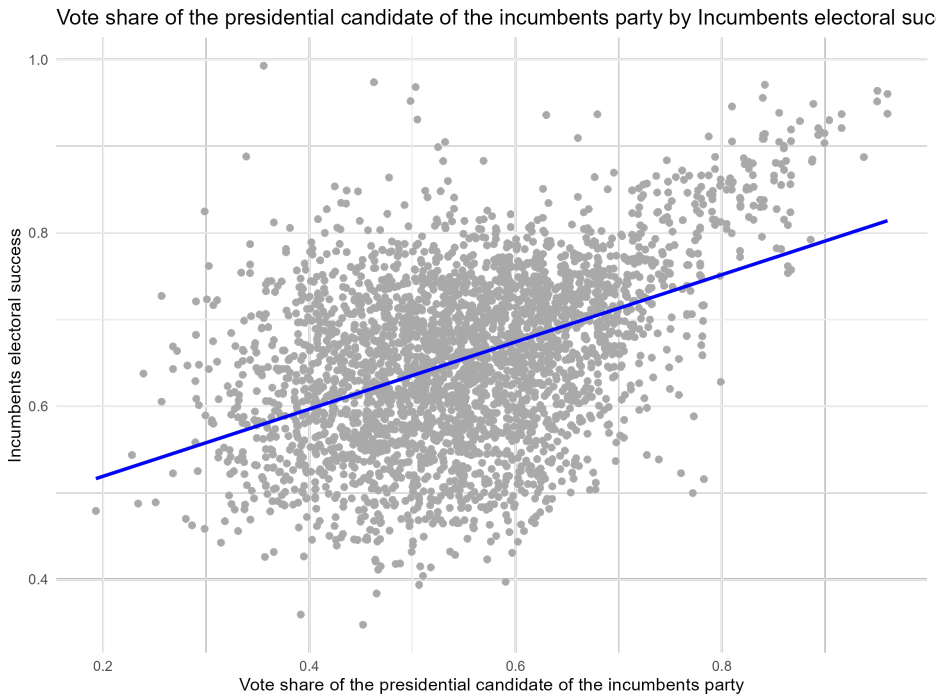
	Dependent variable:
	voteshare
presvote	0.388***
	(0.013)
Constant	0.441***
	(0.008)
Observations	3,193
R^2	0.206
Adjusted R^2	0.206
Residual Std. Error	0.088 (df = 3191)
F Statistic	826.950*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

The summary of this regression shows that the vote share of the presidential candidate of the incumbent's part (presvote) has, on average, a positive effect of approximately 0.388 on the incumbent's electoral success (voteshare). This relationship is statistically significant.

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

```
ggplot3 <- ggplot(inc.sub, aes(x = presvote, y = voteshare)) +
  geom_point(colour = "darkgrey") +
  geom_smooth(method = "lm", colour = "blue", se = FALSE) +
  labs(
```

```
    title = "Vote share of the presidential candidate of the incumbents party by Incumbents electoral success",
    x = "Vote share of the presidential candidate of the incumbents party",
    y = "Incumbents electoral success"
  ) +
  theme_minimal()
ggsave("scatterplot3_PS03.png", plot = ggplot3, width = 8, height = 6, dpi = 300)
```



This scatterplot shows the positive correlation that we had identified with the regression analysis.

3. Write the prediction equation.

   The general formula is Y = a + bX.
   In this case, it is:
   In this case, it is Incumbents electoral success = Intercept + slope * vote share of the presidential candidate
   According to the regression's results:
   Y = 0.441330 + 0.388018 * X
   voteshare = 0.441330 + 0.388018 * 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.

```
model_residuals <- lm(residuals_model_q1 ~ residuals_model_q2, data = inc
    .sub)
summary(model_residuals)
```

