

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`.

```
1 incumbents_subset <- read_csv("GitHub/QTM200Spring2020/problem_sets/PS3/  
  incumbents_subset.csv")  
2  
3 fitmodel <- lm(voteshare ~ difflog, data=incumbents_subset)  
4 summary(fitmodel)  
5  
6 plot(incumbents_subset$voteshare ~ incumbents_subset$difflog)
```

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

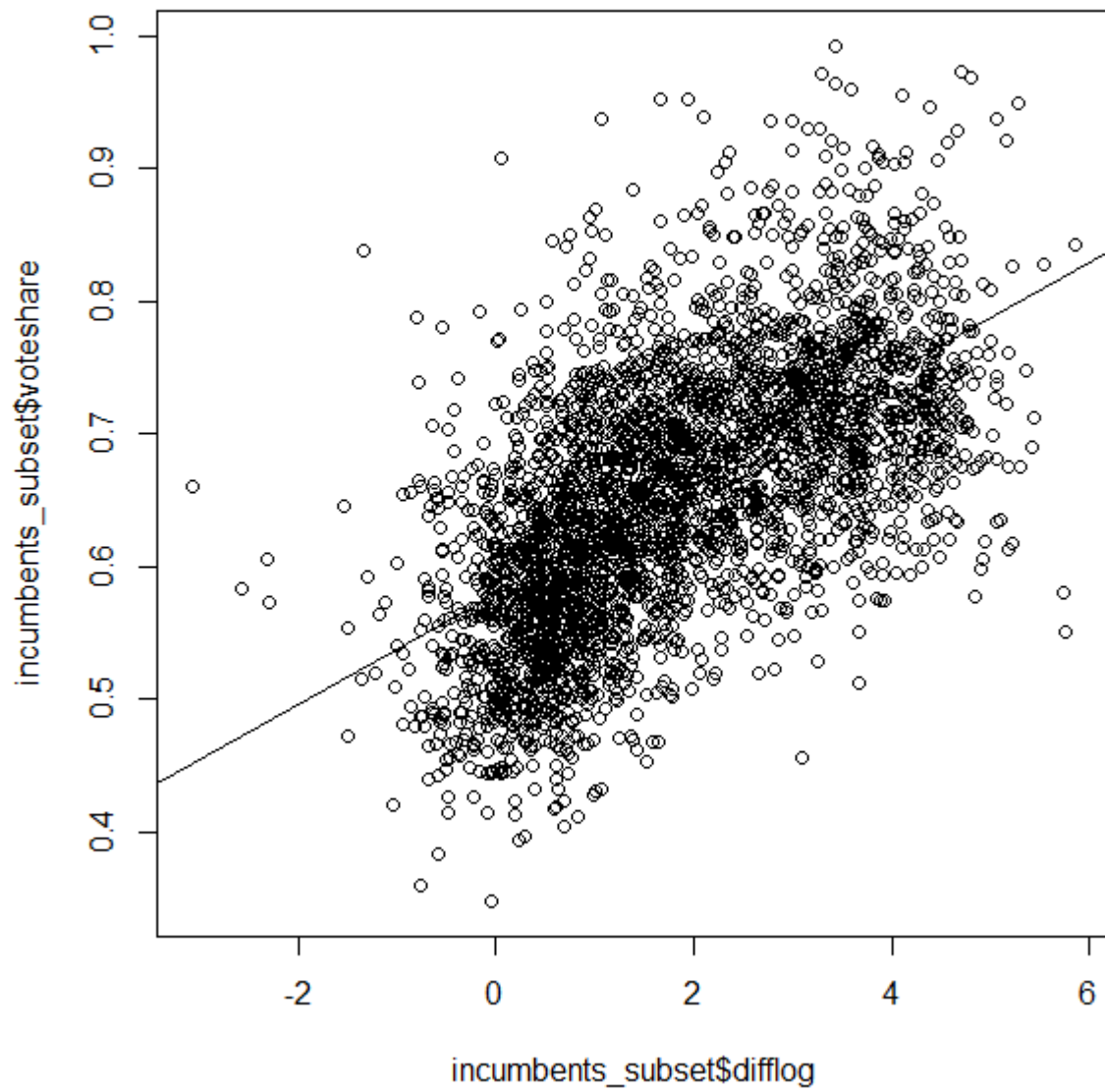
Residual standard error: 0.07867 on 3191 degrees of freedom

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.

```
1 abline(fitmodel)
2
3 incumbents_subset$predicted <- predict(fitmodel)
```

