Problem Set 3

QTM 200: Applied Regression Analysis

Due: February 17, 2020

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 the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Monday, February 17, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

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 (20 points)

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`.
   I first load the dataset into R and rename it as "incum". Then I run the regression with X variable as difflog and Y variable as voteshare. I then check the estimateed coefficients of the model using `summary()`.

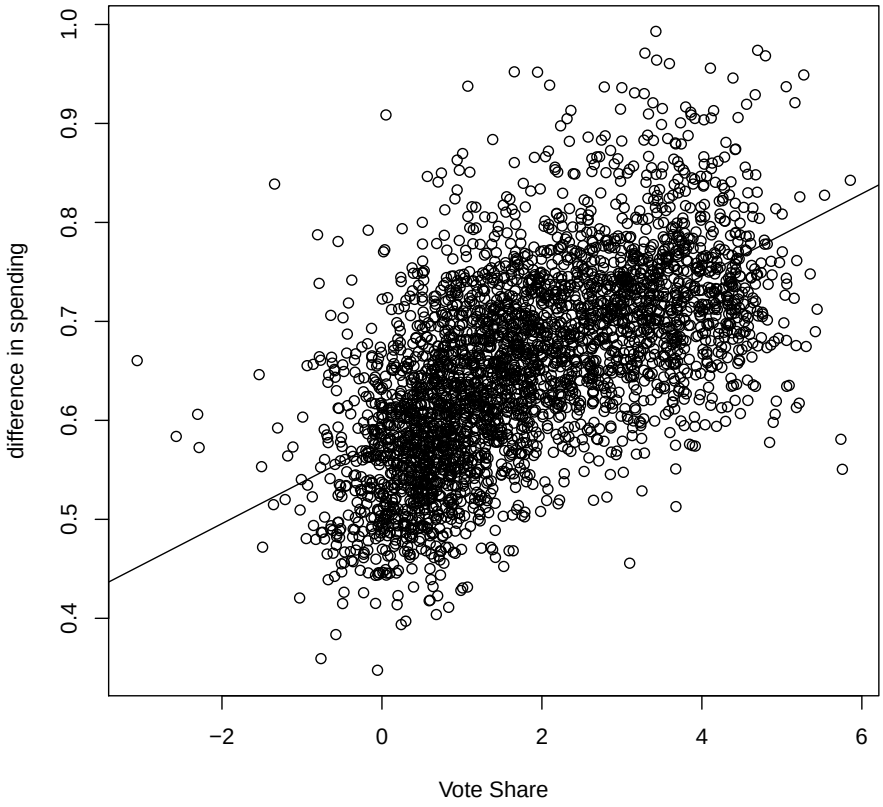
```
incum <- read.csv("incumbents_subset.csv")
#run the regression with y=voteshare, x=difflog
regression_p1 <- lm(voteshare ~ difflog, data=incum)
#check the summary of model with coefficient estimates
summary(regression_p1)
```

```
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.

Figure 1: Scatterplot of `voteshre` and `difflog`.



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

```r
#3. Save the residuals of the model in a separate object.
residuals_1 <- regression_p1$residuals
```

4. Write the prediction equation.

The simple linear prediction equation is $\hat{Y}_i = \beta_0 + \beta_1 * X_i$.
I use the coefficients from the previous question to find β_0, which is the $Y - intercept$ when $X = 0$, and β_0, which is the estimated slope.

```
regression_p1$coefficients #check the values of beta0 and beta1.
# Yˆi = 0.579 + 0.042*Xi
```

The prediction equation is $\hat{Y}_i = 0.579 + 0.042 * X_i$.

Question 2 (20 points)

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`.
   I run the regression with X variable as `difflog` and Y variable as `presvote`. I then check the estimateed coefficients of the model using `summary()`.

```
#run the regression with y=presvote, x=difflog
regression_p2 <- lm(presvote ~ difflog, data=incum)
#check the summary of model with coefficient estimates
summary(regression_p2)
```

```
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.

