Problem Set 3

Applied Stats/Quant Methods 1
Jia Lyu-2337006

Due: November 11, 2024

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 11, 2024. 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`.

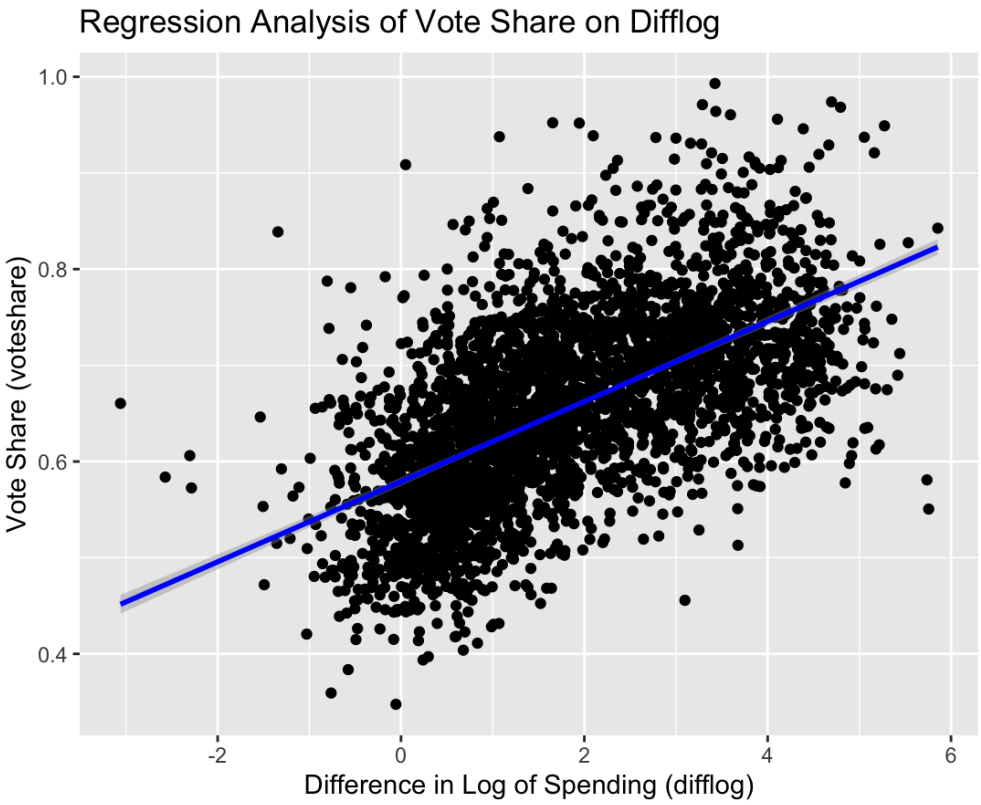
```r
# Running regression analysis
model <- lm(voteshare ~ difflog, data = df)
summary(model)
```

