Problem Set 3

Applied Stats/Quant Methods 1: 16327268, Marcus Ó Faoláin

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`.

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

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

4. Write the prediction equation.

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`.

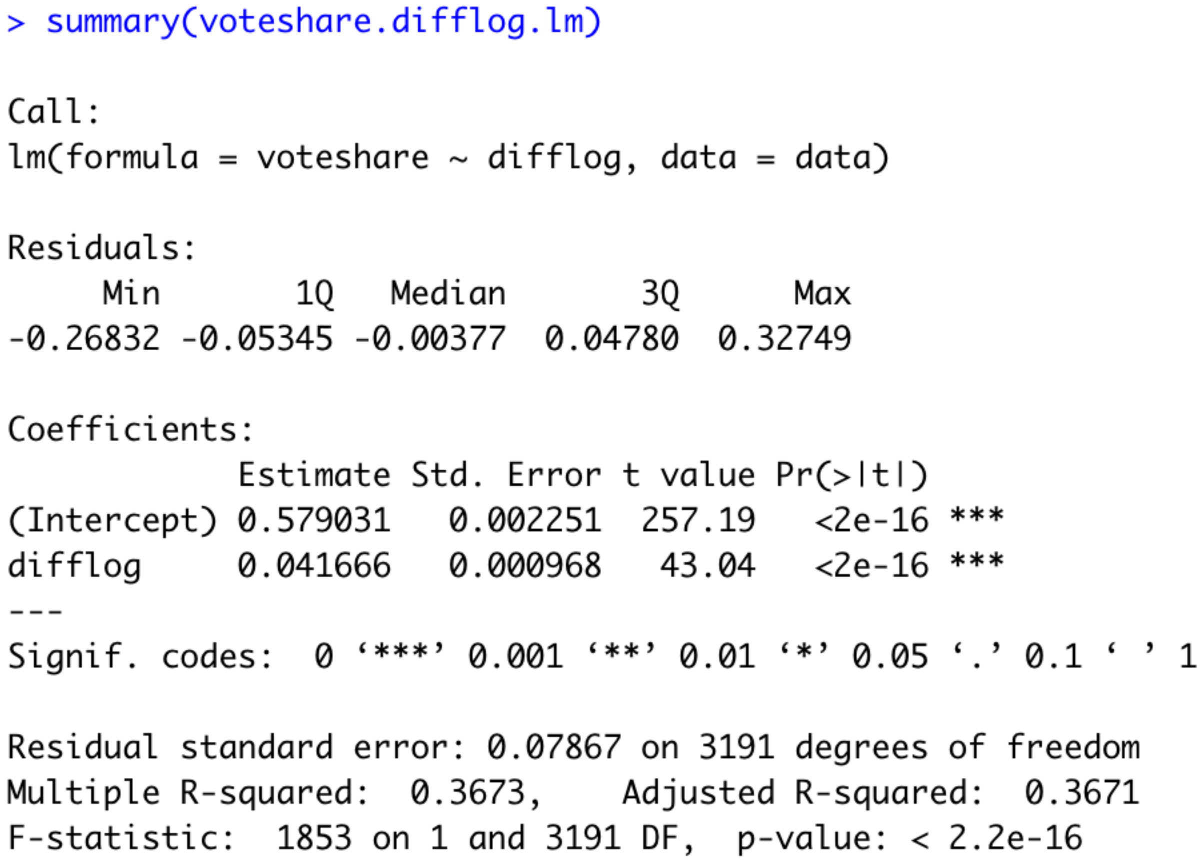
   We run a regression where `voteshare` is the outcome variable (Y) and `difflog` is the explanatory variable (X). We save it in a variable `voteshare.difflog.lm` and use the `summary()` function to view the results. Finally, we generate Latex output using `stargazer` which generates the table below.

```
voteshare.difflog.lm <- lm(voteshare ~ difflog, data = data)
summary(voteshare.difflog.lm)
stargazer(voteshare.difflog.lm, type='latex', summary= FALSE)
```

Table 1:

	Dependent variable:
	voteshare
difflog	0.042^{***}
	(0.001)
Constant	0.579^{***}
	(0.002)
Observations	3,193
R^2	0.367
Adjusted R^2	0.367
Residual Std. Error	0.079 (df = 3191)
F Statistic	$1,852.791^{***}$ (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

Attached is a screenshot of the console output when we run a summary call on the `voteshare.difflog.lm` model.



```
> summary(voteshare.difflog.lm)

Call:
lm(formula = voteshare ~ difflog, data = data)

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
```

The residuals are the difference between the predicted and observed values in the model. The min value for these residuals is -0.26832 and the max value is 0.32749. The median -0.00377 is quite close to 0, which suggests that the residuals are somewhat symmetrical.

The low and signficant p-values in our model for the intercept and for the coefficient `difflog` suggest that the coefficient is not zero and suggest that there is strong evidence that it explains variance in our model.