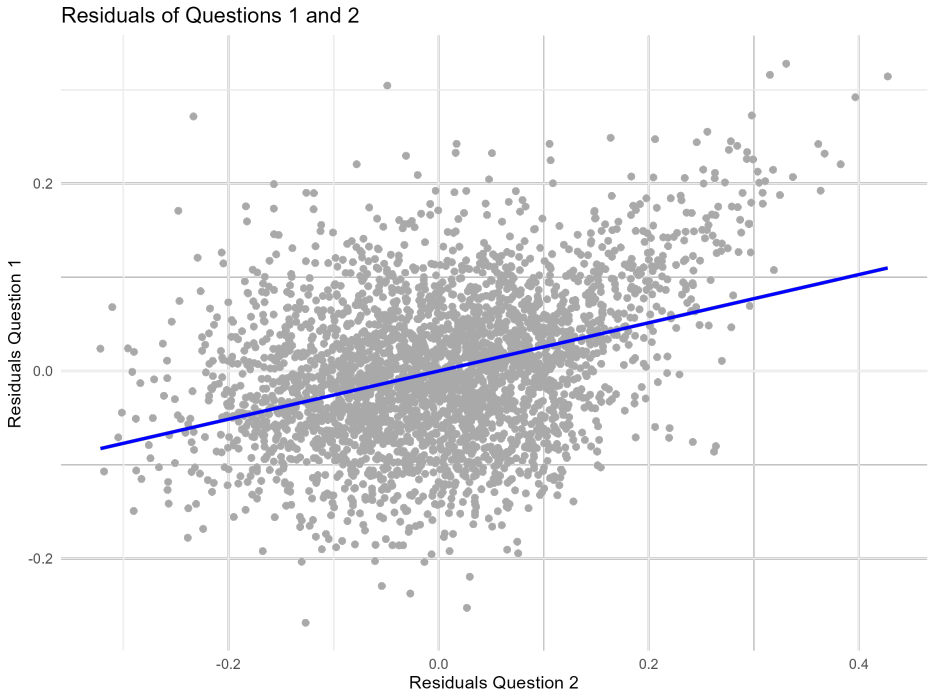
Table 4:

	Dependent variable:
	residuals_model_q1
residuals_model_q2	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

The summary of this regression shows that the residual variation in the incumbent's vote share `voteshare` that is not explained by campaign spending (`difflog`) is positively associated with the residual variation in the presidential candidate's vote share that is also (`presvote`) not explained by campaign spending. This indicates that, once the effect of `difflog` is accounted for, there remains a significant positive relationship (with an increase of 0.257 in residuals-model-q1 per one-unit increase in residuals-model-q2) between the two variables.

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

```
ggplot4 <- ggplot(inc.sub, aes(x = residuals_model_q2, y = residuals_
    model_q1)) +
  geom_point(colour = "darkgrey") +
  geom_smooth(method = "lm", colour = "blue", se = FALSE) +
  labs(
    title = "Residuals of Questions 1 and 2",
    x = "Residuals Question 2",
    y = "Residuals Question 1"
  ) +
  theme_minimal()
```



This scatterplot shows the positive correlation that we had identified with the regression analysis.

3. Write the prediction equation.

The general formula is Y = a + bX.
In this case, it is:
In this case, it is `residuals-model-q1` = Intercept + slope * `residuals-model-q2`
According to the regression's results:

Y = -5.520e-18 + 2.569e-01 * X
`residuals-model-q1` = -5.520e-18 + 2.569e-01 * `residuals-model-q2`

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?

(a) Run a regression where the outcome variable is the incumbent's `voteshare` and the explanatory variables are `difflog` and `presvote`.

```
model_q5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
summary(model_q5)
```

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

The summary of this regression shows that both the difference in campaign spending between the incumbent and the challenger (`difflog`) and the presidential candidate's vote share (`presvote`) have positive and statistically significant effects on the incumbent's vote share (`voteshare`). Holding `presvote` constant, a one-unit increase in `difflog` is associated with an average increase of 0.036 in the incumbent's vote share. At the same time, holding `difflog` constant, a one-unit increase in `presvote` corresponds to an average increase of 0.257 in `voteshare`. Importantly, this model explains about 45% of the variation in the incumbent's vote share, according to the R^2 value.

(b) Write the prediction equation.

The general formula is `Y = a + bX.`
In this case, it is:
`Incumbent's vote share = Intercept + slope1 * president's popularity`
`+ slope2 * difference in spending between incumbent and challenger`
According to the regression's results:
`Y = 0.4486442 + 0.0355431 * X1 + 0.2568770 * X2`
`voteshare = 0.4486442 + 0.0355431 * difflog + 0.2568770 * presvote`

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

What is identical in this output and in Question 4's regression output are the values for residuals-model-q2 (in Question 4's regression) and presvote (in Question 5's regression). This happens because both models are basically accounting for the same correlation, although in different fashions.

In question 5 we regress voteshare on difflog and presvote. Therefore, the coefficient on presvote shows how much this variable has an effect on voteshare once we have controlled for the effect of difflog.

Question 4's model is, however, more complex. In this case, we've regressed residuals-model-q1 on residuals-model-q2. This means that, first, we have removed the effect of difflog on both variables voteshare and presvote, as we have taken the residuals. Then, the remaining variation on voteshare has to be explained by the remaining variation on presvote.