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

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

The regression analysis reveals that the difference in
campaign  spending (‘difflog‘) explains a moderate
portion (36.73%) of the variability in the incumbent’s vote
share   (‘voteshare‘). With a R-squared of 0.3673 and an
adjusted  R-squared of 0.3671, the model suggests that
‘difflog‘ contributes meaningfully to predicting vote share
without indicating overfitting. The residual standard error
(0.07867) indicates a relatively close fit between observed
and predicted values.
A highly significant F-statistic (1853, p < 2.2e-16) further
confirms the model’s relevance, underscoring ‘difflog‘ as
a statistically significant predictor of ‘voteshare‘, although
other factors may also play a role.
```

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

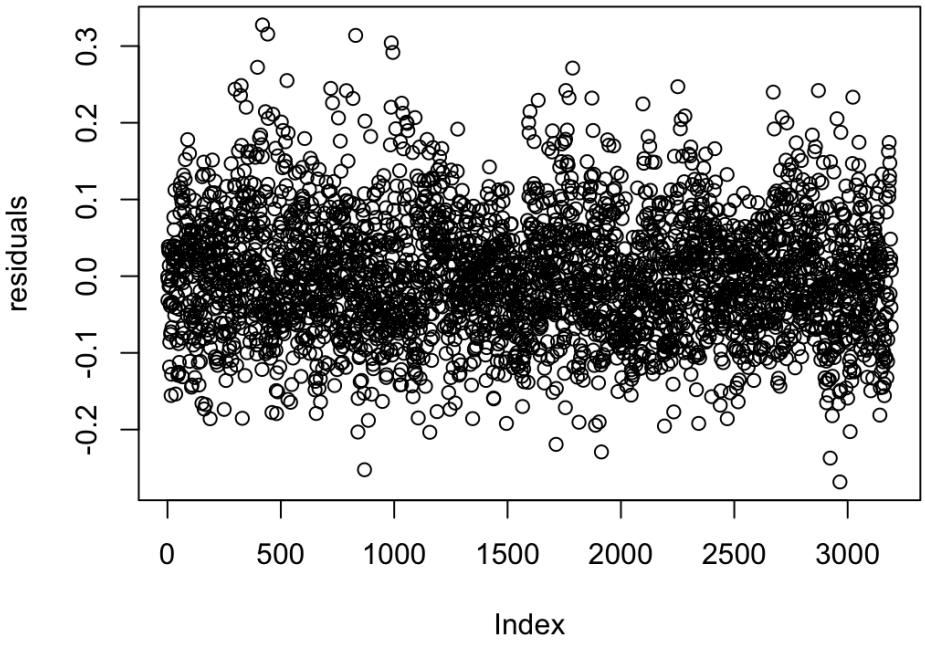
```r
# Scatterplot with regression line
ggplot(df, aes(x = difflog , y = voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Regression Analysis of Vote Share on Difflog",
       x = "Difference in Log of Spending (difflog)",
       y = "Vote Share (voteshare)")
```



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

```r
residuals <- residuals(model)
head(residuals)

plot(residuals)
```

```
          1           2           3
-0.0004227622 -0.0316840149 -0.0045514943
          4           5           6
 0.0386688767  0.0355287965  0.0322832521
```



4. Write the prediction equation.

```r
intercept <- coef(model)
slope <- coef(model)

# Prediction equation
prediction_equation <- paste("voteshare =", as.character(intercept), " +
    ", as.character(slope), "* difflog")
print(prediction_equation)
```

```
"voteshare = 0.579030710920674  +  0.0416663238227399 * 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
model_presvote <- lm(presvote ~ difflog, data = df)
summary(model_presvote)
```

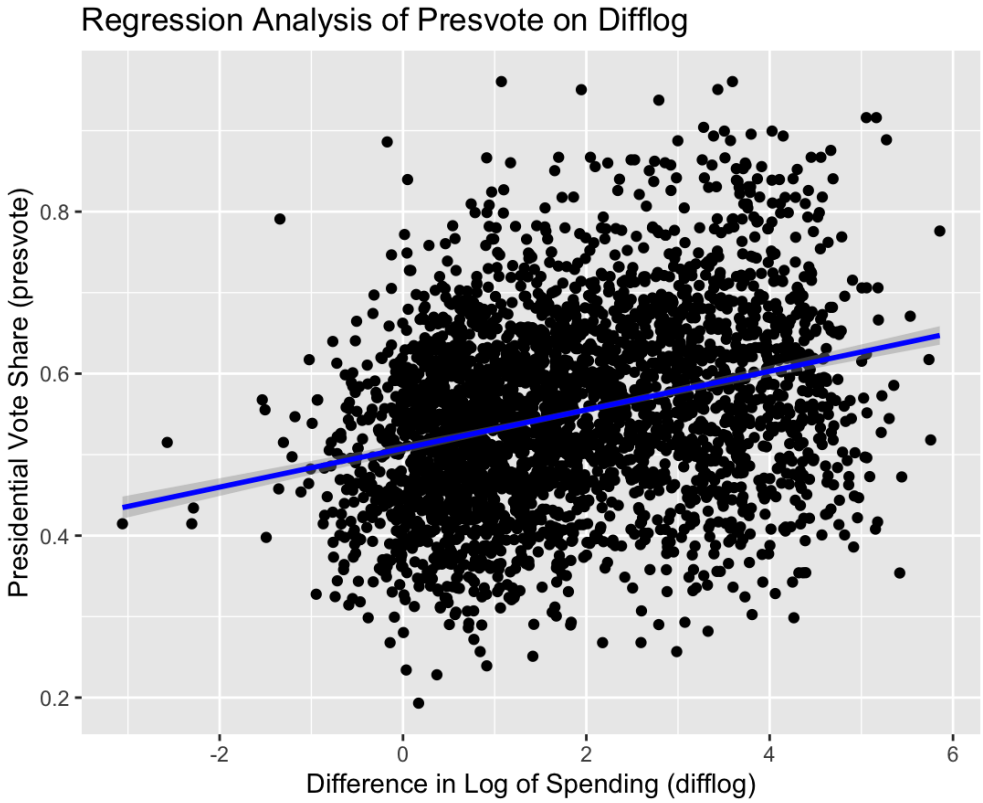
```
Call:
lm(formula = presvote ~ difflog, data = df)