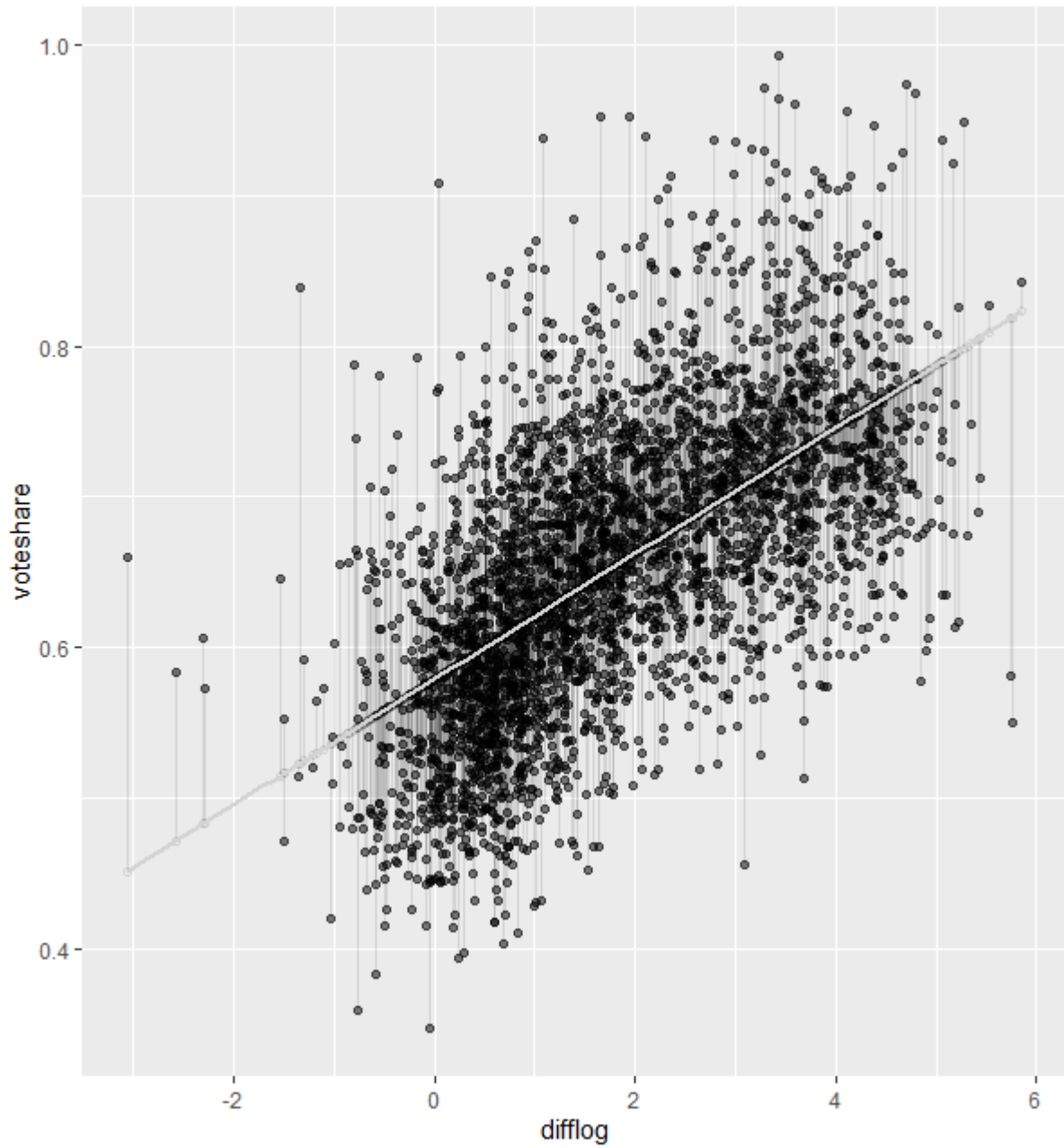


3. Save the residuals of the model in a separate object.

```

1
2 plot_residuals <- ggplot(incumbents_subset, aes(x=difflog, y=votesshare))+
3   geom_point(alpha = I(0.5))+
4   geom_point(aes(y= predicted), shape =1, alpha = I(0.1))+
5   geom_segment(aes(xend = difflog, yend = predicted), alpha=I(0.1))+
6   geom_smooth(method="lm", se=F, color = "lightgrey")
7 plot_residuals
8
9 # Question 2
10 fitmodel2 <- lm(presvote ~ difflog, data = incumbents_subset)
11 summary(fitmodel2)

