Problem Set 3

Applied Stats/Quant Methods 1

Due: November 11, 2024

Name: Ombeline Mussat
Student Number: 24346050

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`.

   Let's run a regression where `voteshare` is the outcome variable and `difflog` is the explanatory variable.

   ```r
   regression_q1 <- lm(voteshare ~ difflog, data = inc.sub)
   summary(regression_q1)
   ```

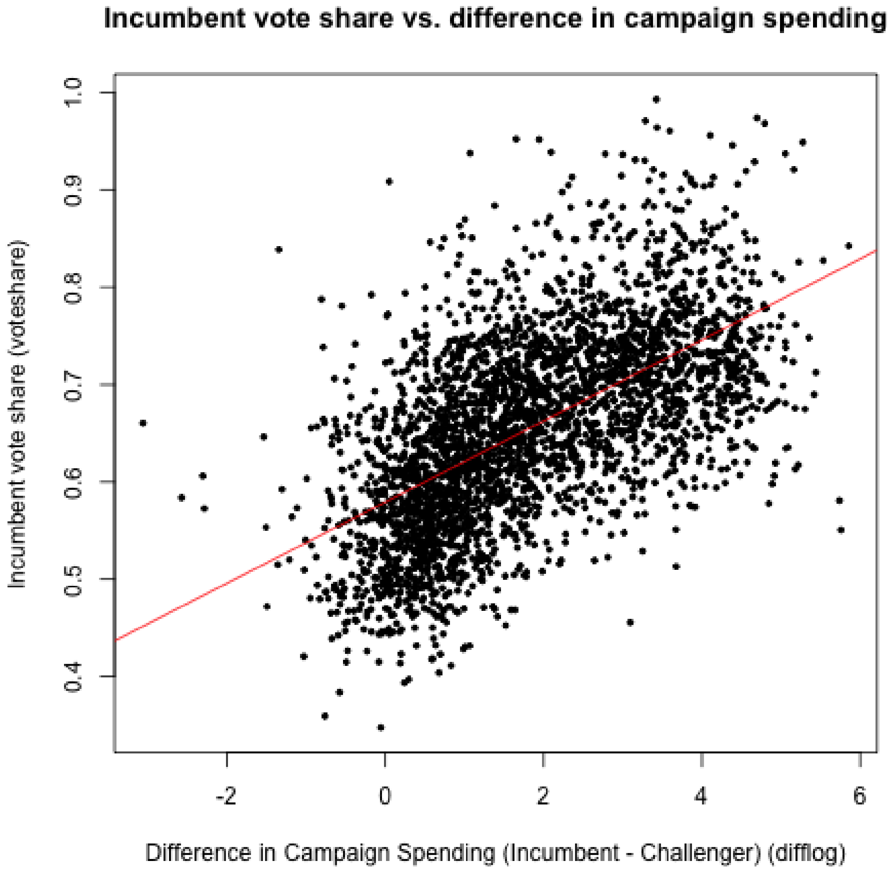
   We get the following results:

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

   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.

   Let's make a scatterplot of the two variables `voteshare` on the y-axis and `difflog` on the x-axis. We will also add the regression line.

```r
#Let's make a scatterplot of the two variables and add the regression line
png(file="scatter_plot_voteshare_difflog.png")
plot(inc.sub$difflog, inc.sub$voteshare,
     xlab = "Difference in Campaign Spending (Incumbent - Challenger) (difflog)",
     ylab = "Incumbent vote share (voteshare)",
     main = "Incumbent vote share vs. difference in campaign spending",
     pch = 19, col = "black", cex = 0.5)

# Add the regression line using abline
abline(regression_q1, col = "red")
dev.off()
```



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

   We can save the residuals of the model in a separate object which we can call `residuals_q1`.

```r
residuals_q1 <- regression_q1$residuals
```

   The object `residuals_q1` contains the differences between the actual vote share (`voteshare`) and the predicted vote share from the regression model. Each residual reflects how much the model's prediction deviates from the observed vote share, with positive values indicating underestimates and negative values indicating overestimates by the model.