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
```

The regression analysis reveals a significant positive relationship between the difference in campaign spending (‘difflog‘) and the incumbent's vote share (‘presvote‘). Specifically, for every one-unit increase in ‘difflog‘, the incumbent's vote share increases by approximately 0.023837 units. The model's intercept is 0.507583, with both the intercept and the slope being statistically significant (p-value < 2e-16). The coefficient of determination (R-squared) is 0.08795, indicating that about 8.8% of the variation in the incumbent's vote share can be explained by the difference in campaign spending. Although the model is statistically significant (F-statistic = 307.7, p-value < 2.2e-16), its explanatory power is limited, suggesting that other factors beyond campaign spending may also influence the incumbent's

vote share.

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

```r
# Create a scatter plot and add the regression line
ggplot(df, aes(x = difflog, y = presvote)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") + # Add regression line
  labs(title = "Regression Analysis of Presvote on Difflog",
       x = "Difference in Log of Spending (difflog)",
       y = "Presidential Vote Share (presvote)")
```



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

```r
# Save the residuals and create a residual plot
residuals_presvote <- residuals(model_presvote)
head(residuals_presvote)

# Create residual plot
plot(residuals_presvote, main = "Residuals of Regression Model",
     xlab = "Index", ylab = "Residuals")
abline(h = 0, col = "red")  # Add a horizontal reference line
```

```
          1           2           3           4           5           6
0.005605594 0.037578519 -0.053134788 -0.052993694 -0.045842994 0.074339701
```

4. Write the prediction equation.

```r
#Write the regression equation
intercept_presvote <- coef(model_presvote)
slope_presvote <- coef(model_presvote
prediction_equation_presvote <- paste("presvote =", round(intercept_
    presvote, 3),
                                      " + ", round(slope_presvote, 3), "*
     difflog")
print(prediction_equation_presvote)
```

```
"presvote = 0.508  +  0.024 * 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 the regression analysis
model_voteshare <- lm(voteshare ~ presvote, data = df)
summary(model_voteshare)
```

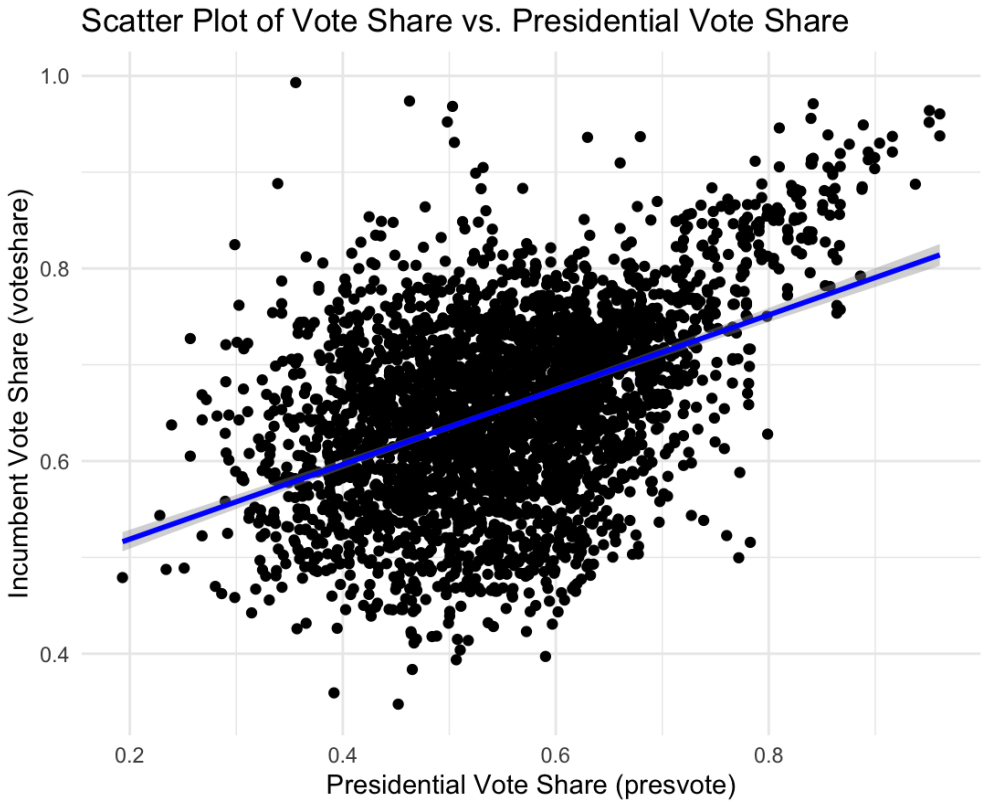
```
Call:
lm(formula = voteshare ~ presvote, data = df)