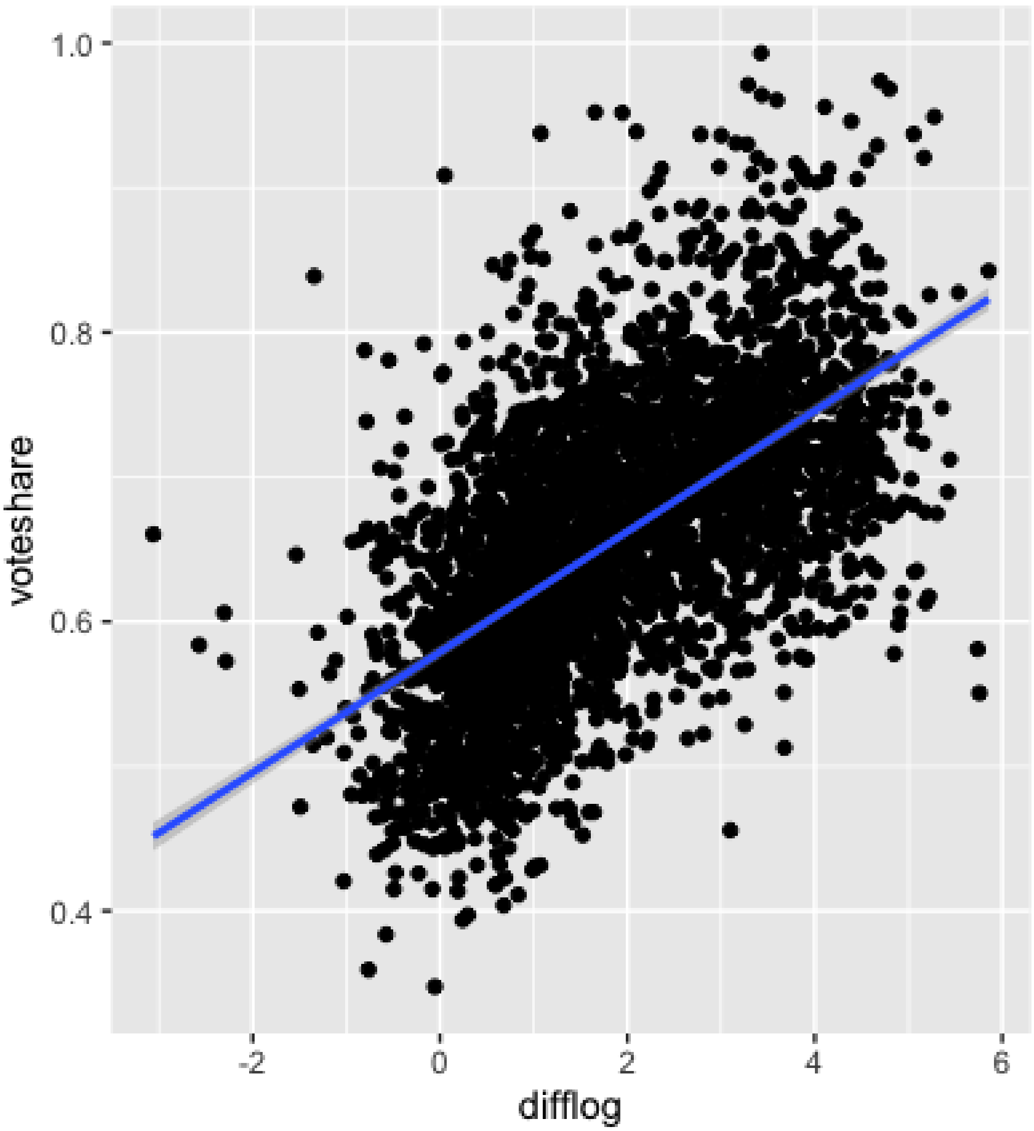
The Multiple R-squared value of this simple linear regression gives us an indication of how much of the outcome variable is explained by the explanatory variable. The Multiple R-squared value of 0.3673 suggests that 36.73 percent of the variation in voteshare is explained by the explanatory variable difflog.

Finally, the low p-value well below 0.05 suggests that there is strong evidence that the coefficient in this model isn't 0, and that the `difflog` variable does have an effect on `voteshare`

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

   We use the ggplot() function to create the scatterplot where the `difflog` is on the x axis and `voteshare` is on the y axis. We use the `geom smooth(method=lm)` function to generate the regression line.

```
ggplot(data, aes(x=difflog, y=voteshare))+
geom_point()+
geom_smooth(method=lm)
```



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

   We save the residuals in a separate variable by running the regression again and assigning the residuals portion of the regression to a variable using `resid`.

```
residualsVoteshareDifflog <- lm(voteshare ~ difflog, data = data)$resid
residualsVoteshareDifflog
```

4. Write the prediction equation.

   The formula for the prediction equation is `y = b0 + b1x`, where b0 is the intercept of the regression and b1, the coefficient of x, is the `difflog` value in the regression.

   As we can see from Table 1, the value for `difflog` (b1) is 0.042 and the value of the intercept (b0) of the regression is 0.579.

   Therefore, the prediction equation is `y = 0.042x + 0.579`.

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`.

