Problem Set 3

Applied Stats/Quant Methods 1

Due: November 20, 2022

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 20, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

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`.

   First I load the data into my R Studio session using the below code:

   ```
   inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_
       Fall2022/main/datasets/incumbents_subset.csv")
   ```

   Then I run the regression with the following code:

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

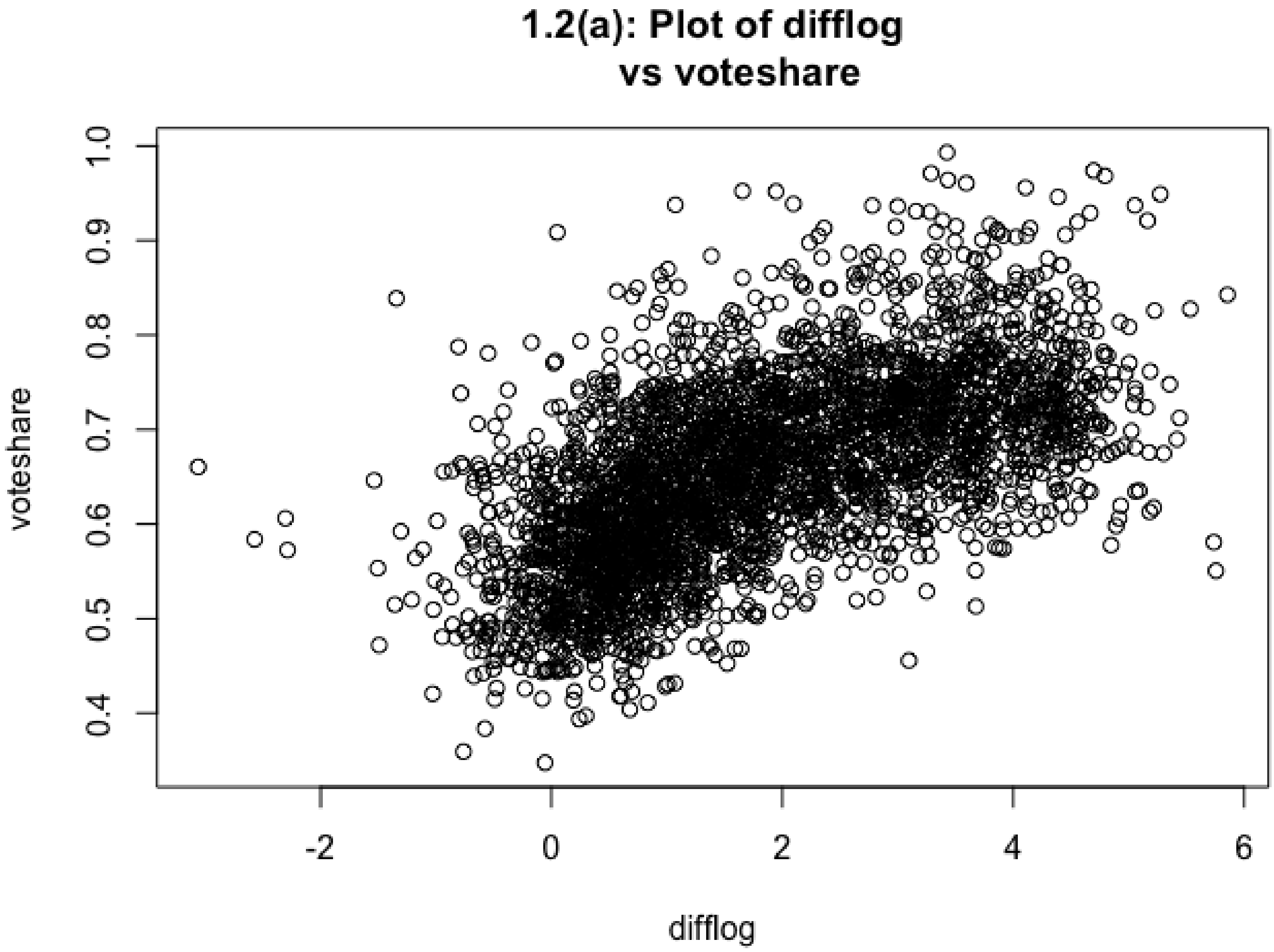
```
Residuals:     Min       1Q   Median       3Q      Max
 -0.26832 -0.05345 -0.00377  0.04780  0.32749
 Coefficients:
             Estimate    Std. Error  t value Pr(>|t|)
 (Intercept) 0.579031   0.002251  257.19   <2e-16 ***
 difflog     0.041666   0.000968   43.04   <2e-16 ***
 ---
 Signif. codes:  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.

   First I made the scatterplot (with x = difflog, y = voteshare) using the following code:

