Problem Set 3

Applied Stats/Quant Methods 1

Zhuo Zhang/23346227

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.

```
#Question1:
model1 <- lm(voteshare ~ difflog, data = inc.sub)

# Summary of the regression model
summary(model1)
```

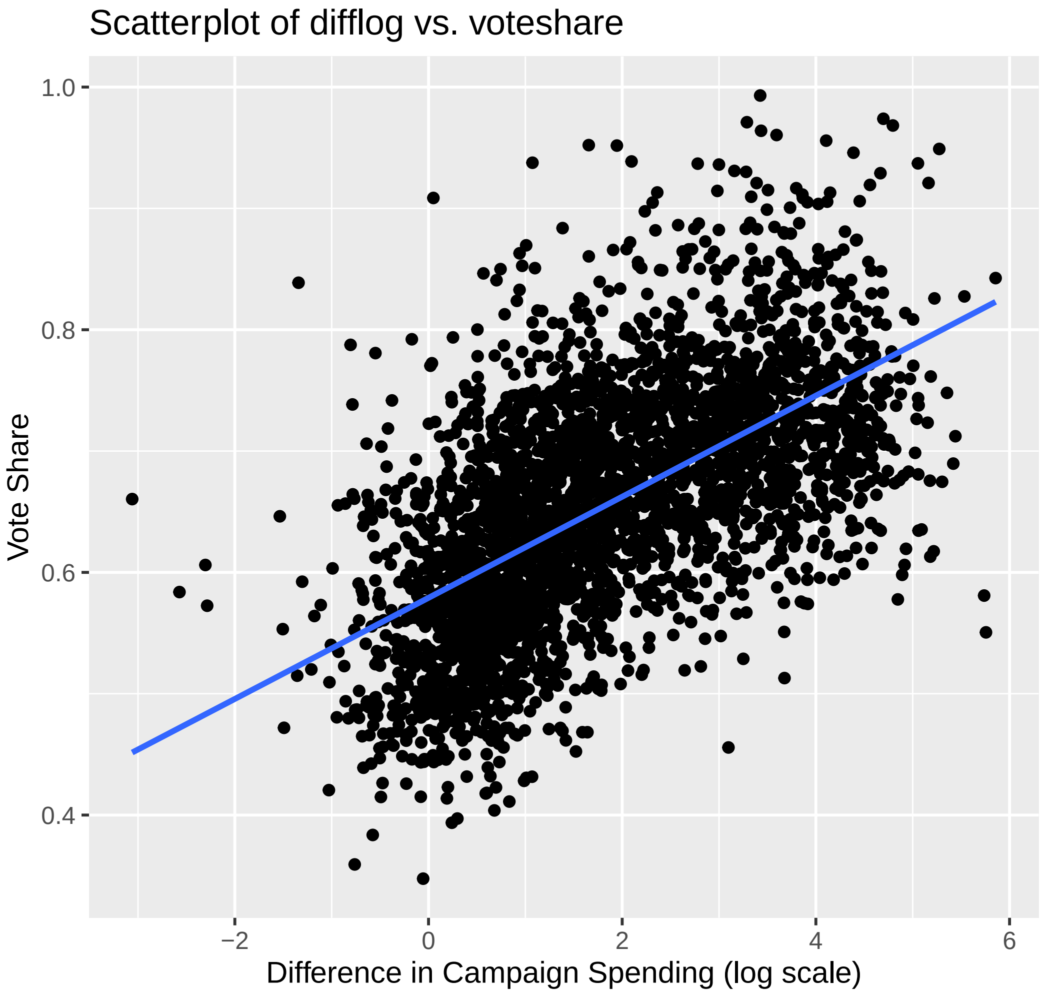
Result:

```
Call:
lm(formula = voteshare ~ difflog, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26832 -0.05345 -0.00377  0.04780  0.32749 

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
```

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

```r
# Create a scatterplot of difflog and voteshare
ggplot(inc.sub, aes(x = difflog, y = voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatterplot of difflog vs. voteshare",
       x = "Difference in Campaign Spending (log scale)",
       y = "Vote Share")
```

Result:



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

```r
# Save residuals of the model
residuals <- resid(model1)
write.csv(data.frame(residuals), ”residuals.csv”)
```