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
```

The regression analysis reveals a significant positive relationship between ‘presvote‘ and the incumbent's ‘voteshare‘. Specifically, for every one-unit increase in the presidential vote share, the incumbent's vote share increases by approximately 0.2465.
The model explains about 8.7% of the variability in the incumbent's vote share, suggesting that other factors beyond ‘presvote‘ may also influence the incumbent's electoral success.

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

```r
# Create a scatterplot with the regression line
ggplot(df, aes(x = presvote, y = voteshare)) +
```

```
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Scatter Plot of Vote Share vs. Presidential Vote Share",
       x = "Presidential Vote Share (presvote)",
       y = "Incumbent Vote Share (voteshare)")
```



3. Write the prediction equation.

```
# Write the regression equation
intercept_voteshare <- coef(model_voteshare)
slope_voteshare <- coef(model_voteshare)
prediction_equation_voteshare <- paste("voteshare =", round(intercept_voteshare, 3),
                                       " + ", round(slope_voteshare, 3), "* presvote")
print(prediction_equation_voteshare)
```

```
"voteshare = 0.441  +  0.388 * 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.

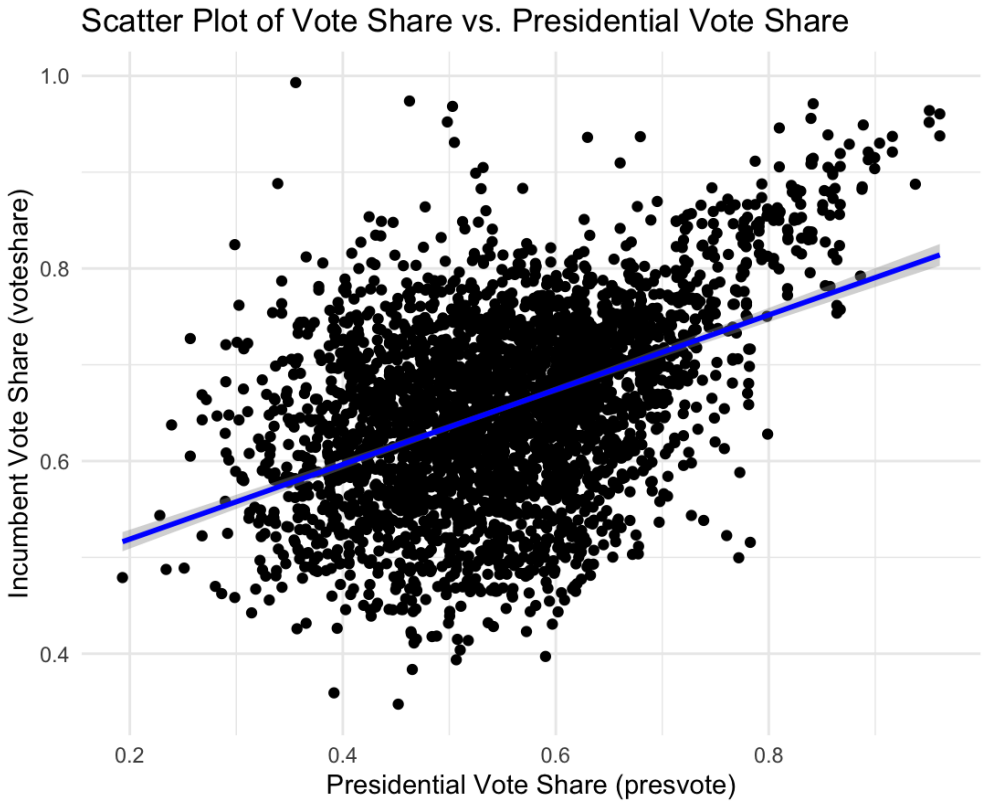
```r
# Run the regression on the residuals
model_residuals <- lm(voteshare ~ presvote, data = df)

summary(model_residuals)
```