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

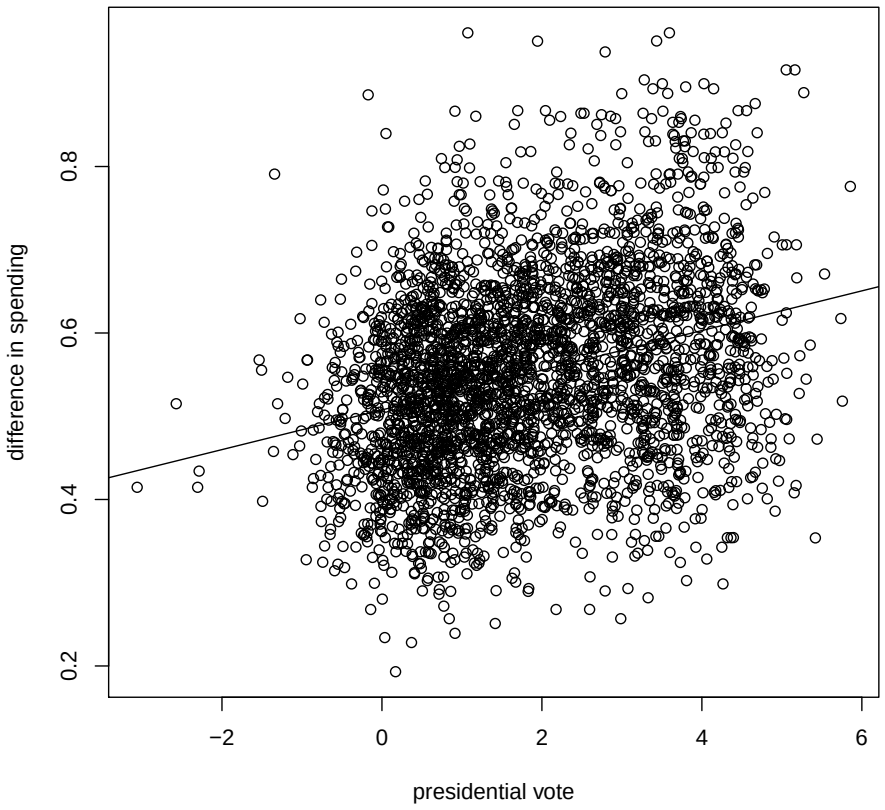
```
#3. Save the residuals of the model in a separate object.
residuals_2 <- regression_p2$residuals
```

4. Write the prediction equation.

```
regression_p2$coefficients #check the values of beta0 and beta1.
# Yˆi = 0.507 + 0.024*Xi
```

   The prediction equation is $\hat{Y}_i = 0.507 + 0.024 * X_i$.

Figure 2: Scatterplot of `difflog` and `presvote`.



Question 3 (20 points)

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`.
   I run the regression with X variable as `presvote` and Y variable as `voteshare`. I then check the estimateed coefficients of the model using `summary()`.

```
#run the regression with y=voteshare, x=presvote
regression_p3 <- lm(voteshare ~ presvote, data=incum)
#check the summary of model with coefficient estimates
summary(regression_p3)
```

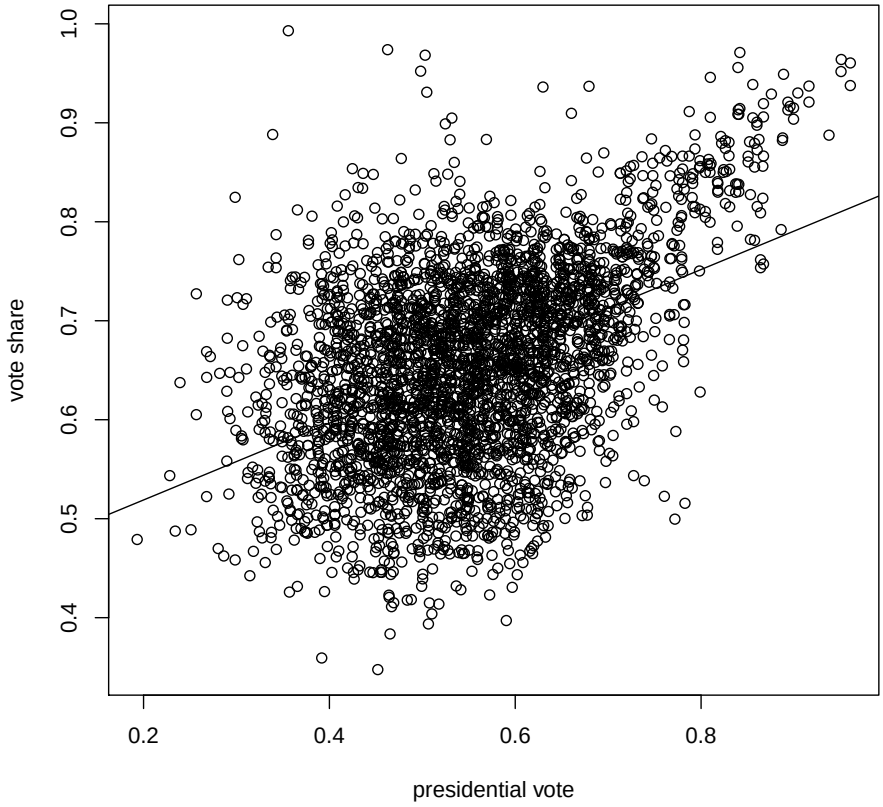
```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.441330   0.007599   58.08   <2e-16 ***
```

```
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.

Figure 3: Scatterplot of `presvote` and `voteshare`.



3. Write the prediction equation.

