Problem Set 3 Clare Zureich

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`.

   There is a positive relationship between the amount of incumbent and challenger spending and the incumbent's vote share. The incumbent's vote share will, on average, increase by .04 for a one unit increase in the difference in spending (logged). The relationship is significantly significant as the p-value is low.

   Below is the R Code and summary output of the regression:

```
regression_model1 <- lm(voteshare ~ difflog, data = inc.sub)
summary(regression_model1)
```

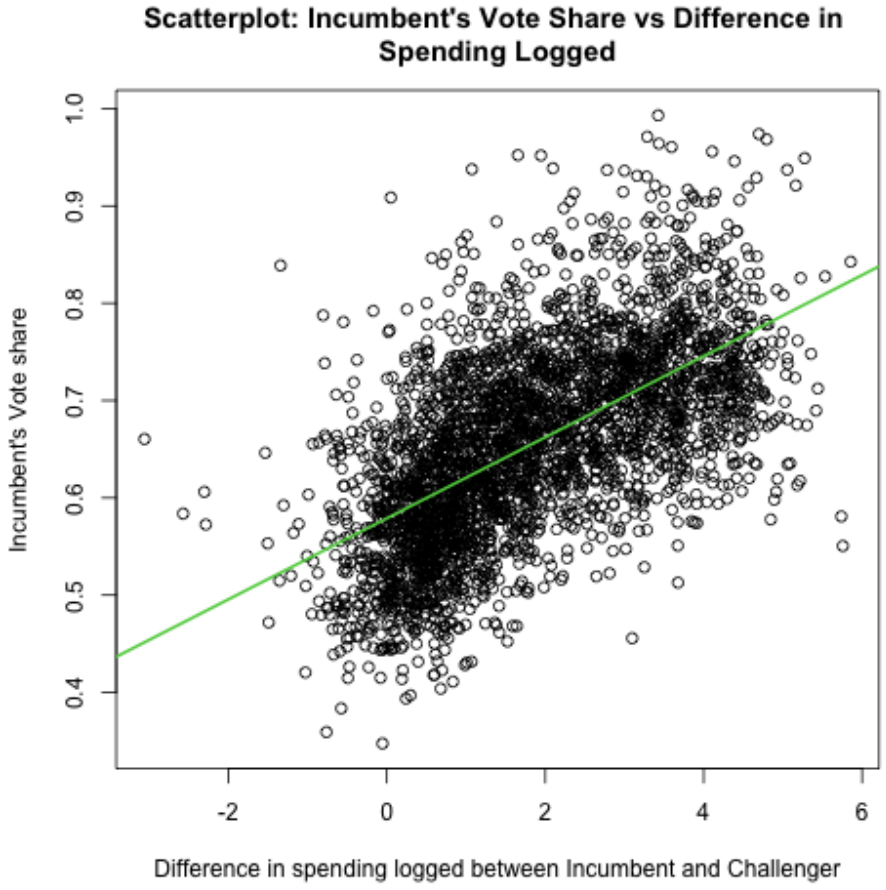
```
Call:
lm(formula = voteshare ~ difflog, data = inc.sub)

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.



```
png(file = "votesshare_vs_difflog.png")
plot(inc.sub$difflog, inc.sub$voteshare,
     ylab = "Incumbent's Vote share",
     xlab = "Difference in spending logged between Incumbent and
    Challenger",
     main = "Scatterplot: Incumbent's Vote Share vs Difference in \
    nSpending Logged")
abline(regression_model1, col=3, lwd = 2)
```

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

```
regression_residuals1 <- residuals(regression_model1)
summary(regression_residuals1)
```

```
      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.268319 -0.053454 -0.003769  0.000000  0.047798  0.327488
```

4. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{difflog}$$

voteshare = 0.579 + 0.042 x 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`.

   There is a positive relationship between the amount of incumbent and challenger spending and the presidential vote share. The presidential vote share will, on average, increase by .024 for a one unit increase in the difference in spending (logged). The relationship is significantly significant as the p-value is low.

   Below is the R Code and summary output of the regression:

```r
regression_model2 <- lm(presvote ~ difflog, data = inc.sub)
summary(regression_model2)
```

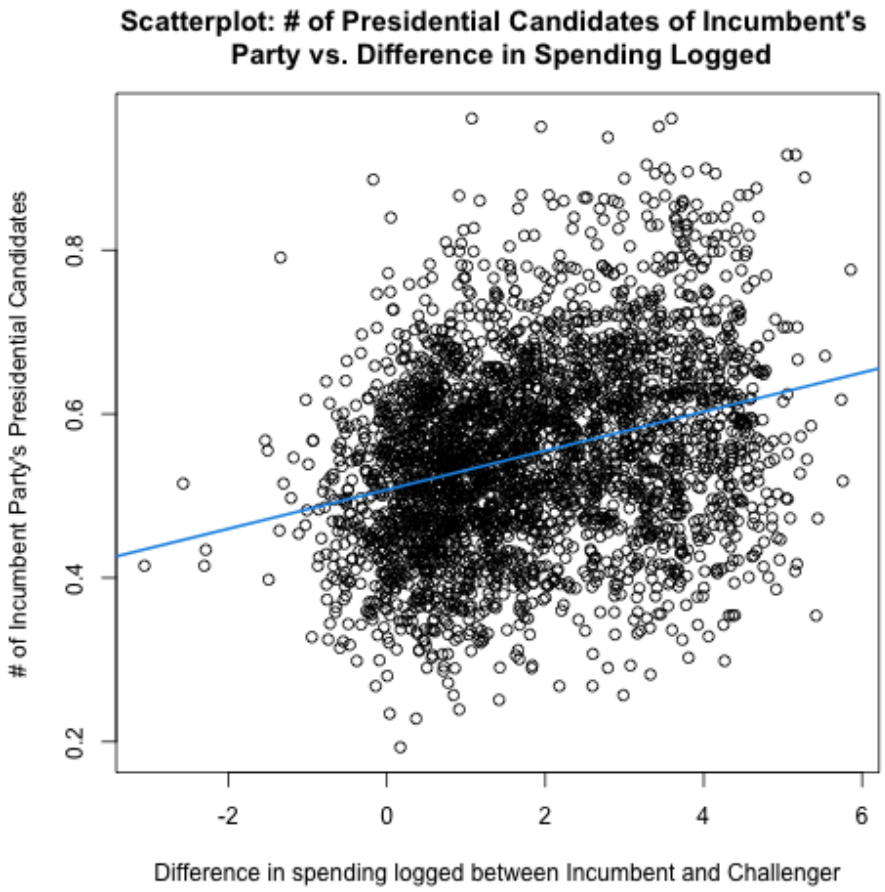
```
Call:
lm(formula = presvote ~ difflog, data = inc.sub)

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.




```
png(file = "presvote_vs_difflog.png")
plot(inc.sub$difflog, inc.sub$presvote,
     ylab = "# of Incumbent Party's Presidential Candidates",
     xlab = "Difference in spending logged between Incumbent and
   Challenger",
     main = "Scatterplot: # of Presidential Candidates of Incumbent's \n
   Party vs. Difference in Spending Logged")
abline(regression_model2, col=4, lwd = 2)
```

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

```
regression_residuals2 <- residuals(regression_model2)
summary(regression_residuals2)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.321965 -0.074069 -0.001018  0.000000  0.071507  0.427435
```

4. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{difflog}$$

presvote = 0.508 + 0.024 x 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`.

   There is a positive relationship between the presidential vote share and the incumbent's vote share. The incumbent's vote share will, on average, increase by .388 for a one unit increase in the presidential vote share. The relationship is significantly significant as the p-value is low.