```r
plot(inc.sub$difflog, inc.sub$voteshare, main = "1.2(a): Plot of difflog
     vs voteshare",
     xlab = "difflog", ylab = "voteshare")
```

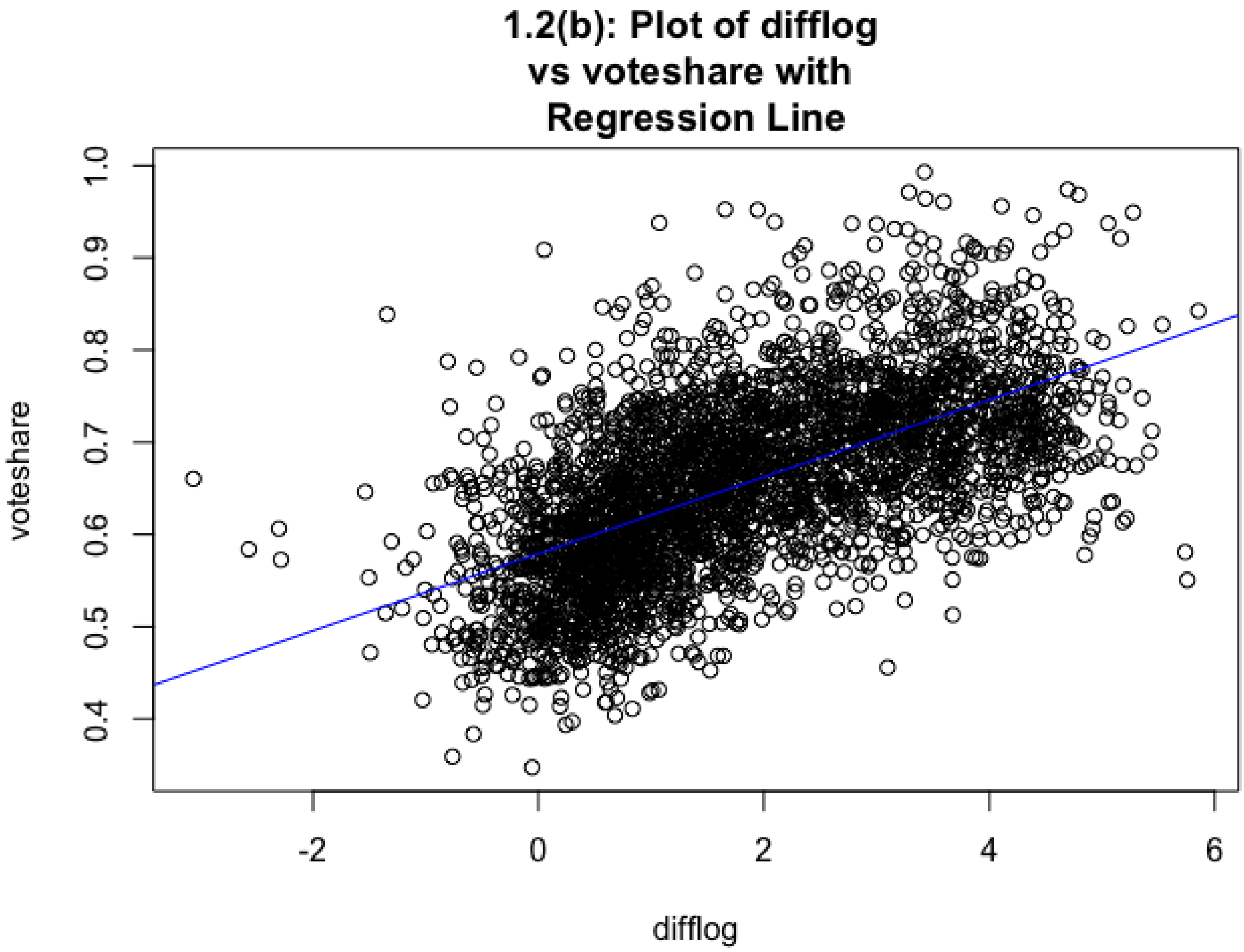
   The scatterplot is below:



Then I added the Regression Line to this plot with the following code:

```
plot(inc.sub$difflog, inc.sub$voteshare, main = "1.2(b): Plot of difflog
vs voteshare with
Regression Line",
     xlab = "difflog", ylab = "voteshare")
abline(lm(voteshare ~ difflog, data = inc.sub), col = "blue")
```

Which produced the following plot:



3. Save the residuals of the model in a separate object. I do this with the following code:

```
Res_lm_vs_dl <- lm_vs_dl$residuals
```

4. Write the prediction equation.

   y hat = intercept + slope*x

   Predicted voteshare = 0.579 + 0.042(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`.

   Code for regression with outcome variable presvote, explanatory variable:

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

   Summary of this regression output is below:

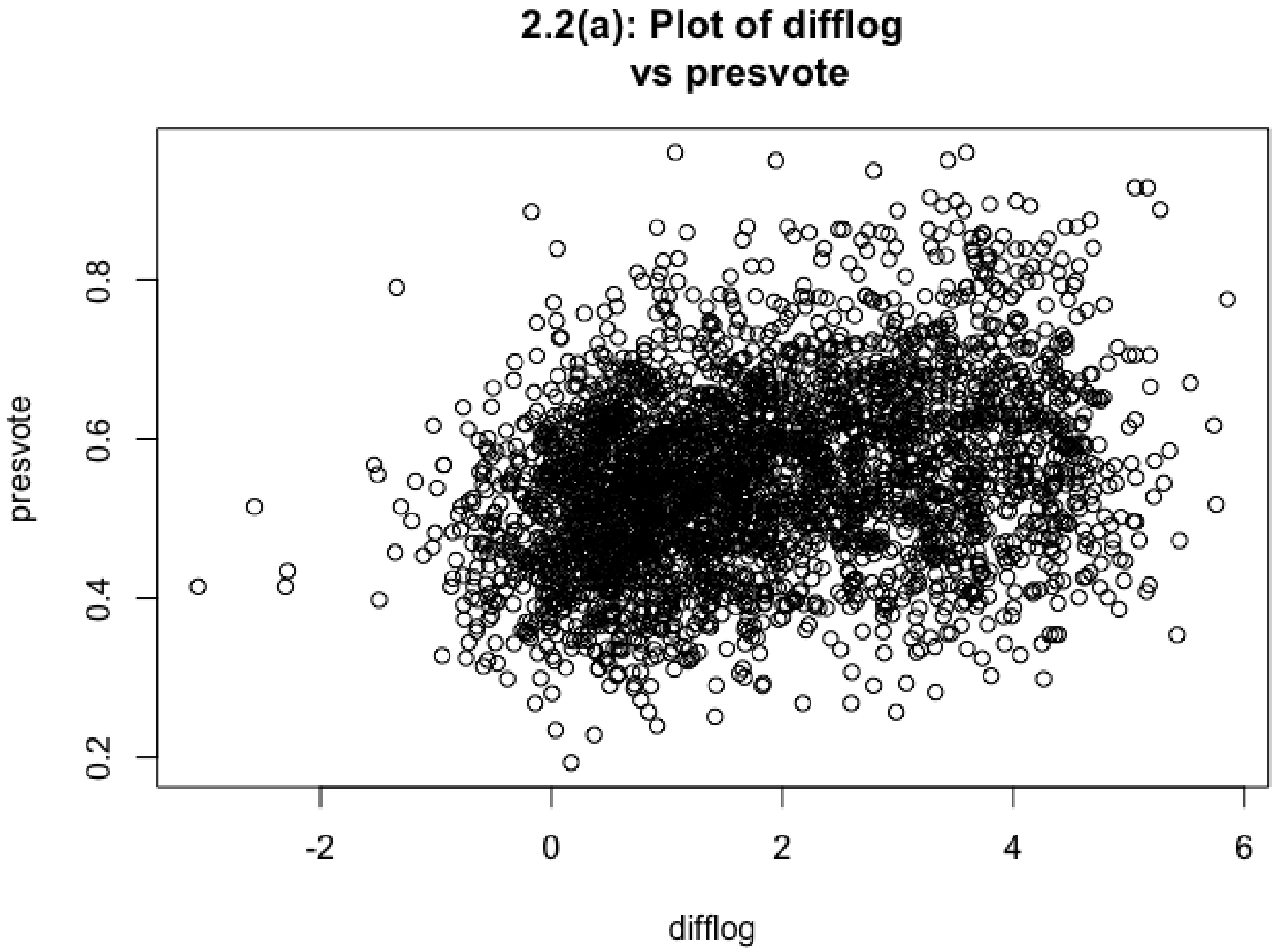
   ```
   Residuals:     Min       1Q   Median       3Q      Max
   -0.32196 -0.07407 -0.00102  0.07151  0.42743
    Coefficients:
             Estimate Std. Error t value Pr(>|t|)
       (Intercept) 0.507583   0.003161  160.60   <2e-16 ***
       difflog     0.023837   0.001359   17.54   <2e-16 ***
       ---
       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.