4. Write the prediction equation.

   The prediction equation is:

$$\text{voteshare} = 0.579031 + 0.041666 \times \text{difflog}$$

   This equation indicates that the expected value of voteshare increases by approximately 0.042 for each one-unit increase in difflog. When difflog is 0, voteshare is equal to approximately 0.58.

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`.

   Let's run a regression where `presvote` is the outcome variable and `difflog` is the explanatory variable.

   ```
   regression_q2 <- lm(presvote ~ difflog, data = inc.sub)
   summary(regression_q2)
   ```

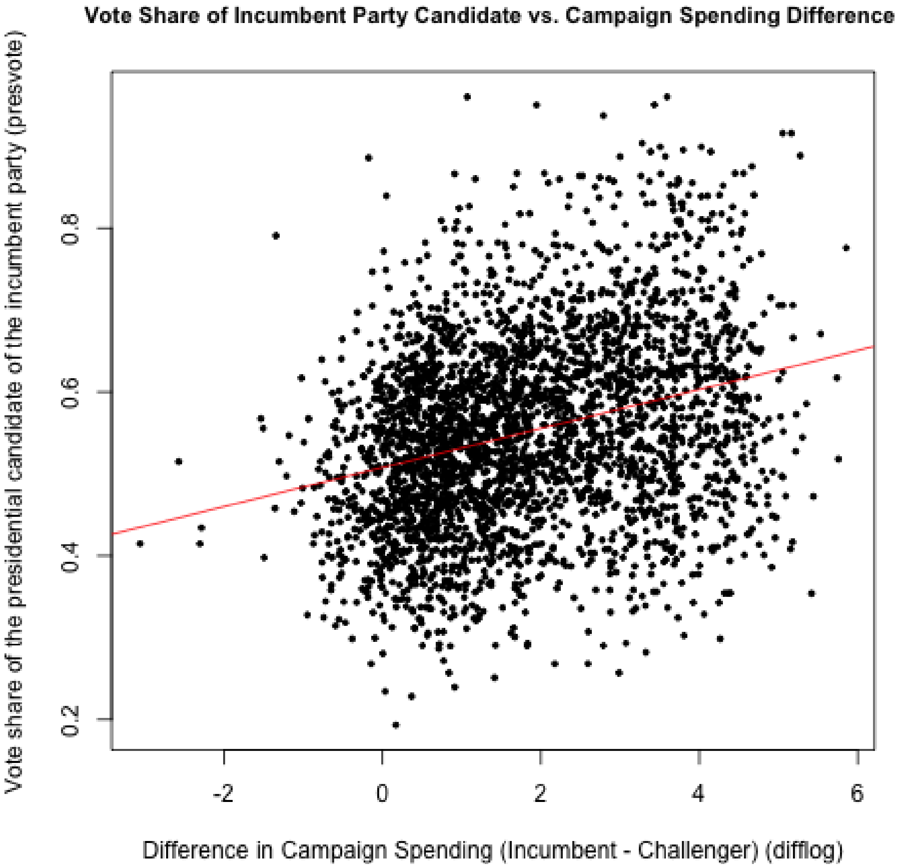
   We get the following results:

   ```
   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.

   Let's make a scatterplot of the two variables `presvote` on the y-axis and `difflog` on the x-axis. We will also add the regression line.