   Below is the R Code and summary output of the regression:

```
regression_model3 <- lm(voteshare ~ presvote, data = inc.sub)
summary(regression_model3)
```

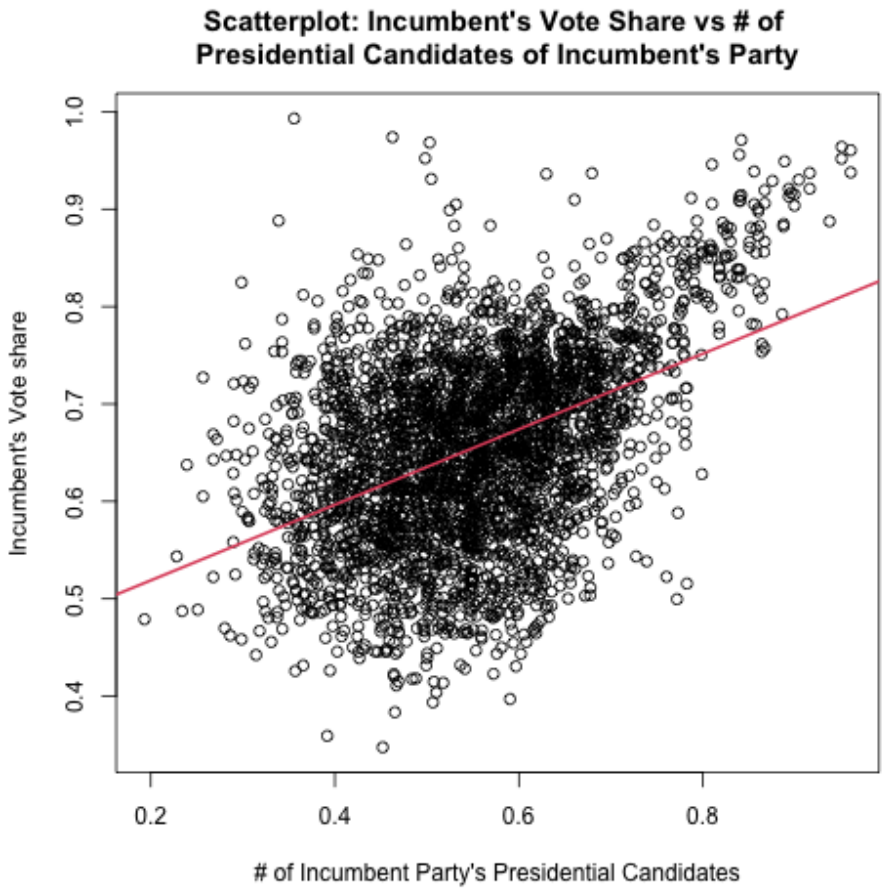
```
Call:
lm(formula = voteshare ~ presvote, data = inc.sub)

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
png(file = "votesshare_vs_presvote.png")
plot(inc.sub$presvote, inc.sub$voteshare,
     ylab = "Incumbent's Vote share",
     xlab = "# of Incumbent Party's Presidential Candidates",
     main = "Scatterplot: Incumbent's Vote Share vs # of \nPresidential
   Candidates of Incumbent's Party")
abline(regression_model3, col=2, lwd = 2)
```

3. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{presvote}$$

voteshare = 0.441 + 0.388 x 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.

   There is a positive relationship between the residuals from regression 1 (vote share vs difflog) and the residuals from regression 2 (presvote vs difflog). The residual error of regression 1 will, on average, increase by .257 for a one unit increase in regression 2 residuals error. The relationship is significantly significant as the p-value is low.

```r
residual_regression_model <- lm(regression_residuals1 ~ regression_
    residuals2)
summary(residual_regression_model)
```