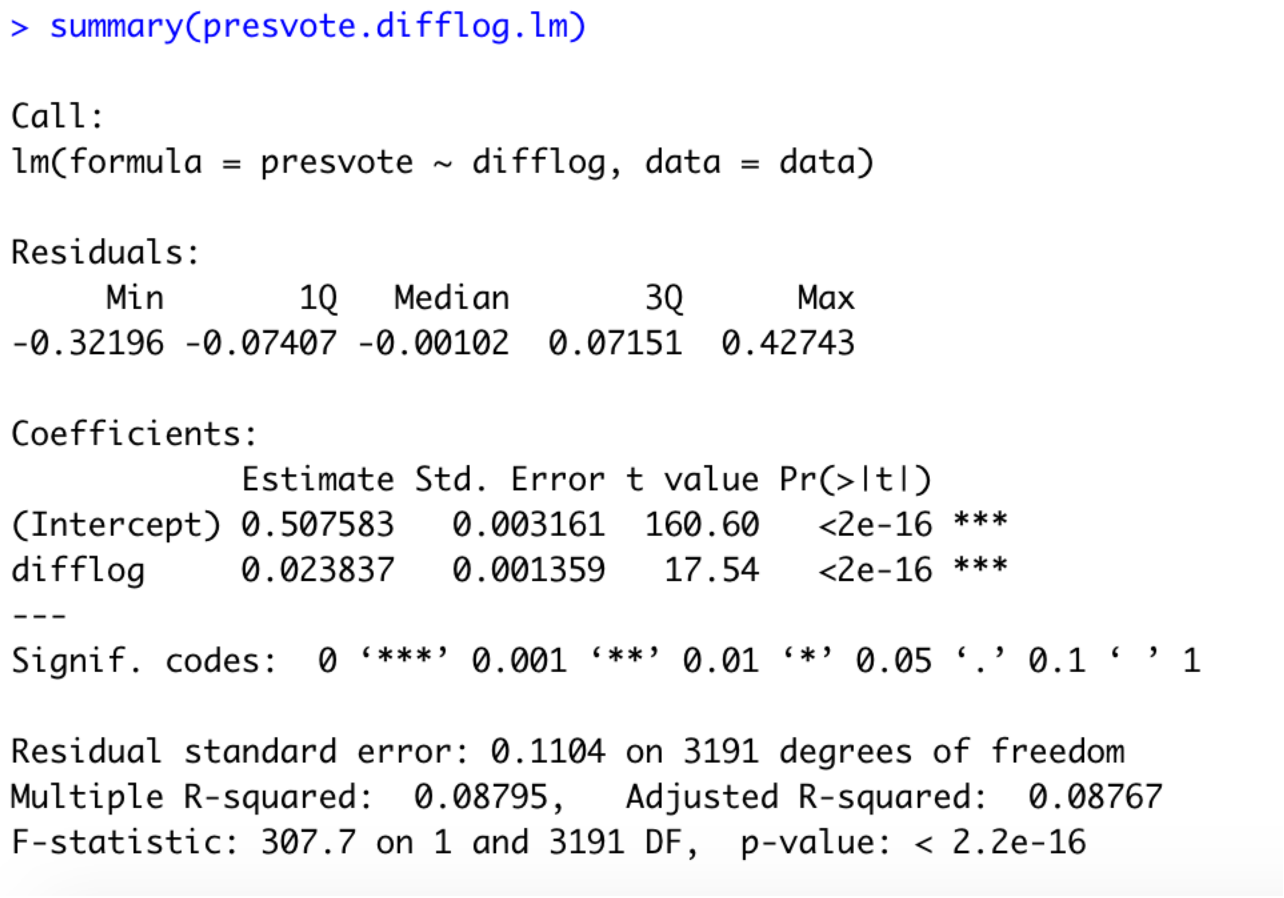
   We run a regression where the outcome variable (Y) is `presvote` and the explanatory variable (x) is `difflog`. We save it in a variable called `presvote.difflog.lm` and use the summary function to view the results. Finally, we generate Latex output using `stargazer`, which generate the below output.

```
presvote.difflog.lm <- lm(presvote ~ difflog, data = data)
summary(presvote.difflog.lm)
stargazer(voteshare.difflog.lm, type='latex', summary= FALSE)
```

Table 2:

	Dependent variable:
	presvote
difflog	0.024^{***}
	(0.001)
Constant	0.508^{***}
	(0.003)
Observations	3,193
R^2	0.088
Adjusted R^2	0.088
Residual Std. Error	0.110 (df = 3191)
F Statistic	307.715^{***} (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

Attached below is a screenshot of the output generated when the summary call is used on the `presvote.difflog.lm` regression.

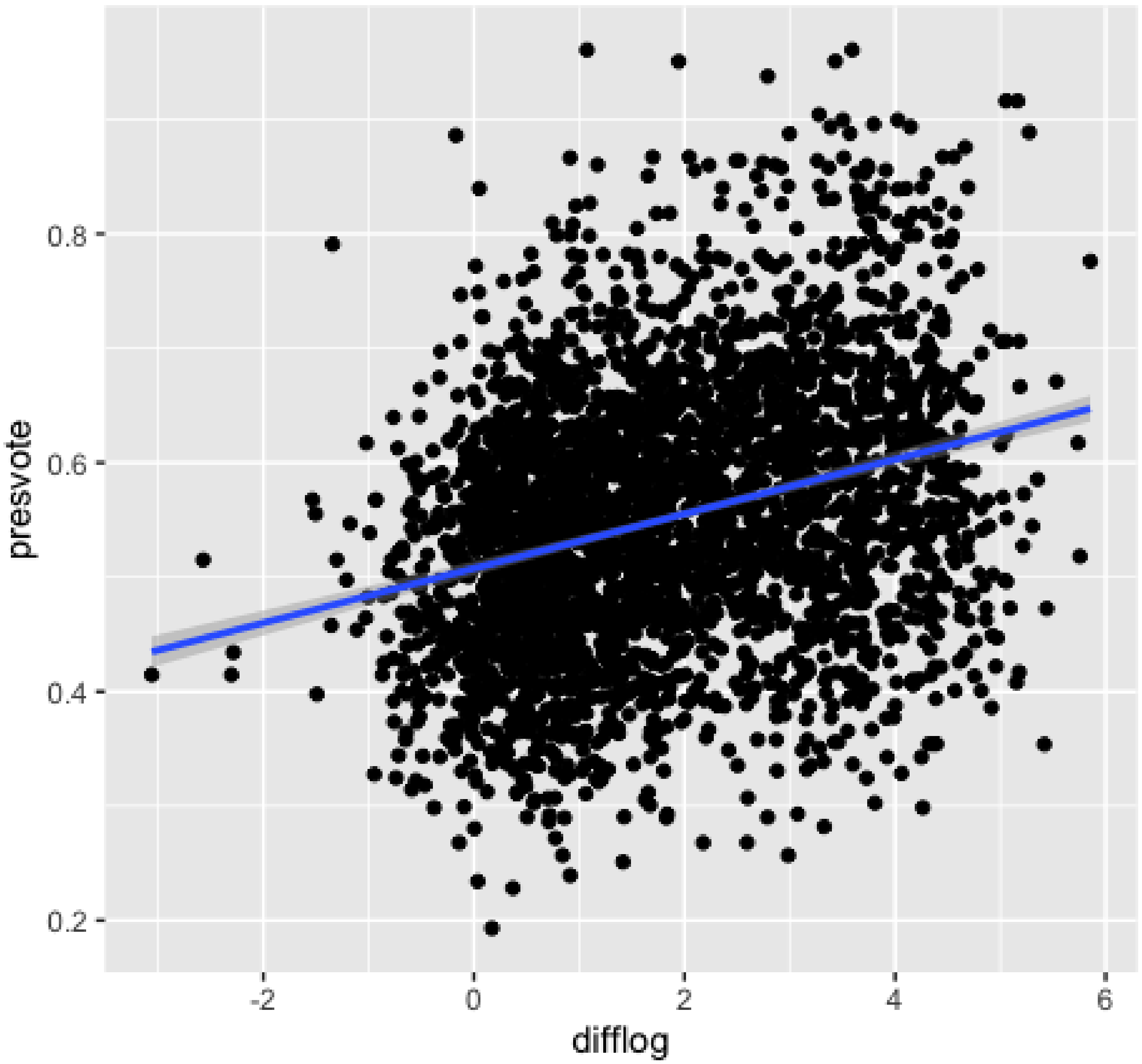


```
> summary(presvote.difflog.lm)

Call:
lm(formula = presvote ~ difflog, data = data)

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.

   We use the ggplot() function to create the scatterplot where `difflog` is on the x axis and `presvote` is on the y axis. We use the `geom smooth(method = lm)` function to generate a regression line