   To first plot the two variables I input the following code:

   ```
   plot(inc.sub$difflog , inc.sub$presvote , main = ”2.2(a): Plot of difflog
        vs presvote”,
        xlab = ”difflog”, ylab = ”presvote”)
   ```

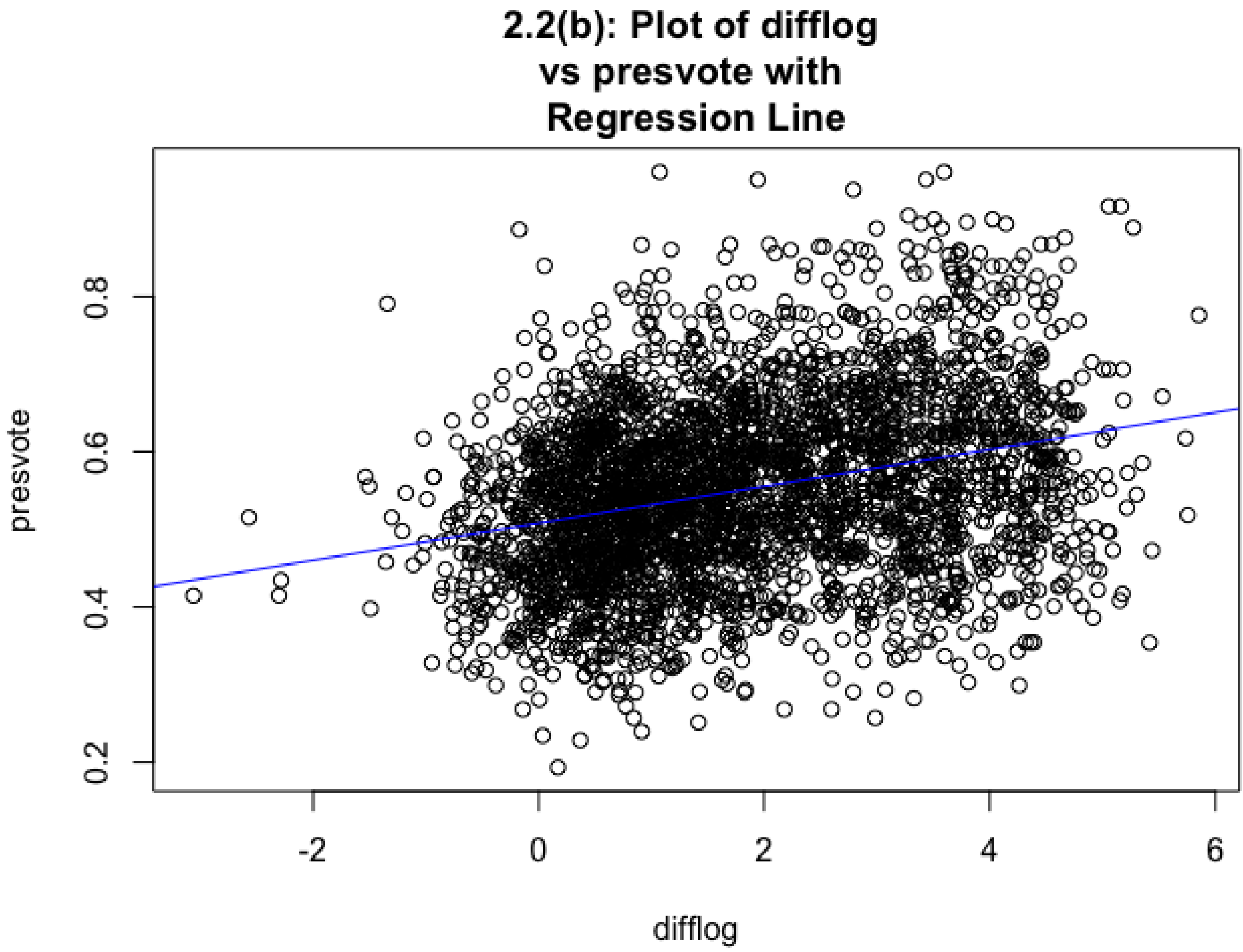
   The plot of these two variables is below:



Then I added the regression line with the following code:

```
plot(inc.sub$difflog, inc.sub$presvote, main = "2.2(b): Plot of difflog
vs presvote with
Regression Line",
     xlab = "difflog", ylab = "presvote")
abline(lm(presvote ~ difflog, data = inc.sub), col = "blue")
```

The plot produced with the regression line included is below:



3. Save the residuals of the model in a separate object. I did this with the following code:

```r
Res_lm_pv_dl <- lm_pv_dl$residuals
```

4. Write the prediction equation.

   y hat = intercept + slope*x

   Predicted 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`.