```



4. Write the prediction equation.

$$y = 0.04167x + 0.579$$

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`.

```
1 fitmodel2 <- lm(presvote ~ difflog, data = incumbents_subset)
2 summary(fitmodel2)
```

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

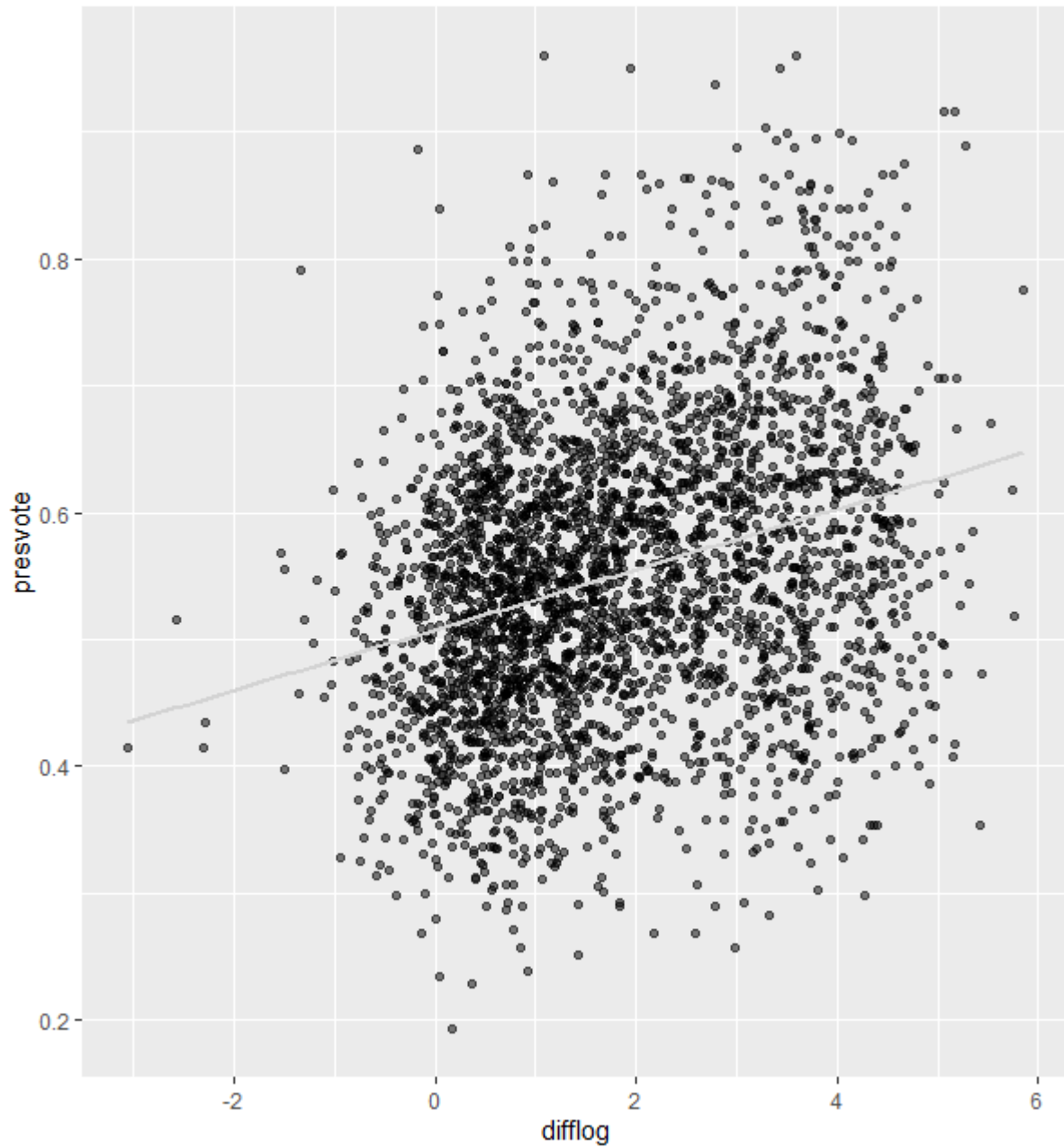
Residual standard error: 0.1104 on 3191 degrees of freedom

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.

```
1 plot2 <- ggplot(incumbents_subset, aes(x=difflog, y=presvote)) +
2   geom_point(alpha=I(0.5)) +
3   geom_smooth(method = "lm", se = F, color = "lightgrey")
4 plot2
```