```
#Let's make a scatterplot of the two variables
png(file=”scatter_plot_presvote_difflog.png”)
plot(inc.sub$difflog, inc.sub$presvote,
     xlab = ”Difference in Campaign Spending (Incumbent − Challenger) (difflog)”,
     ylab = ”Vote share of the presidential candidate of the incumbent party (presvote)”,
     main = ”Vote Share of Incumbent Party Candidate vs. Campaign Spending Difference ”,
     cex.main = 0.95, #we want a smaller size for the title so it fits on the graph
     pch = 19, col = ”black”, cex = 0.5)

#Add the regression line using abline
abline(regression_q2, col = ”red”)
dev.off()
```



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

   We can save the residuals of the model in a separate object which we can call `residuals_q2`.

   ```
   residuals_q2 <- regression_q2$residuals
   ```

   The object `residuals_q2` contains the differences between the actual value of `presvote` and the predicted values of `presvote` from the regression model. Each residual reflects how much the model's prediction deviates from the observed value, with positive values indicating underestimates (predicted value is too low) and negative values indicating overestimates by the model (predicted value is too high).

4. Write the prediction equation.

   The prediction equation is:

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

   This equation shows that for each one-unit increase in difflog, presvote increases by approximately 0.0238. The intercept of 0.5076 represents the estimated value of presvote when $\text{difflog} = 0$.

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`.

   Let's run a regression where `voteshare` is the outcome variable and `presvote` is the explanatory variable.

   ```
   regression_q3 <- lm(voteshare ~ presvote, data = inc.sub)
   summary(regression_q3)
   ```

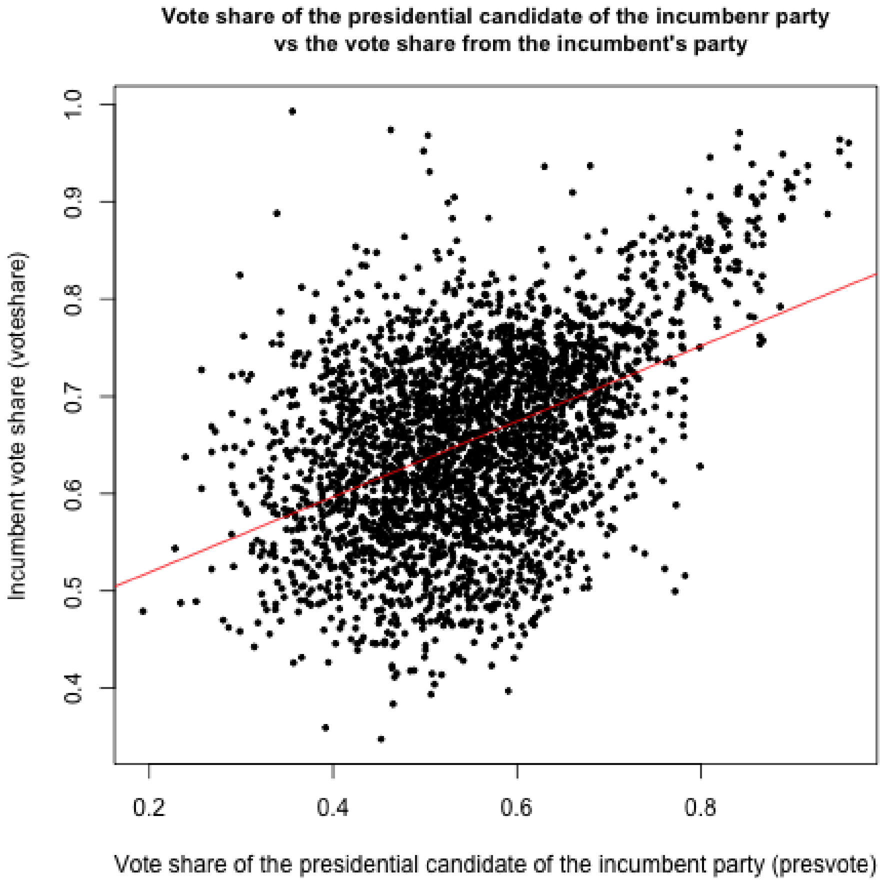
   We get the following results:

   ```
   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.