```
ggplot(data, aes(x=difflog, y=presvote))+
geom_point()+
geom_smooth(method=lm)
```



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

   We save the residuals of the regression into a separate variable by selecting the `resid` portion of the regression and assigning it to a new variable.

   ```
   residualsPresvoteDifflog <- lm(presvote ~ difflog, data = data)$resid
   residualsPresvoteDifflog
   ```

4. Write the prediction equation.

   As in Question 1, the prediction equation is given by `y = b0 + b1x`, where b0 is the intercept of the regression and b1, the coefficient of x, is the `difflog` value in the regression.

   As we can see in Table 2, the value for `difflog` (b1) is 0.024 and the value of the intercept (b0) is 0.508.

   Therefore the prediction equation is:

   `y = 0.024x + 0.508`.

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`.

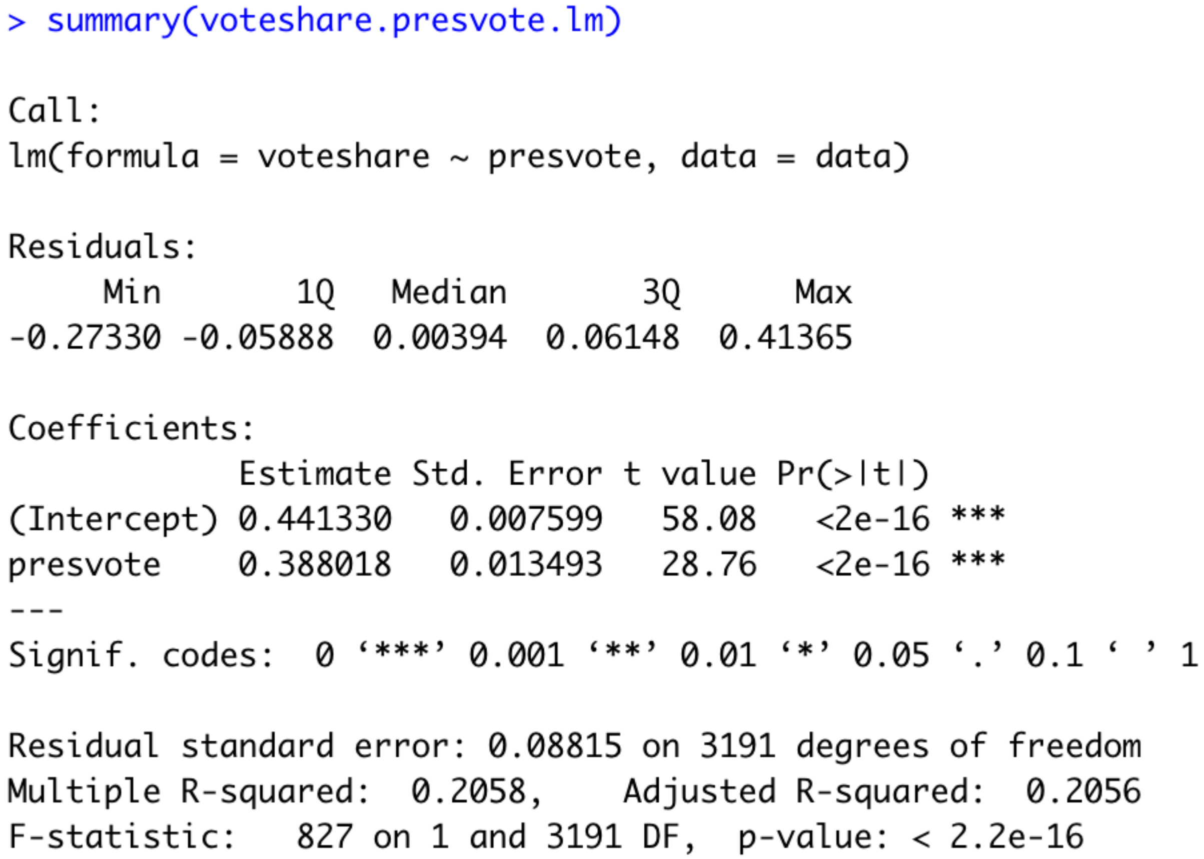
   We run a regression where the outcome variable (y) is `voteshare` and the explanatory variable (x) is `presvote`. We save it in a variable called `voteshare.presvote.lm` and use the summary function to view the results. Finally, using the `stargazer` function we generate the following Latex output.

