Problem Set 3

Applied Stats/Quant Methods 1

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
#Question 1
# Run the regression
model <- lm(voteshare ~ difflog, data = inc.sub)

# Display the summary of the regression results
summary(model)
```

   Then we can get

   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

We can see the coefficient for difflog is positive and statistically significant, so higher campaign spending relative to the challenger (a higher difflog value) is associated with an increased vote share for the incumbent.

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

```r
pdf("Scatterplot with regression line.pdf")
# Create scatterplot with regression line
ggplot(inc.sub, aes(x = difflog, y = voteshare)) +
  geom_point(color = "blue", alpha = 0.6) +          # Scatter plot with
    semi-transparent blue points
  geom_smooth(method = "lm", color = "red", se = TRUE) + # Regression
    line in red with confidence interval
  labs(title = "Scatterplot of Vote Share vs. Difflog",
       x = "Difflog",
       y = "Incumbent's Vote Share") +
  theme_minimal()          # A clean, minimal theme for the plot
dev.off()
```

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

```r
# Save the residuals in a separate object
residuals_model <- residuals(model)

# Display the first few residuals to confirm
head(residuals_model)
```

```
          1           2           3
-0.0004227622 -0.0316840149 -0.0045514943

          4           5           6
 0.0386688767  0.0355287965  0.0322832521
```

4. Write the prediction equation.
   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`.

```r
#Question 2
# Run the regression
model2 <- lm(presvote ~ difflog, data = inc.sub)

# Display the summary of the regression results
summary(model2)
```

Then we can get

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:

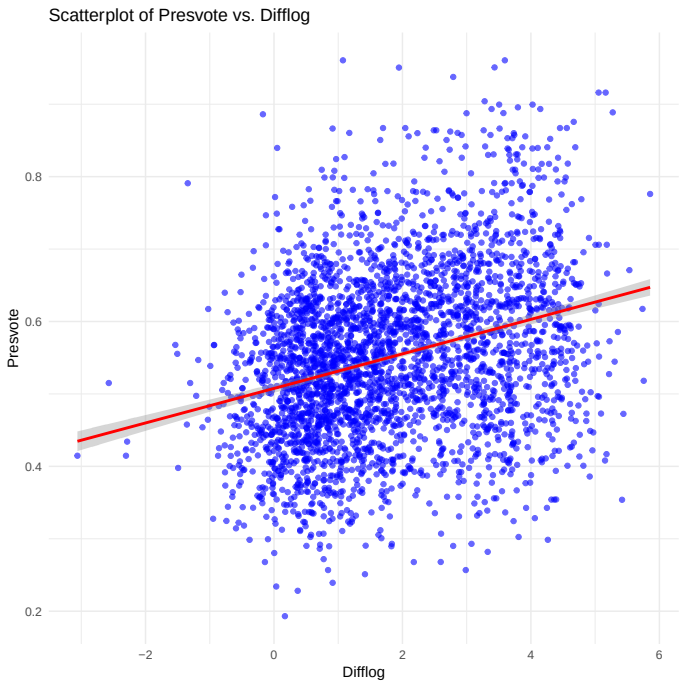
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.507583   0.003161  160.60   <2e-16 ***
difflog     0.023837   0.001359   17.54   <2e-16 ***
```

—
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

We can see the coefficient for difflog is positive and statistically significant, so higher campaign spending relative to the challenger (a higher difflog value) is associated with presvote.

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



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

```r
# Save the residuals in a separate object
residuals_model2 <- residuals(model2)

# Display the first few residuals to confirm
head(residuals_model2)
```

```
          1           2           3
0.005605594 0.037578519 -0.053134788

          4           5           6
-0.052993694 -0.045842994 0.074339701
```

4. Write the prediction equation.

   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`.

```r
#Question 3
# Run the regression
model3 <- lm(voteshare ~ presvote, data = inc.sub)

# Display the summary of the regression results
summary(model3)
```

Then we can get

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

We can see the coefficient for presvote is positive and statistically significant, so presvote is associated with voteshare.

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

3. Write the prediction equation.

   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.

```r
#Question 4
# Run the regression
model4 <- lm(residuals_model ~ residuals_model2, data = inc.sub)

# Display the summary of the regression results
summary(model4)
```

Then we can get

Call:
lm(formula = residuals_model ~ residuals_model2, data = inc.sub)

Residuals:

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

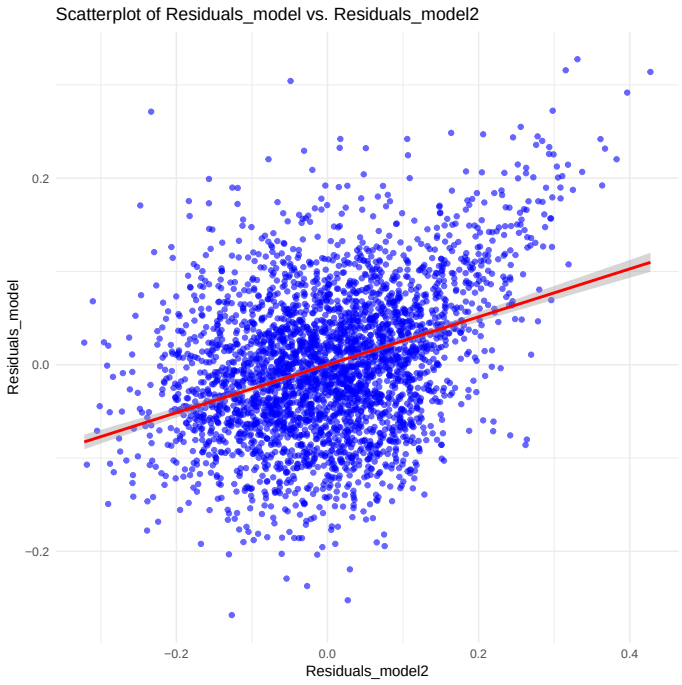
Coefficients:

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.942e-18  1.299e-03    0.00        1
residuals_model2  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.



3. Write the prediction equation.

   the residuals from Question 1=-1.942e-18+2.569e-01×the residuals from Question 2

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
#Question 5
# Run the regression
model5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)

# Display the summary of the regression results
summary(model5)
```

Then we can get

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.

   voteshare=0.4486442+0.0355431×difflog+0.2568770×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?

   The coefficient for presvote in the multiple regression of voteshare ~ difflog + presvote should be identical to the coefficient from the residuals regression in Question 4.

   This process is known as partial regression or adjustment for other predictors in multiple regression, which explains why the coefficient of presvote in this multiple regression is identical to the coefficient in the residual regression from Question 4.