4. Write the prediction equation.

```r
# Getting the coefficients of the regression model
coefficients <- coef(model1)

# Extract the intercept and slope
intercept <- coefficients[1]
slope <- coefficients[2]

# Output prediction equation
```

```
cat("prediction equation: voteshare =", round(intercept, 2), "+", round(slope, 2), "* difflog\n")

#If the lm function is not used, it can be explained in detail with the following procedure.
# Extraction of relevant variables
x <- inc.sub$difflog
y <- inc.sub$voteshare

# Calculate the mean value
x_mean <- mean(x)
y_mean <- mean(y)

# Calculation of regression coefficients
beta_1 <- sum((x - x_mean) * (y - y_mean)) / sum((x - x_mean)^2)
beta_0 <- y_mean - beta_1 * x_mean

# Print Factor
cat("ratio (beta_1):", beta_1, "\n")
cat("intercept (beta_0):", beta_0, "\n")
cat("prediction equation: voteshare =", round(beta_0, 2), "+", round(beta_1, 2), "* difflog\n")
```

Result:
prediction equation: voteshare = 0.58 + 0.04 * difflog
ratio (beta-1): 0.04166632
intercept (beta-0): 0.5790307
prediction equation: voteshare = 0.58 + 0.04 * difflog

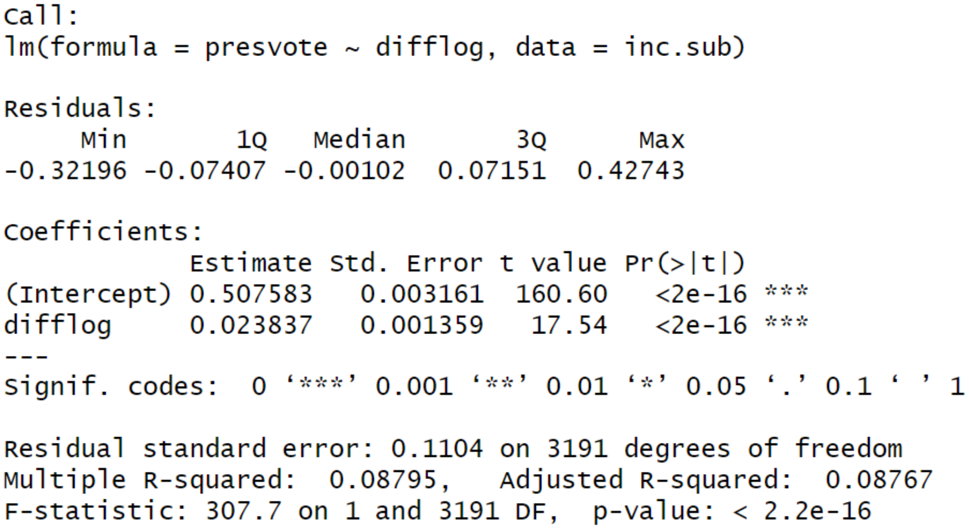
Question 2

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`.

```
# Run a regression where the outcome variable is presvote and the
  explanatory variable is difflog
model2 <- lm(presvote ~ difflog, data = inc.sub)

# Summary of the regression model
summary(model2)
```

Result:

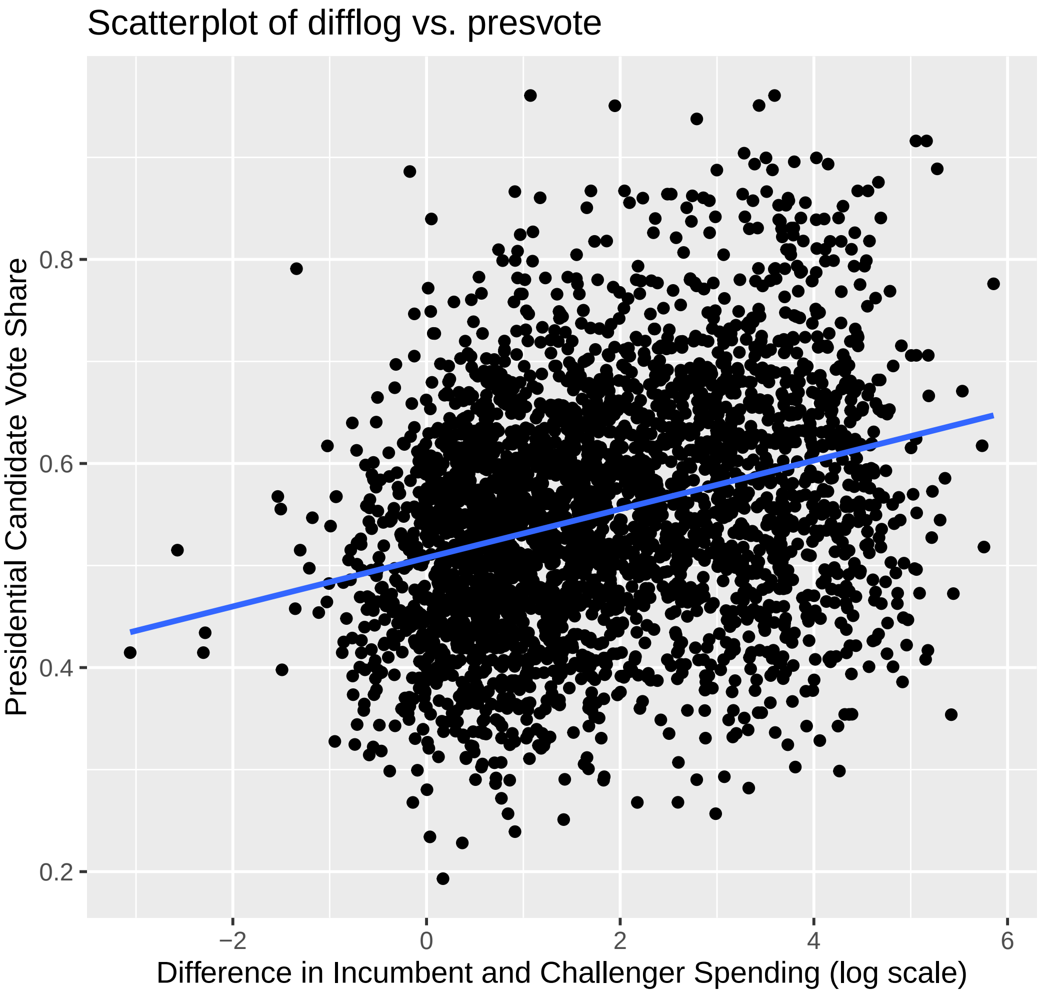


```
Call:
lm(formula = presvote ~ difflog, data = inc.sub)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32196 -0.07407 -0.00102  0.07151  0.42743 

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
```

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

```
# Create a scatterplot with a regression line
ggplot(inc.sub, aes(x = difflog, y = presvote)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatterplot of difflog vs. presvote",
       x = "Difference in Incumbent and Challenger Spending (log scale)",
       y = "Presidential Candidate Vote Share")
```

Result:



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

```
# Save residuals of the model
residuals <- resid(model2)
```

4. Write the prediction equation.

```
# Extract the intercept and slope
coefficients <- coef(model2)
intercept <- coefficients[1]
slope <- coefficients[2]

# Output prediction equation
cat("prediction equation: presvote =", round(intercept, 2), "+", round(
    slope, 2), "* difflog\n")
```