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

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
```

```
The regression analysis shows a significant positive relationship
between the residuals of the presidential vote share and the
incumbent's vote share. Specifically, for every one-unit increase
in the residuals of the presidential vote share, the incumbent's
residuals increase by 0.2569 units. The model explains
approximately 13% of the variance in the residuals, with a
p-value less than 2.2e-16, indicating statistical significance.
However, the model's limited explanatory power suggests
that other factors may also affect the residuals.
```

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

```r
# Create a scatter plot and add the regression line
ggplot(data.frame(voteshare, presvote), aes(x = presvote, y = voteshare)) +
  geom_point() +  # Create scatter plot
  geom_smooth(method = "lm", color = "blue") +  # Add regression line
  labs(title = "Regression Analysis of Residuals from Voteshare on Residuals from Presvote",
       x = "Residuals from Presvote (presvote)",
       y = "Residuals from Voteshare (voteshare)")  # Set title and axis labels
```



3. Write the prediction equation.

```r
# Write the regression equation
intercept_residuals <- coef(model_residuals)[1]
slope_residuals <- coef(model_residuals)[2]
prediction_equation_residuals <- paste("voteshare =", round(intercept_residuals, 3),
                                       " + ", round(slope_residuals, 3), "* presvote")
print(prediction_equation_residuals)
```

```
voteshare = 0.441  +  0.388 * presvote
```

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 the regression where the outcome variable is voteshare and the
    explanatory variables are difflog and presvote
model_voteshare <- lm(voteshare ~ difflog + presvote, data = df)
summary(model_voteshare)
```

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

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
```

The regression analysis reveals a significant positive relationship between the incumbent's vote share, campaign spending difference (‘difflog‘), and the president's vote share (‘presvote‘). A one-unit increase in ‘difflog‘ results in a 0.0355 increase in the incumbent's vote share, while a one-unit increase in ‘presvote‘ results in a 0.2569 increase. The model explains 44.96% of the variation in vote share R-squared = 0.4496), with an F-statistic of 1303 and p-value < 2.2e-16, indicating statistical significance. However, other factors may still influence the incumbent's vote share.

2. Write the prediction equation.

```r
# Write the prediction equation
intercept_voteshare <- coef(model_voteshare)
slope_difflog <- coef(model_voteshare)
slope_presvote <- coef(model_voteshare)
prediction_equation_voteshare <- paste("voteshare =", round(intercept_
    voteshare, 3),
                                       " + ", round(slope_difflog, 3), "*
     difflog",
                                       " + ", round(slope_presvote, 3), "
    * presvote")
print(prediction_equation_voteshare)
```

```
voteshare = 0.449  +  0.036 * difflog  +  0.257 * 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?

Similarities:
1. Residuals:The statistical data for residuals in both models are identical: minimum value: -0.25928, first quartile: -0.04737, median: -0.00121, third quartile: 0.04618, and maximum value: 0.33126.
2. Residual Standard Error:The residual standard errors are nearly identical (0.07339 and 0.07338), indicating similar model fits.
3. F-statistic and p-value:Both models show significant F-statistics and p-values, suggesting they are statistically significant.

Reasons:
1. Same Dataset:Both models use the same dataset, resulting in similar residual distributions.
2. Similar Error Structures:The error terms in both models likely follow similar distributions.
3. Same Sample Size:The similar degrees of freedom (3190 and 3191) contribute to comparable residual standard errors.
4. Model Significance:Both models effectively explain the variability in the dependent variables, supported by significant F-statistics.
5. Linear Relationships: Both models capture linear relationships, leading to similar residual patterns.
6. Random Error: In large samples, random error can produce similar residual patterns, even with different models.