```
voteshare.presvote.lm <- lm(voteshare ~ presvote, data = data)
summary(voteshare.presvote.lm)
stargazer(voteshare.presvote.lm, type='latex', summary= FALSE)
```

Table 3:

	Dependent variable:
	voteshare
presvote	0.388***
	(0.013)
Constant	0.441***
	(0.008)
Observations	3,193
R^2	0.206
Adjusted R^2	0.206
Residual Std. Error	0.088 (df = 3191)
F Statistic	826.950*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

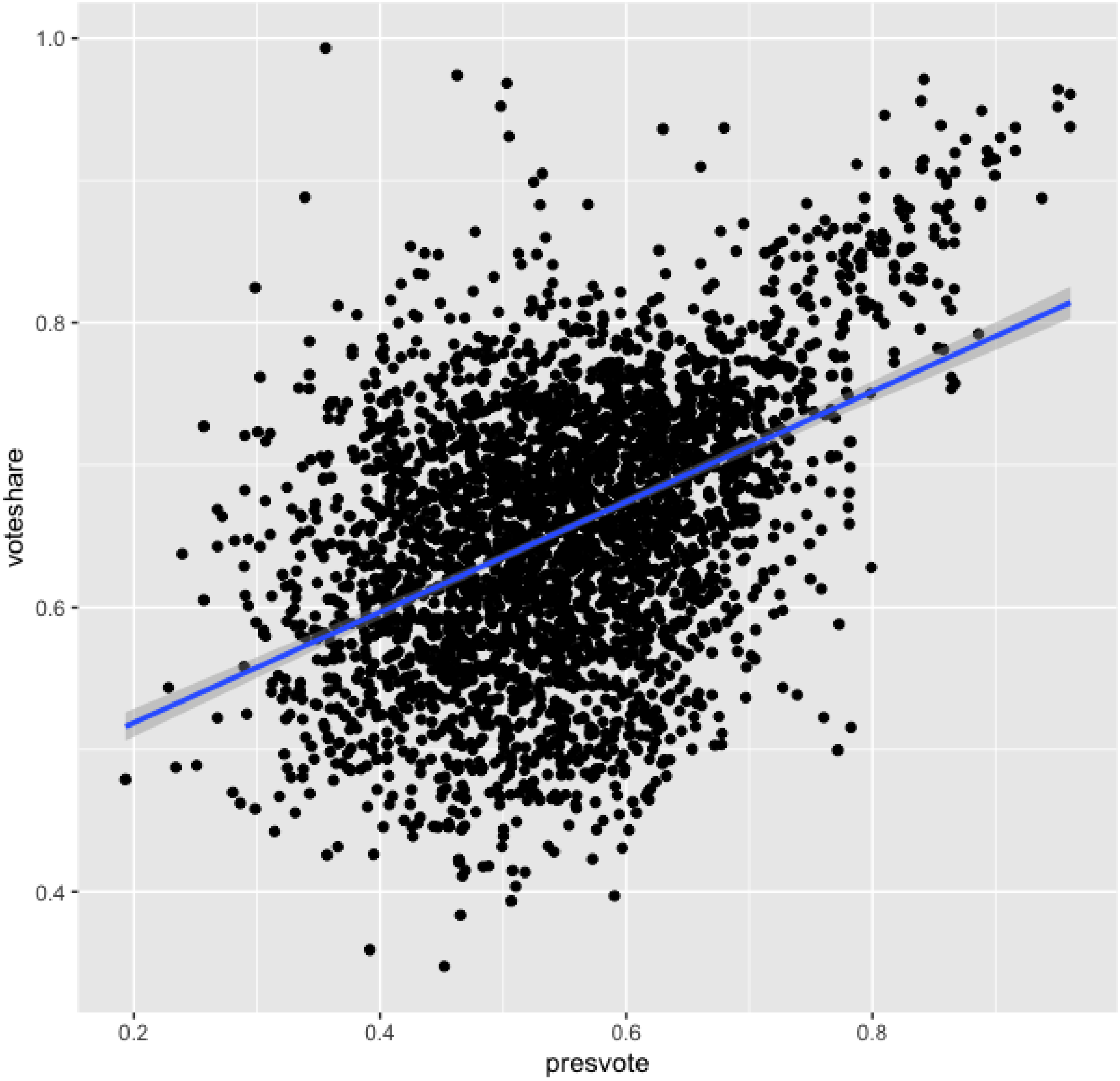
Attached below is a screenshot of the output generated in the R Studio console when `summary()` is called on the `voteshare.presvote.lm` regression.



```
> summary(voteshare.presvote.lm)

Call:
lm(formula = voteshare ~ presvote, data = data)

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. We use the `ggplot()` function to create a scatterplot where `voteshare` is on the y axis and `presvote` is on the x axis.

```
ggplot(data, aes(x=presvote, y=voteshare))+
geom_point()+
geom_smooth(method=lm)
```



3. Write the prediction equation.

   As in Questions 1 and 2, the prediction equation is given by `y = b0 + b1x`, where b0 is the intercept of the regression and b1, the coefficient of x, is the `difflog` value in the regression.

   As we can see in Table 3, the value for `presvote` (b1) is 0.388 and the value of the intercept (b0) is 0.441.

   Therefore, the prediction equation is:

   `y = 0.388x + 0.441`.

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.