Result:
prediction equation: presvote = 0.51 + 0.02 * difflog

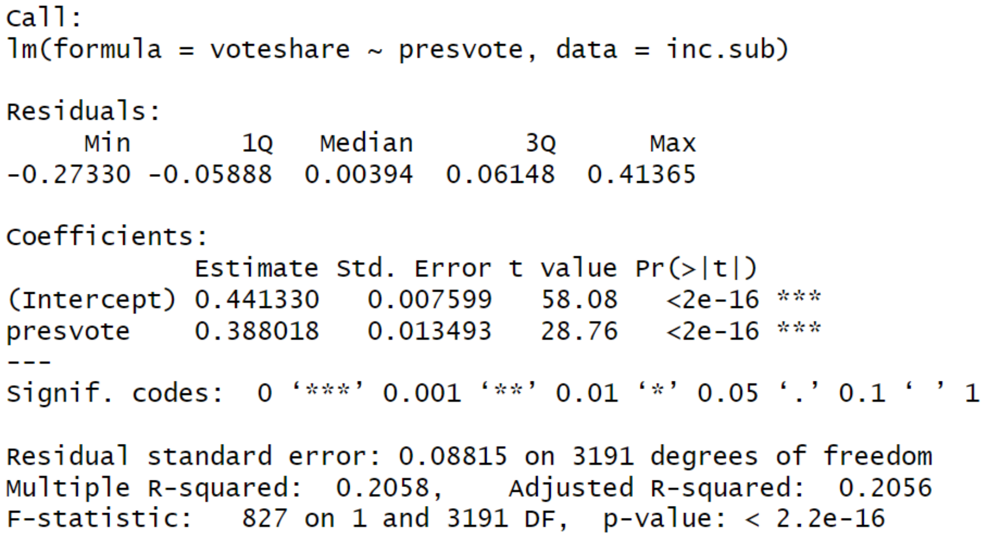
Question 3

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`.

```r
# Run a regression where the outcome variable is voteshare and the explanatory variable is presvote
model3 <- lm(voteshare ~ presvote, data = inc.sub)

# Summary of the regression model
summary(model3)
```

Result:

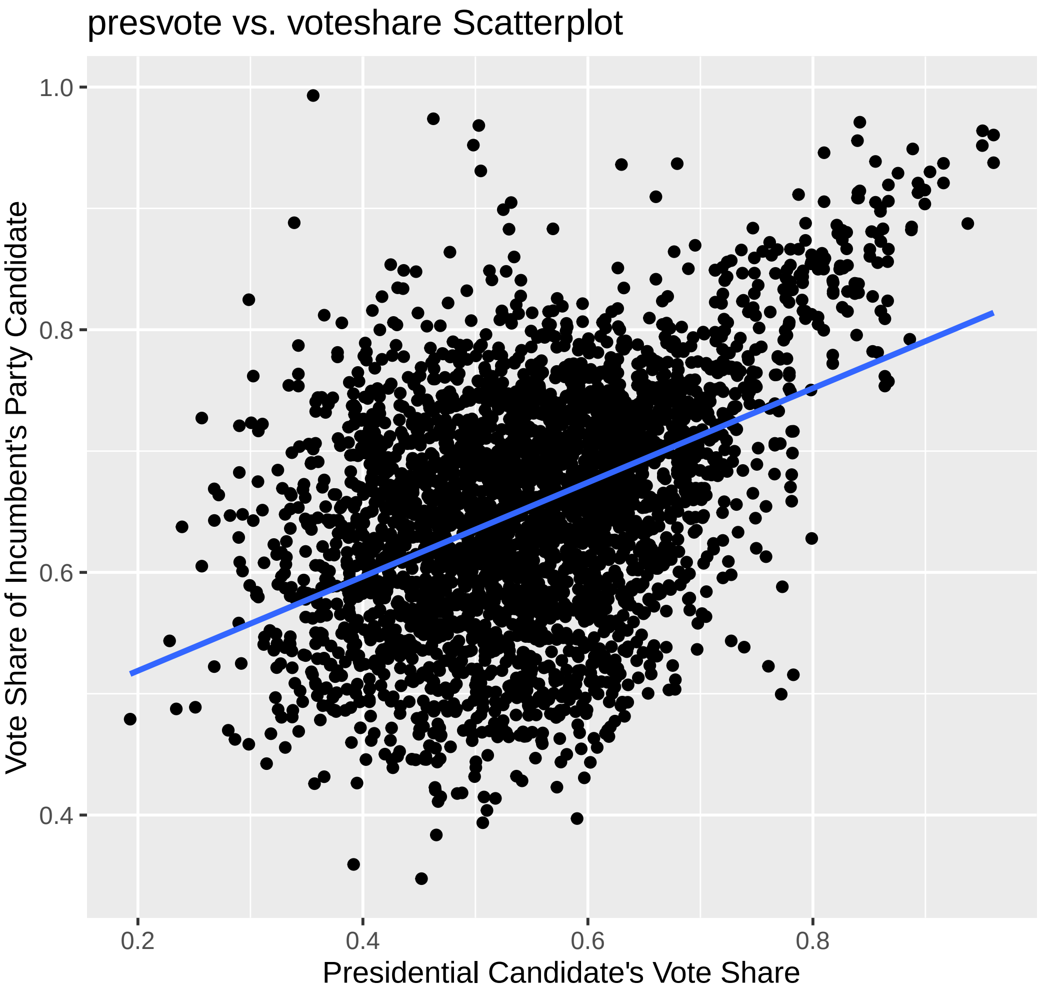


```
Call:
lm(formula = voteshare ~ presvote, data = inc.sub)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27330 -0.05888  0.00394  0.06148  0.41365 

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
```