   Regression run with those variables with the below code:

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

   Summary output of this regression:

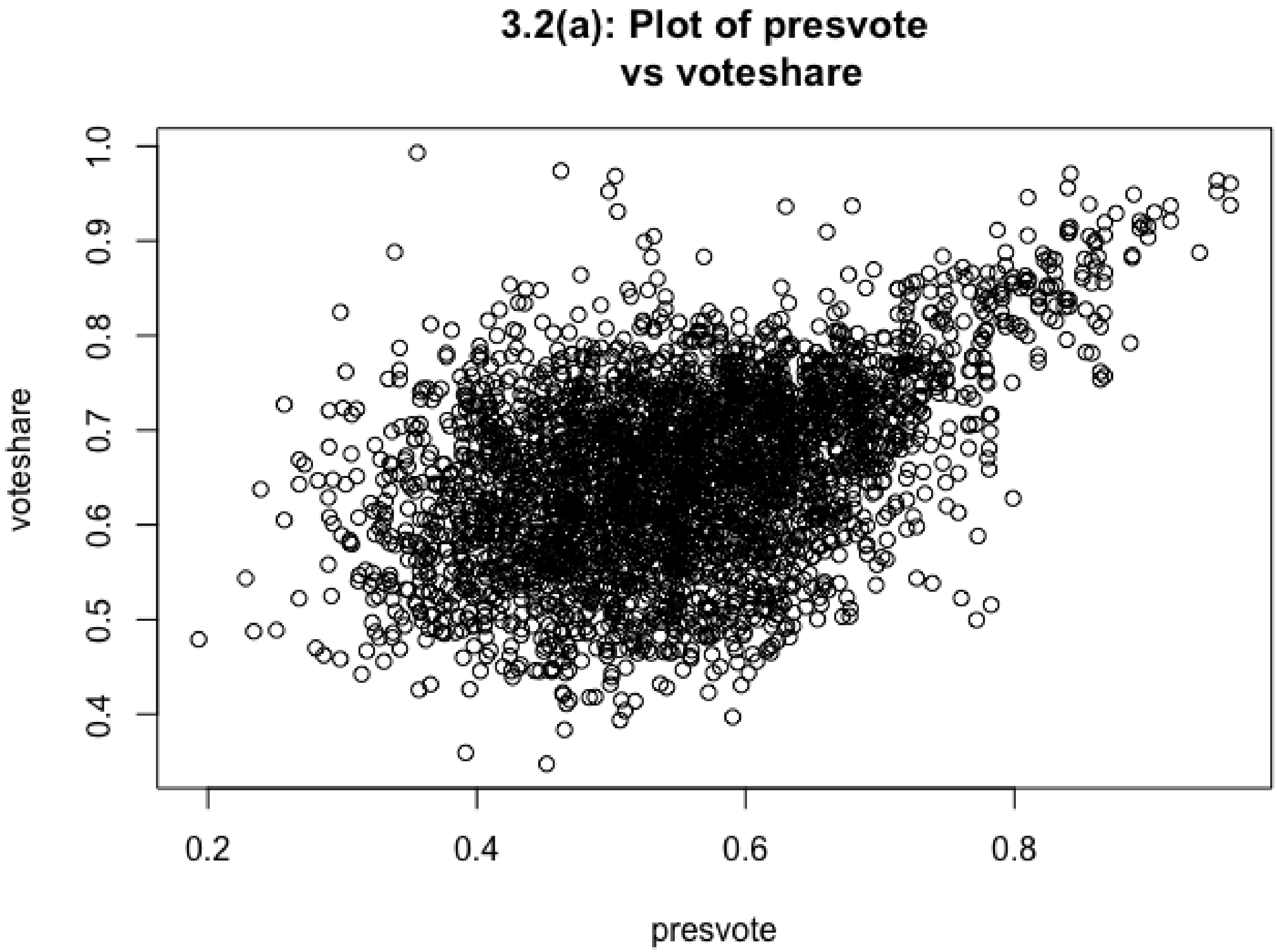
   ```
   Residuals:     Min       1Q   Median       3Q      Max
   -0.27330 -0.05888  0.00394  0.06148  0.41365
   Coefficients:
   Estimate Std. Error t value Pr(>|t|)
   (Intercept) 0.441330   0.007599   58.08   <2e-16
   presvote    0.388018   0.013493   28.76   <2e-16