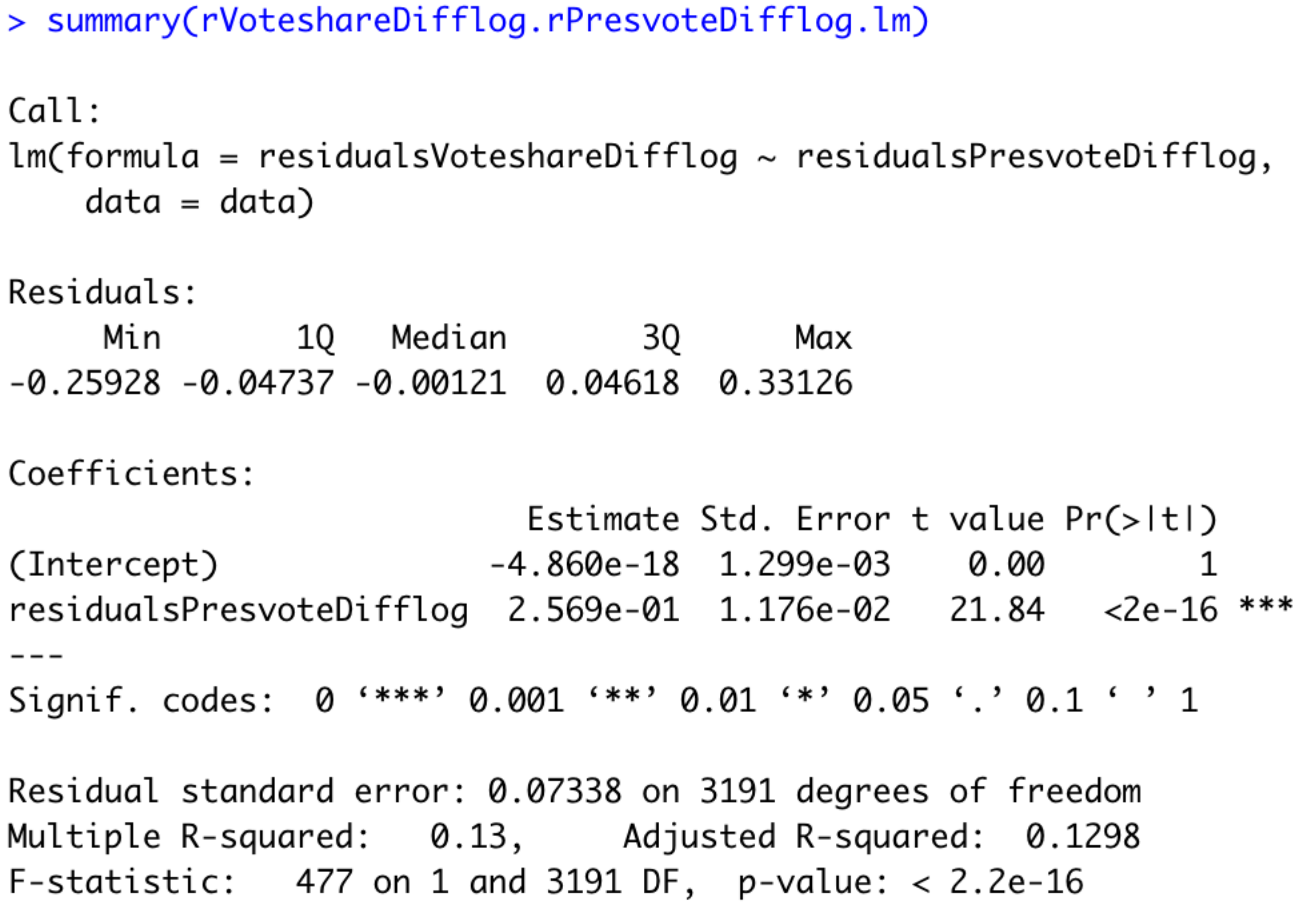
   We run a regression where the outcome variable (y) is the residuals from Question 1 (`residualsVoteshareDifflog`) and the explanatory variable (x) is the residuals from Question 2 (`residualsPresvoteDifflog`).
   We save it in a variable called `rVoteshareDifflog.rPresvoteDifflog.lm`. Finally, we use the stargazer function to generate the following Latex output.

```
rVoteshareDifflog.rPresvoteDifflog.lm <- lm(residualsVoteshareDifflog ~
    residualsPresvoteDifflog, data = data)
summary(rVoteshareDifflog.rPresvoteDifflog.lm)
stargazer(rVoteshareDifflog.rPresvoteDifflog.lm, type='latex', summary=
    FALSE)
```

Table 4:

	Dependent variable:
	residualsVoteshareDifflog
residualsPresvoteDifflog	0.257***
	(0.012)
Constant	−0.000
	(0.001)
Observations	3,193
R^2	0.130
Adjusted R^2	0.130
Residual Std. Error	0.073 (df = 3191)
F Statistic	476.975*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

Below is a screenshot of the output from the console in R Studio of the regression model.