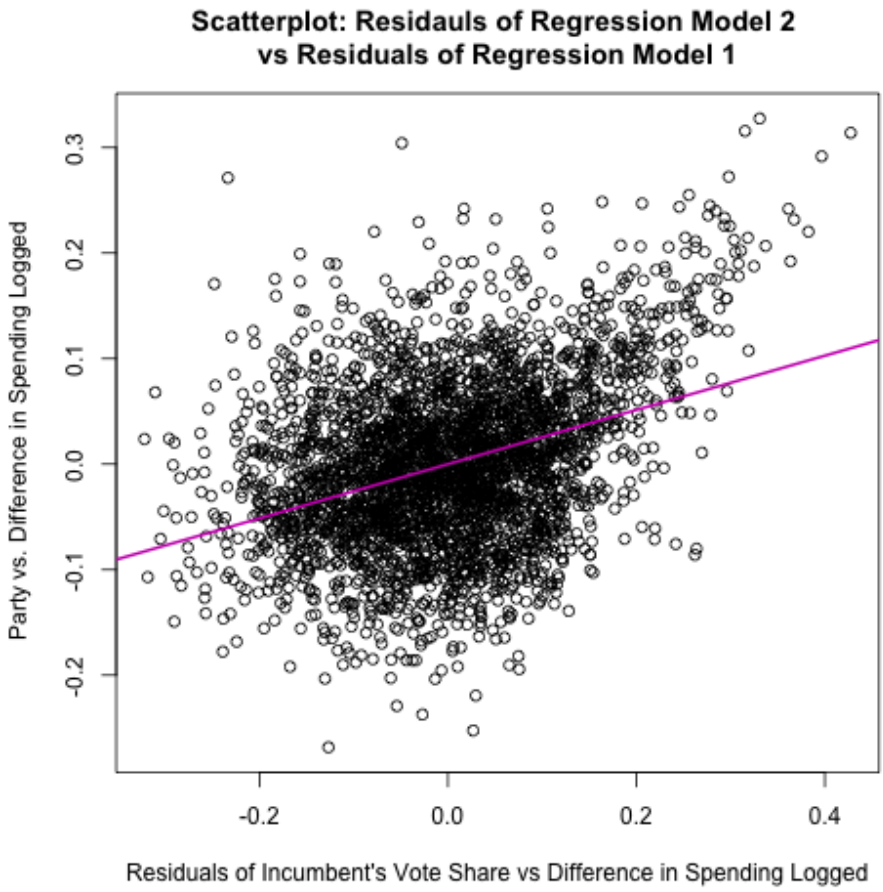
```
Call:
lm(formula = regression_residuals1 ~ regression_residuals2)

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
regression_residuals2  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.




```
png(file = "Q1_vs_Q2.png")
plot(regression_residuals2, regression_residuals1,
     ylab = "Residuals of # of Presidential Candidates of Incumbent's \n
    Party vs. Difference in Spending Logged",
     xlab = "Residuals of Incumbent's Vote Share vs Difference in
    Spending Logged",
     main = "Scatterplot: Residauls of Regression Model 2 vs Residuals of
     Regression Model 1")
abline(residual_regression_model, col=6, lwd = 2)
```

3. Write the prediction equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{ResidualsQ2}$$

ResidualsQ1= 0 + 0.257 ResidualsQ2

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`.

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

```
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.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{difflog} + \hat{\beta}_2 \times \text{presvote}$$

voteshare = .449 + .036 x difflog + .257 x 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 slope coefficient of the regressionresidual2 variable in the residual regression model (.257) is equal to the slope coefficient of the presvote variable in the multivariate linear model (and prediction equation) (.257). This is because the slope coefficients in both models represent the effect of presvote on voteshare after taking out the effects of difflog from both voteshare and presvote. This is the partial effect of presvote.

Step by step analysis: In regressionresiduals1, we first found the residual of the linear relationship between votesahre and difflog (regressionmodel1), which is the part of voteshare that is not linearly related to difflog. In regressionresiduals2, we found the residual of the linear relationship between presvote and difflog (regressionmodel2), which is the part of presvote that is not linearly related to difflog. We then found the linear relationship between the voteshare residual (regressionresidual1) and the presvote residual (regressionresidual2). The result is the coefficient which represents the effect of presvote on voteshare after taking out the effects of difflog from voteshare and presvote