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

```r
# Make a scatterplot of the two variables and add the regression line
ggplot(inc.sub, aes(x = presvote, y = voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "presvote vs. voteshare Scatterplot",
       x = "Presidential Candidate's Vote Share",
       y = "Vote Share of Incumbent's Party Candidate")
```

Result:



3. Write the prediction equation.

```
# Extract coefficients
coefficients <- coef(model3)
intercept <- coefficients[1]
slope <- coefficients[2]

# Write prediction equation
cat("prediction equation: voteshare =", round(intercept, 2), "+", round(
  slope, 2), "* presvote\n")
```

Result:
prediction equation: voteshare = 0.44 + 0.39 * presvote

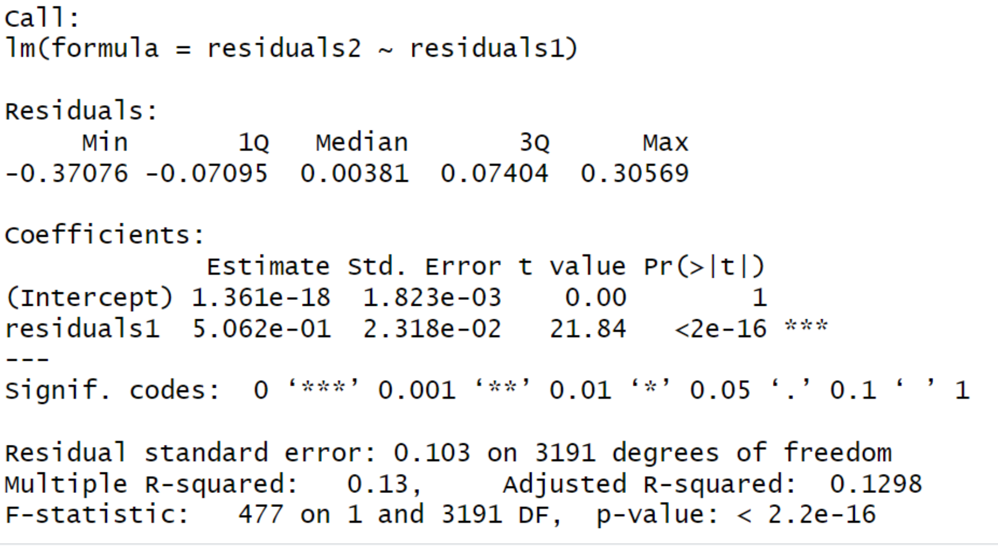
Question 4

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.

```r
# Extract residuals from the first regression
residuals1 <- resid(model1)

# Extract residuals from the second regression
residuals2 <- resid(model2)

# Run a regression where the outcome variable is the residuals from Question 1 and the explanatory variable is the residuals from Question
model4 <- lm(residuals2 ~ residuals1)

# Summary of the regression model
summary(model4)
```

Result:

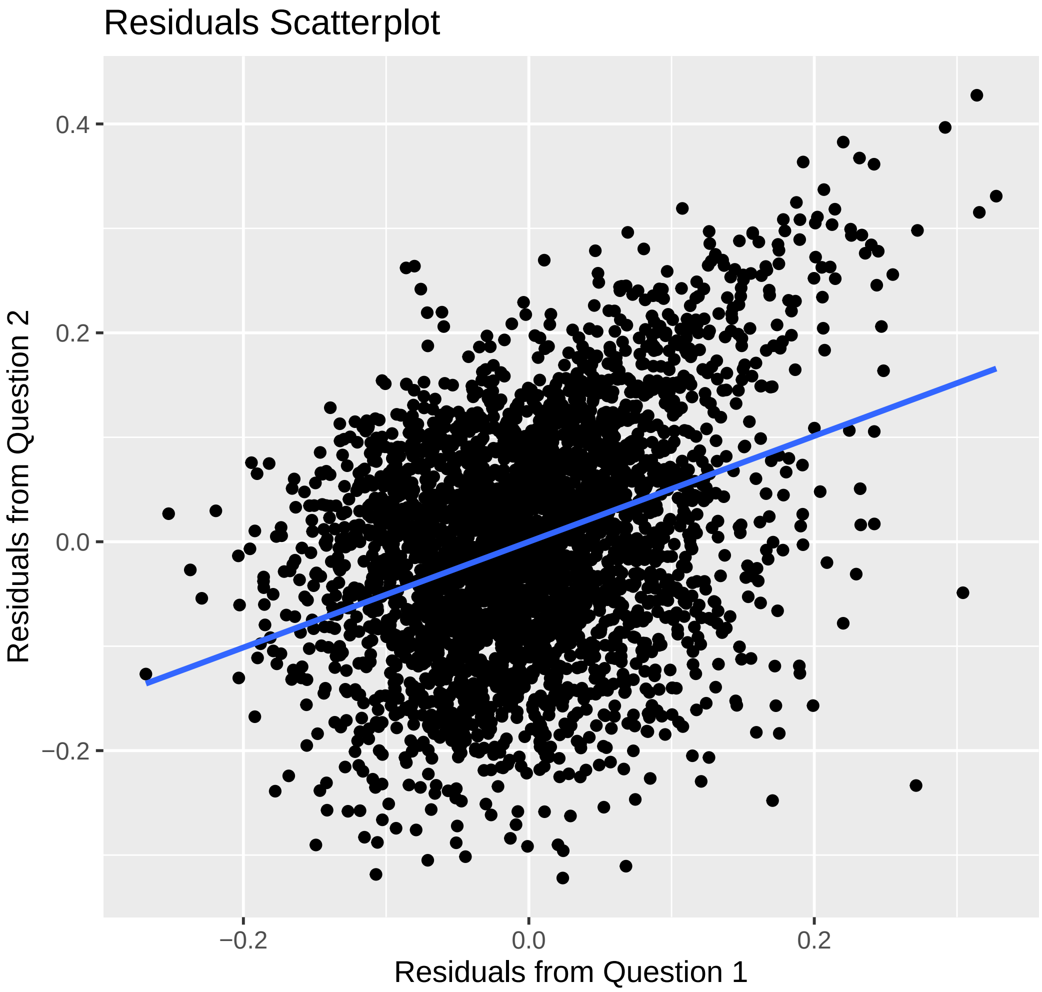


```
Call:
lm(formula = residuals2 ~ residuals1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37076 -0.07095  0.00381  0.07404  0.30569 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.361e-18  1.823e-03    0.00        1    
residuals1  5.062e-01  2.318e-02   21.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.103 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
```

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

```r
# Scatterplot with regression line for residuals
ggplot(data.frame(residuals1, residuals2), aes(x = residuals1, y = residuals2)) +
  geom_point() +
  geom_smooth(method = ”lm”, se = FALSE) +
  labs(title = "Residuals Scatterplot",
       x = "Residuals from Question 1",
       y = "Residuals from Question 2")
```

Result:



3. Write the prediction equation.

```
# Extract coefficients for the regression with residuals
coefficients_residuals <- coef(model4)
intercept_residuals <- coefficients_residuals[1]
slope_residuals <- coefficients_residuals[2]

# Write prediction equation for residuals
cat("Prediction equation for residuals: residuals2 =", round(intercept_residuals, 2), "+", round(slope_residuals, 2), "* residuals1\n")
```

Result:
Prediction equation for residuals: residuals2 = 0 + 0.51 * residuals1

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`.

```r
# Run a regression where the outcome variable is the incumbent's voteshare and the explanatory variables are difflog and presvote.
model_multiple <- lm(voteshare ~ difflog + presvote, data = inc.sub)

# Summary of the regression model
summary(model_multiple)
```

Result:

```
Call:
lm(formula = voteshare ~ difflog + presvote, data = inc.sub)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25928 -0.04737 -0.00121  0.04618  0.33126 

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
```

2. Write the prediction equation.

```r
# Extract coefficients for the multiple regression
coefficients_multiple <- coef(model_multiple)
intercept_multiple <- coefficients_multiple[1]
slope_difflog <- coefficients_multiple[2]
slope_presvote <- coefficients_multiple[3]

# Write prediction equation for multiple regression
cat("Prediction equation for multiple regression: voteshare =", round(intercept_multiple, 2), "+", round(slope_difflog, 2), "* difflog +", round(slope_presvote, 2), "* presvote\n")
```

Result:
Prediction equation for multiple regression: voteshare = 0.45 + 0.04 * difflog + 0.26 * 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?
   Result:
   Both questions 4 and 5 see the incumbent's share of the vote affected by and the difference in spending between the incumbent and the challenger. And there is a linear relationship, with more spending leading to a higher vote share. But as can also be seen in question 5, the incumbent's share of the vote is affected by both presidential popularity and the difference in spending between the incumbent and the challenger.