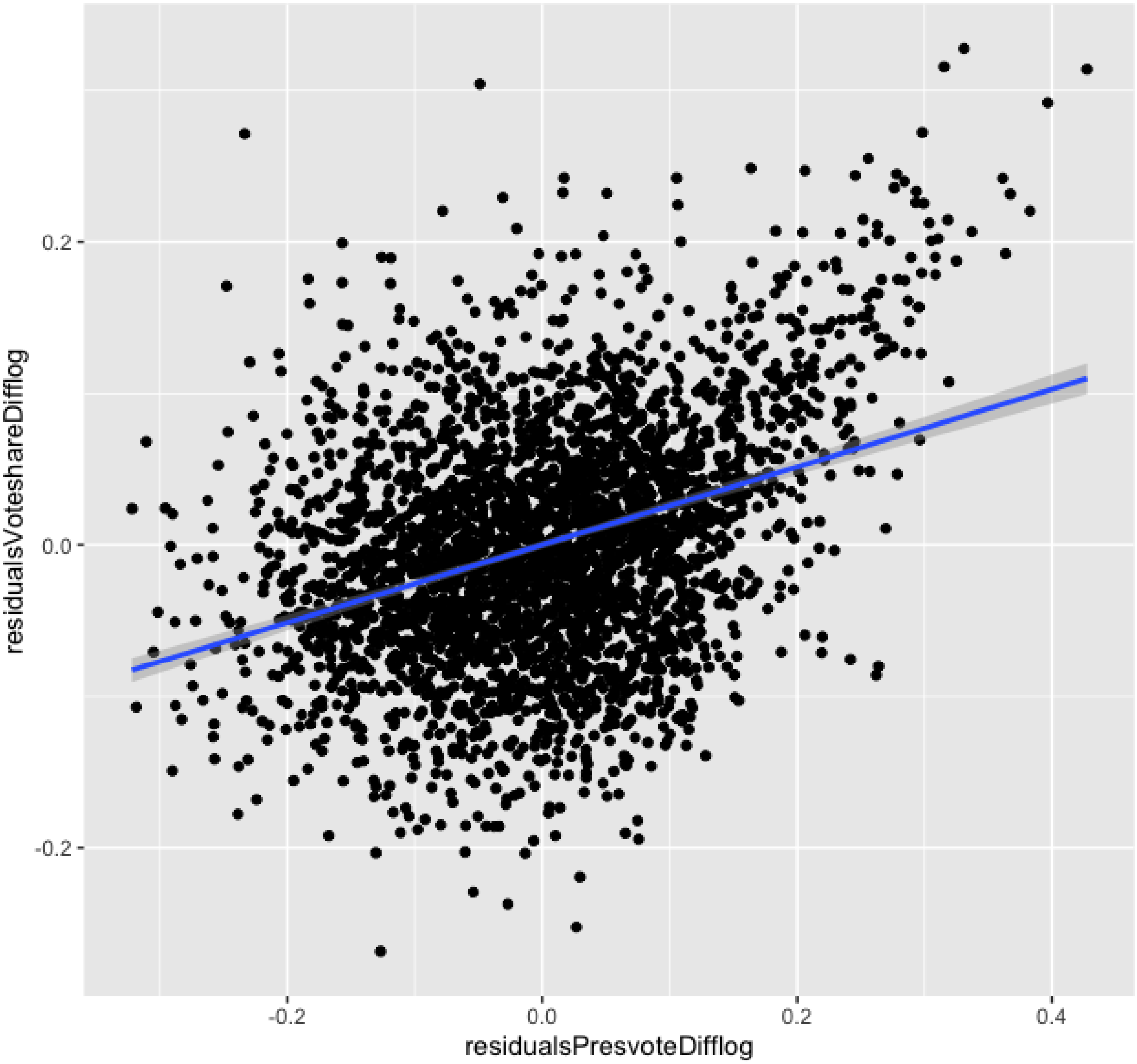
```
> summary(rVoteshareDifflog.rPresvoteDifflog.lm)

Call:
lm(formula = residualsVoteshareDifflog ~ residualsPresvoteDifflog, 
    data = data)

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    
residualsPresvoteDifflog  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
```

The residuals of the regression are the difference between the observed and predicted values of the regression. We can see a min residual value of -0.25928, and a max residual value of 0.33126. The median residual value is -0.00121, which is quite near 0.

The p-value for the intercept at 1 is extremely high. This suggests that it is very unlikely to be signficant to say the intercept value is 0.

The Multiple R squared value is 0.13, which suggests that around 13 percent of the variance of `ResidualsVoteshareDifflog` is explained by `ResidualsPresvoteDifflog`.

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



3. Write the prediction equation.

   As in the previous questions, the prediction equation is given by `y = b0 + b1x`, where b0 is the intercept of the regression and b1, the coefficient of x, is the `residualsVoteshareDifflog` value in the regression.

   As we can see in Table 4, the value for `residualsPresvoteDifflog` (b1) is 0.257 and the value of the intercept (b0) is -0.000, or 0.

   Therefore, the prediction equation is:

   ```
   y = 0.257x - 0.000
   ```

   Or, more simply:

   ```
   y = 0.257x
   ```

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`.