   Let's make a scatterplot of the two variables `voteshare` on the y-axis and `presvote` on the x-axis. We will also add the regression line.

```r
#Let's make a scatterplot of the two variables and add the regression line
png(file="scatter_plot_presvote_voteshare.png")
plot(inc.sub$presvote, inc.sub$voteshare,
     xlab = "Vote share of the presidential candidate of the incumbent party (presvote)",
     ylab = "Incumbent vote share (voteshare)",
     main = " Vote share of the presidential candidate of the incumbenr party
     vs the vote share from the incumbent's party",
     cex.main = 0.95,
     pch = 19, col = "black", cex = 0.5)

# Add the regression line using abline
abline(regression_q3, col = "red")
dev.off()
```



3. Write the prediction equation. The prediction equation is:

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

This equation indicates that `voteshare` increases by 0.388 for each one-unit increase in `presvote`. The intercept, 0.4413, represents the estimated `voteshare` when `presvote` is 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.

   Let's run a regression where `residuals_q1` is the outcome variable and `residuals_q2` is the explanatory variable.

   ```
   regression_q4 <- lm(residuals_q1 ~ residuals_q2)
   summary(regression_q4)
   ```

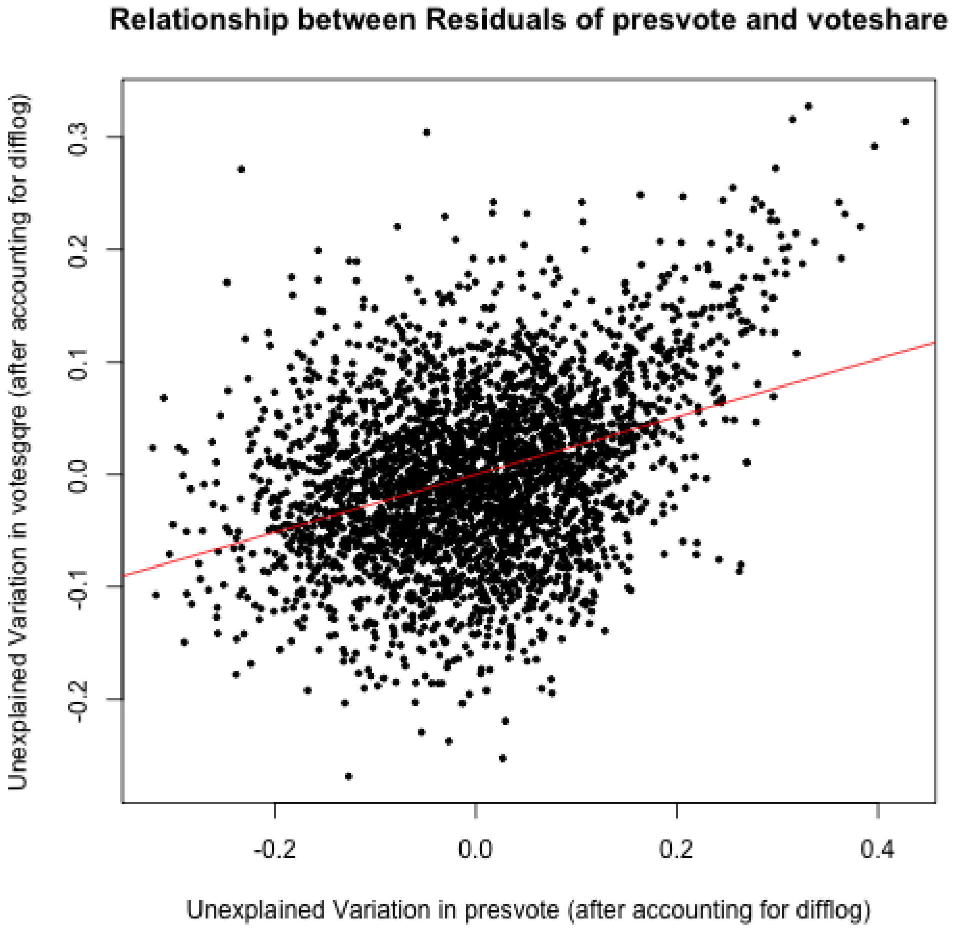
   We get the following results:

   ```
   Call:
   lm(formula = residuals_q1 ~ residuals_q2)

   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)  -1.942e-18  1.299e-03    0.00        1
   residuals_q2  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
   ```

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

   Let's make a scatterplot of the two residuals `residuals_q1` on the y-axis and `residuals_q2` on the x-axis. We will also add the regression line.

```
#Let's make a scatterplot of the two variables and add the regression line
png(file="scatter_plot_residuals_q2_residuals_q1.png")
plot(residuals_q2, residuals_q1,
     xlab = "Unexplained Variation in presvote (after accounting for difflog)",
     ylab = "Unexplained Variation in voteshare (after accounting for difflog)",
     main = "Relationship between Residuals of presvote and voteshare",
     pch = 19, col = "black", cex = 0.5)

# Add the regression line using abline
abline(regression_q4, col = "red")
dev.off()
```



3. Write the prediction equation.

   The prediction equation is:

   $$\text{residuals\_q1} = 0 + 0.2569 \times \text{residuals\_q2}$$

   This equation shows how much of the unexplained variation in `voteshare` (captured by `residuals_q1`) can be explained by the unexplained variation in `presvote` (captured by `residuals_q2`). The coefficient of 0.2569 indicates that for every 1-unit increase in the residuals from the `presvote` model (`residuals_q2`), the residuals from the `voteshare` model (`residuals_q1`) will increase by 0.2569 units.
   This positive association suggests that some variation in `voteshare`, which was initially unexplained by `difflog`, can be explained by variation in `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`.

   Let's run a regression where the incumbent's `voteshare` is the outcome variable and`difflog` and `presvote` is the explanatory variable.

   ```
   regression_q5 <- lm(voteshare ~ difflog + presvote, data= inc.sub)
   summary(regression_q5)
   ```

   We get the following results:

   ```
   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.

   The prediction equation is:

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

   This equation indicates that `voteshare` increases by 0.0355 for each one-unit increase in `difflog` and by 0.2569 for each one-unit increase in `presvote`. The intercept, 0.4486, represents the estimated `voteshare` when both difflog `difflog` and presvote `presvote` are zero.

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

   The coefficient 0.2569 is the same in both the equation in Question 4 ($\text{residuals\_q1} = 0+0.2569\times\text{residuals\_q2}$) and the equation in Question 5 ($\text{voteshare} = 0.4486+0.0355\times \text{difflog} + 0.2569 \times \text{presvote}$). This identical coefficient represents the same relationship between `voteshare` (the incumbent's vote share) and `presvote` (the vote share of the presidential candidate from the same party). This happens because both models focus on the link between `voteshare` and `presvote` and control for `difflog` (a measure of campaign spending difference).

   In Question 5, the model includes both `difflog` and `presvote` to explain `voteshare`. Here, the coefficient 0.2569 shows how much `presvote` affects `voteshare` while keeping `difflog` constant. It explains the effect of `presvote` on `voteshare` after excluding the effect of `difflog`.

   In Question 4, instead of directly including `difflog`, we control for it by using residuals. We run separate regressions of `voteshare` and `presvote` on `difflog` and then take the residuals. These residuals represent the part of `voteshare` and `presvote` that `difflog` does not explain, removing its influence from both variables. The coefficient 0.2569 shows the relationship between the residuals of `voteshare` and `presvote`, focusing on how they are related once `difflog` has been removed. This allows us to look at their direct relationship, without the effect of campaign spending differences.

   To conclude,this identical coefficient captures the relationship between `voteshare` and `presvote` in two different models, but each model controls for `difflog` in a different way, one directly and the other through residuals.

   This relationship makes sense because we would expect these variables to be related. If the presidential candidate from a particular party performs well, it reflects support that would likely benefit other candidates from the same party, like incumbents in Congressional races. The correlation shown by 0.2569 indicates how party-level support, represented by `presvote`, translates into local-level success for incumbents, represented by `voteshare`.