```
regression_p3$coefficients #check the values of beta0 and beta1.
# Yˆi = 0.441 + 0.388*Xi
```

The prediction equation is $\hat{Y}_i = 0.441 + 0.388 * X_i$.

Question 4 (20 points)

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.
   I run the regression with X variable as `residuals2` and Y variable as `residuals1`. I then check the estimateed coefficients of the model using `summary()`.

```
#run the regression with y=residuals_1, x=residuals_2
regression_p4 <- lm(residuals_1 ~ residuals_2)
#check the summary of model with coefficient estimates
summary(regression_p4)
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.860e-18  1.299e-03    0.00        1
residuals_2  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
```

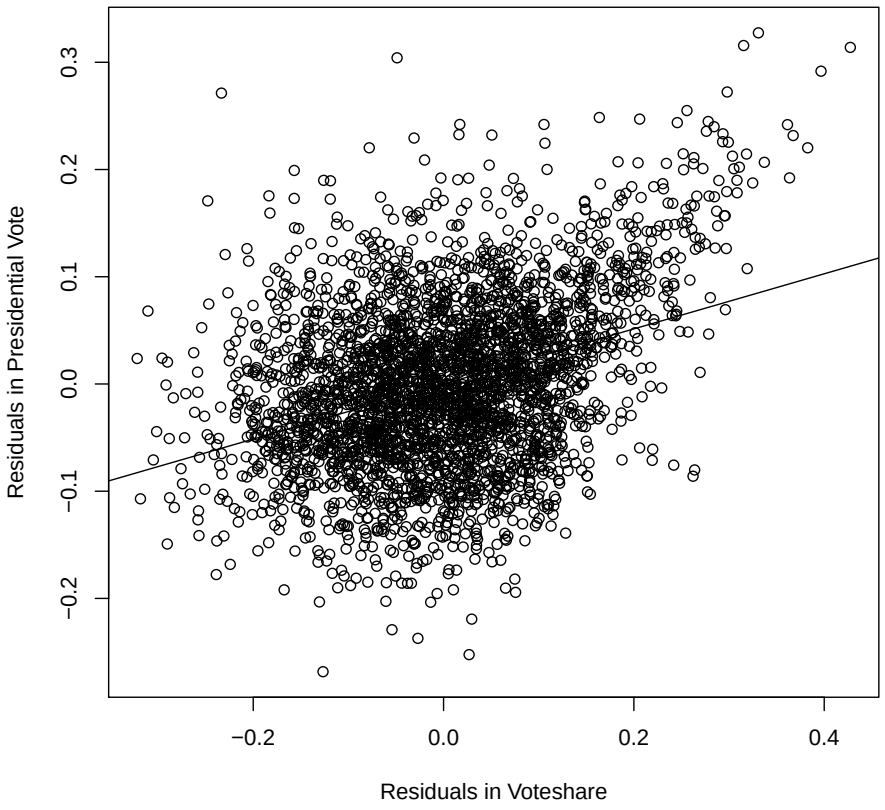
2. Make a scatterplot of the two residuals and add the regression line.

3. Write the prediction equation.

```
#3. check the values of beta0 and beta1.
regression_p4$coefficients
```

   The prediction equation is $\hat{Y}_i = -4.860 + 0.256 * X_i$.

Figure 4: Scatterplot of `residuals2` and `residuals1`.



Question 5 (20 points)

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`.
   I run the regression with X variable as `difflog` and `presvote` and Y variable as `voteshare`. I then check the estimateed coefficients of the model using `summary()`.

```r
#run the regression with y=voteshare, x=difflog and presvote
regression_p5 <- lm(voteshare ~ difflog+presvote, data=incum)
#check the summary of model with coefficient estimates
summary(regression_p5)
```

```
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.

```
regression_p5$coefficients #check the values of beta0 and beta1.
# Yˆi = 0.4486 + 0.0355*Xi(difflog) + 0.2569*Xi(presvote)
```

The prediction equation is $\hat{Y}_i = 0.4486 + 0.0355 * X_(difflog) + 0.2569 * X_(presvote)$.

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 coefficient of variable `presvote` is identical to the coefficient of `residual2`. It makes sense because `residual2` means how much of the variation in `presvote` is not explained by the model.