   (Intercept) ***
   presvote    ***
   ---
   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.

   I plot the two variables using the following code:

   ```
   plot(inc.sub$presvote, inc.sub$voteshare, main = "3.2(a): Plot of
       presvote
         vs voteshare",
         xlab = "presvote", ylab = "voteshare")
   ```

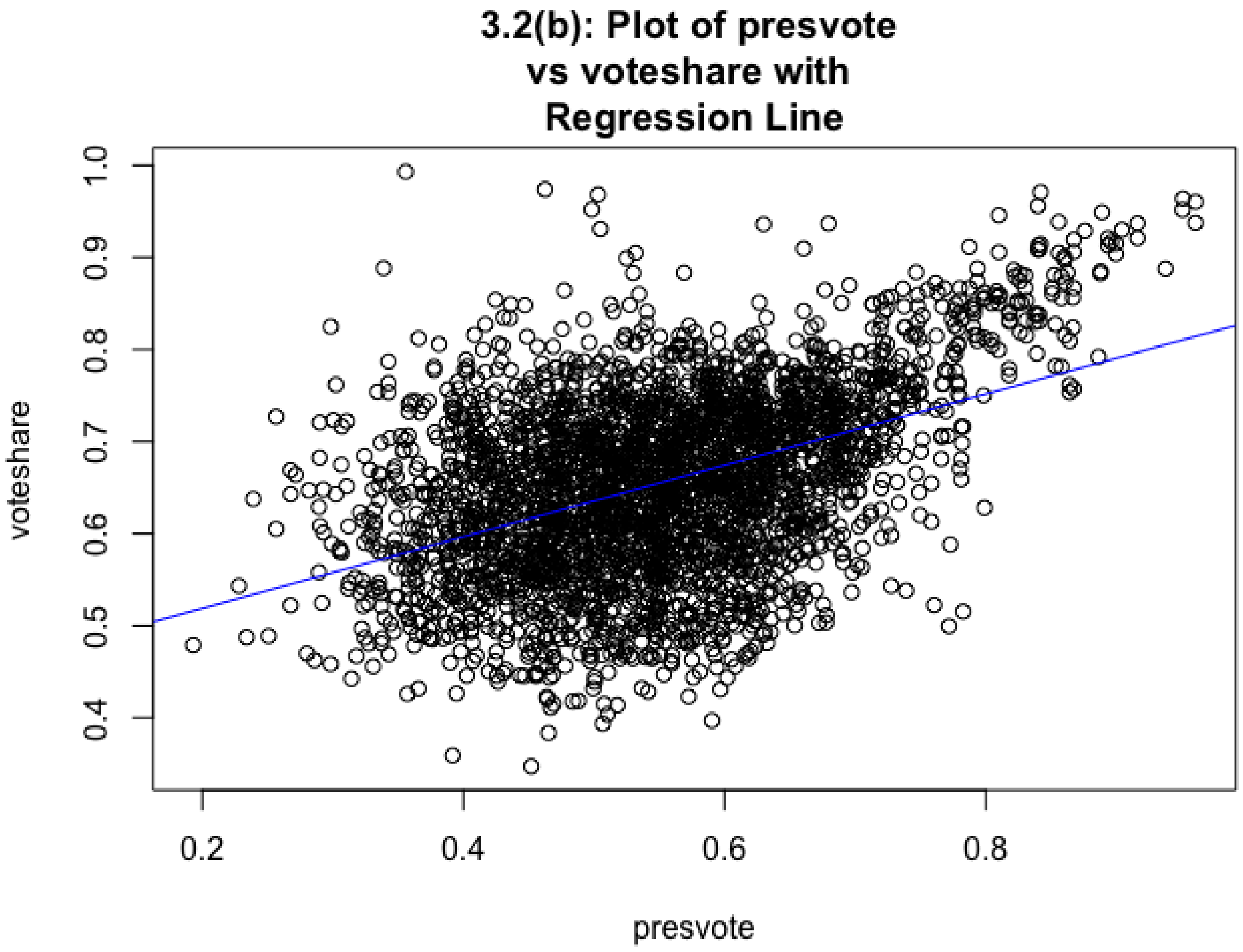
   The plot of these two variables is below:



Then I added the regression line with the following code:

```
plot(inc.sub$presvote, inc.sub$voteshare, main = "3.2(b): Plot of presvote
vs voteshare with
Regression Line",
     xlab = "presvote", ylab = "voteshare")
abline(lm(voteshare ~ presvote, data = inc.sub), col = "blue")
```

The plot produced with the regression line included is below:



3. Write the prediction equation.

   y hat = intercept + slope*x

   Predicted 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.

   To run this regression I input the following code:

   ```
   lm_rq1_rq2 <- lm(Res_lm_vs_dl~Res_lm_pv_dl)
   ```

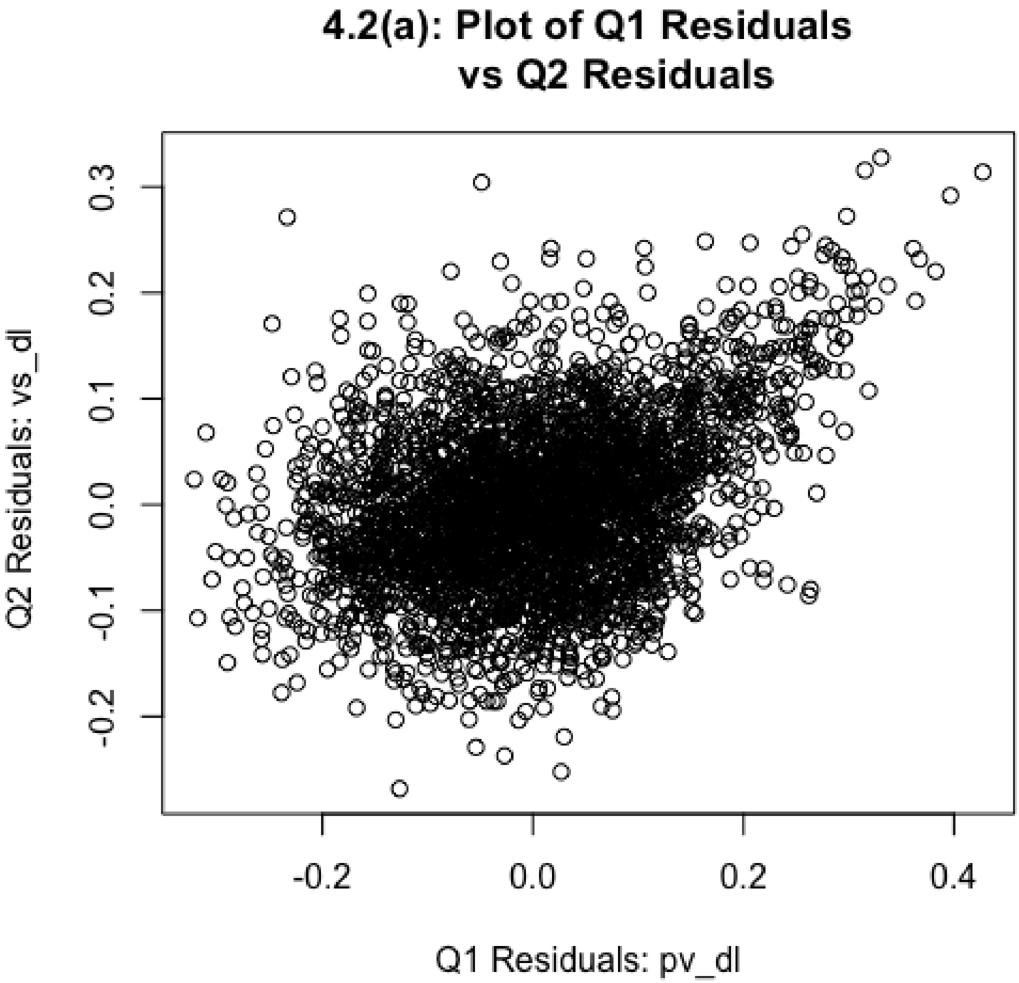
   Which output the following regression summary:

   ```
   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)  -4.860e-18  1.299e-03    0.00        1
     Res_lm_pv_dl  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. I make the scatterplot with the following code:

   ```
   plot(Res_lm_pv_dl, Res_lm_vs_dl, main = "4.2(a): Plot of Q1 Residuals
        vs Q2 Residuals",
        xlab = "Q1 Residuals: pv_dl", ylab = "Q2 Residuals: vs_dl ")
   ```