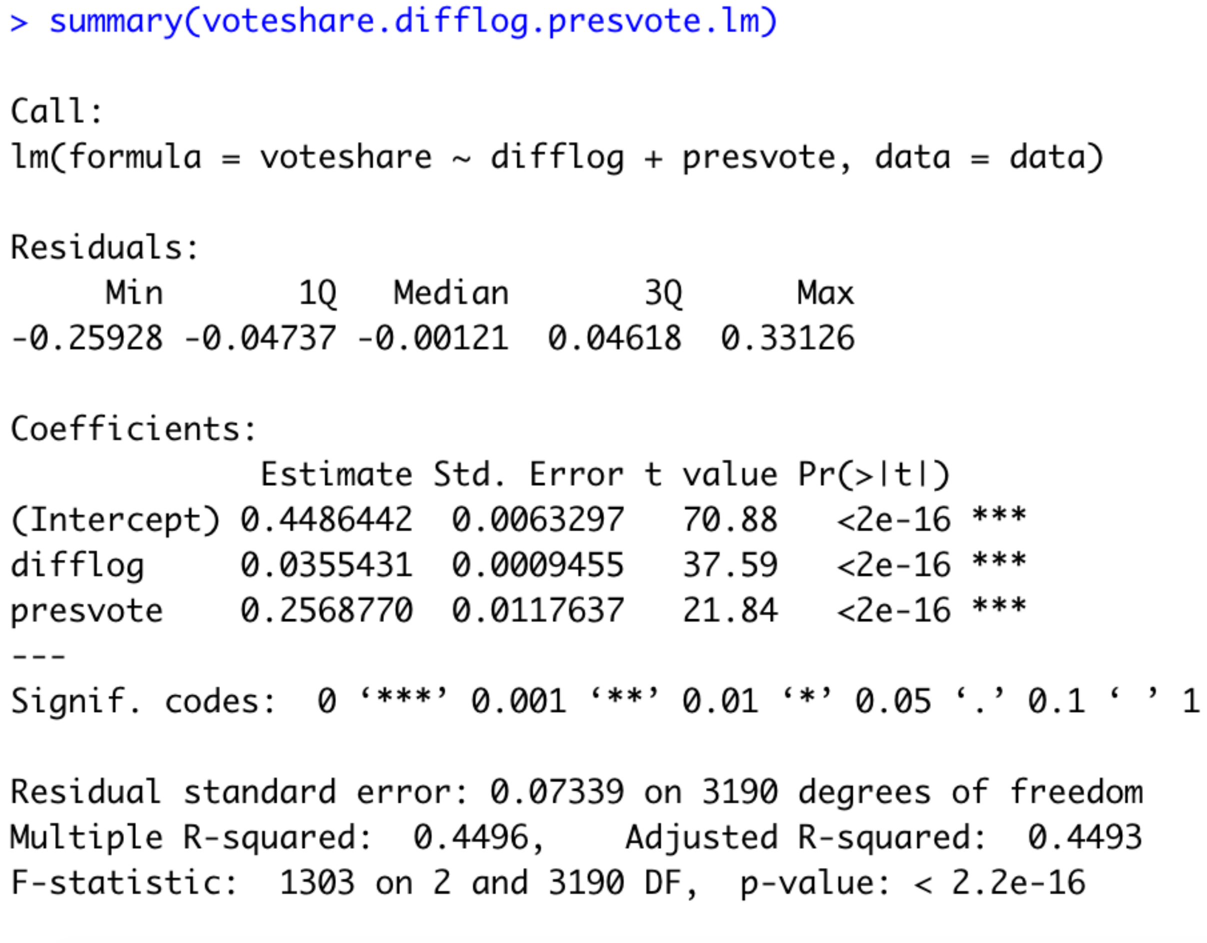
   We run a regression where the outcome variable (y) is the incumbent's `voteshare` and the explantory variables (x) are `difflog` and `presvote`. We save it a variable called `voteshare.difflog.presvote.lm`. Finally, we use the `stargazer` function to yield Latex output as follows:

```
voteshare.difflog.presvote.lm <- lm(voteshare ~ difflog + presvote, data
    = data)
summary(voteshare.difflog.presvote.lm)
stargazer(voteshare.difflog.presvote.lm, type='latex', summary= FALSE)
```

Table 5:

	Dependent variable:
	voteshare
difflog	0.036***
	(0.001)
presvote	0.257***
	(0.012)
Constant	0.449***
	(0.006)
Observations	3,193
R^2	0.450
Adjusted R^2	0.449
Residual Std. Error	0.073 (df = 3190)
F Statistic	1,302.947*** (df = 2; 3190)
Note:	*p<0.1; **p<0.05; ***p<0.01

Attached is a screenshot of the output from the R studio console which summarises the regression model.



```
> summary(voteshare.difflog.presvote.lm)

Call:
lm(formula = voteshare ~ difflog + presvote, data = data)

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.

Since we have two explanatory variables, difflog and presvote, we are running a multiple linear regression as opposed to a simple linear regression in the previous 4 questions. As such, we use a different prediction equation to account for the extra explanatory variable.

We use the following equation:

```
y = b0 + b1x1 + b2x2
```

Where y is the outcome variable voteshare, b0 is the intercept of model or the constant, b1 is the coefficient of the first explanatory variable difflog, x1 is the value of the first explanatory variable difflog, b2 is the coefficient of the second explanatory variable presvote and x2 is the value of the second explantory variable presvote.

In our regression model, as seen from Table 5, the coefficient of difflog is 0.036, the coefficient of presvote is 0.257 and the constant/intercept is 0.449.

Therefore, our prediction equation for this model is:

```
y = (0.036)x1 + (0.257)x2 + 0.449
```

Or:

```
voteshare = (0.036)difflog + (0.257)presvote + 0.449
```

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

We can see that the coefficient estimate, standard error, t-value and p-value of the `presvote` variable in Question 5 are identical to the coefficient estimate, standard error, t-value and p-value of the Question 4 regression of residuals of `presvote` and the residuals of `difflog`.

The regression in Question 4 plots the residuals of `Voteshare` against `Difflog` against the residuals of `Presvote` against `Difflog`. This regression plots the unexplained variation in `voteshare.difflog.lm` against the unexplained variation in `presvote.difflog.lm` to see if there is a relationship between the two unexplained variations. The plot in Question 4 measures the amount of variation `voteshare` that is explained by presvote without any influence from the `difflog` variable. It shows, essentially, how much voteshare is influence by `presvote` on it's own, without `difflog`.

This is the same thing that the regression in Question 5 shows. The regression separates the two variables and shows how much variance in `voteshare` is explained by `difflog` on it's own, namely a 0.036 increase in `voteshare` for a one unit increase in `difflog`.

As `difflog` and `presvote` are separated in the multiple linear regression of Question 5, the coefficient of `presvote` in Question 5 therefore shows how much variation is explained by `presvote` without any influence from `difflog`. This means that a 1 unit increase in `presvote` on its own will result in a 0.257 increase in `voteshare`.

The min, Q1, median, Q3 and max values of the residuals of Question 4 are also identical to the min, Q1, median, Q3 and max values of the residuals of Question 5, for the same reasons.