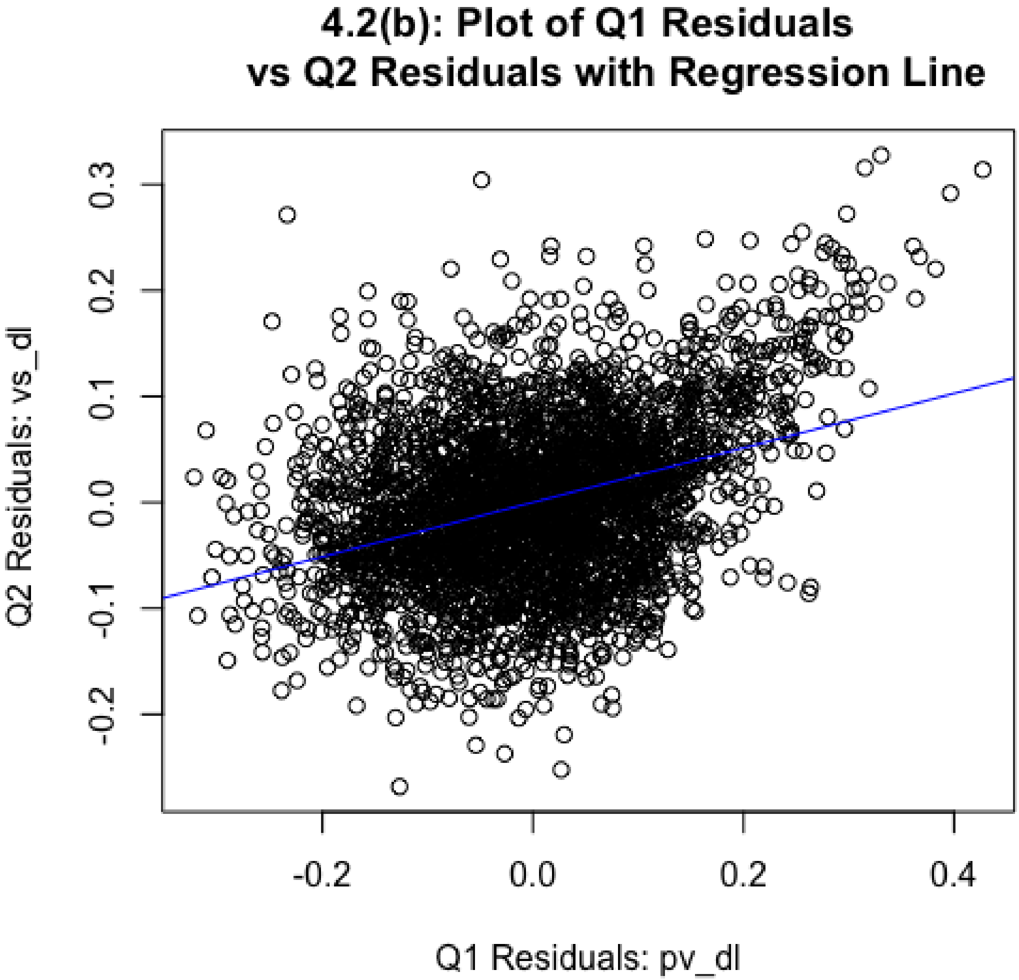
   This produces the scatterplot below:



I add the regression line with the following code:

```
plot(Res_lm_pv_dl, Res_lm_vs_dl, main = ”4.2(b): Plot of Q1 Residuals
     vs Q2 Residuals with Regression Line”,
     xlab = ”Q1 Residuals: pv_dl”, ylab = ”Q2 Residuals: vs_dl ”)
abline(lm(Res_lm_vs_dl~Res_lm_pv_dl), col = ”blue”)
```

This produces the scatterplot plus regression line below:



3. Write the prediction equation.

   The prediction equation is: y hat = intercept + slope*x

   (Amount of variation in vote share not explained by the diff in spending between the incumbent and challenger) = -4.860e-18 + 2.569e-1(variation in presvote not explained by the difference in spending between the incumbent and challenger)

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`. To run this regression I input the following code:

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

   Which results in this output:

   ```
   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
   (Intercept) ***
   difflog     ***
   presvote    ***
   ---
   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. Prediction equation is: The prediction equation is: y hat = intercept + slope*x

   (predicted 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?

   The residuals are identical in this output and that of Question 4. This is because Question 4 runs a regression that's output shows how much that variation in voteshare not explained by difflog is explained by the variation in presvote that is also not explained by difflog. Then, when the residuals are taken from this Question 4 output, these residuals show how much of the voteshare variation is NOT explained both by difflog and presvote. Then, the additive linear regression model ran in Question 5 shows how much variation in voteshare is explained by the difflog plus presvote variables. Thus, the residuals of this regression model (from Question 5) are essentially showing the same thing as the residuals in Question 4: the variation in voteshare that is NOT explained by both difflog and presvote. Essentially, these residuals were obtained by running regressions of two different models, can be interpreted as showing the same thing.