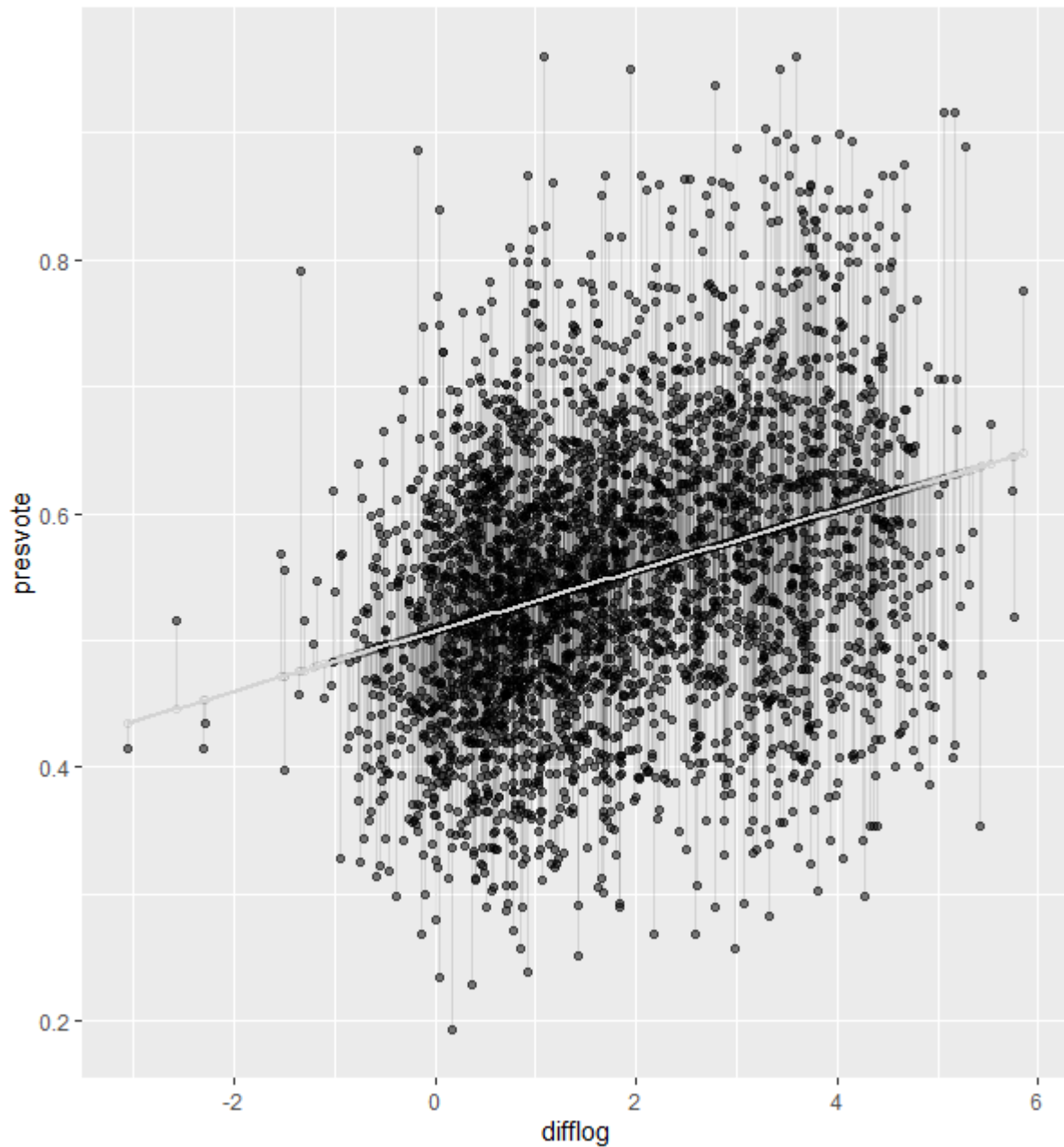
3. Save the residuals of the model in a separate object.

```
1 incumbents_subset$prediction <- predict(fitmodel2)
2 incumbents_subset$residuals2 <- residuals(fitmodel2)
3
```

```

4 plot2_residuals <-ggplot(incumbents_subset, aes(x=difflog, y=presvote))+
5   geom_point(alpha = I(0.5))+
6   geom_point(aes(y= prediction), shape =1, alpha = I(0.1))+
7   geom_segment(aes(xend = difflog, yend = prediction), alpha=I(0.1))+
8   geom_smooth(method="lm", se=F, color = "lightgrey")
9 plot2_residuals

```



4. Write the prediction equation.

$$y = 0.0238x + 0.508$$

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**.

```
1 fitmodel3 <- lm(voteshare ~ presvote, data=incumbents_subset)
2 summary(fitmodel3)
```

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

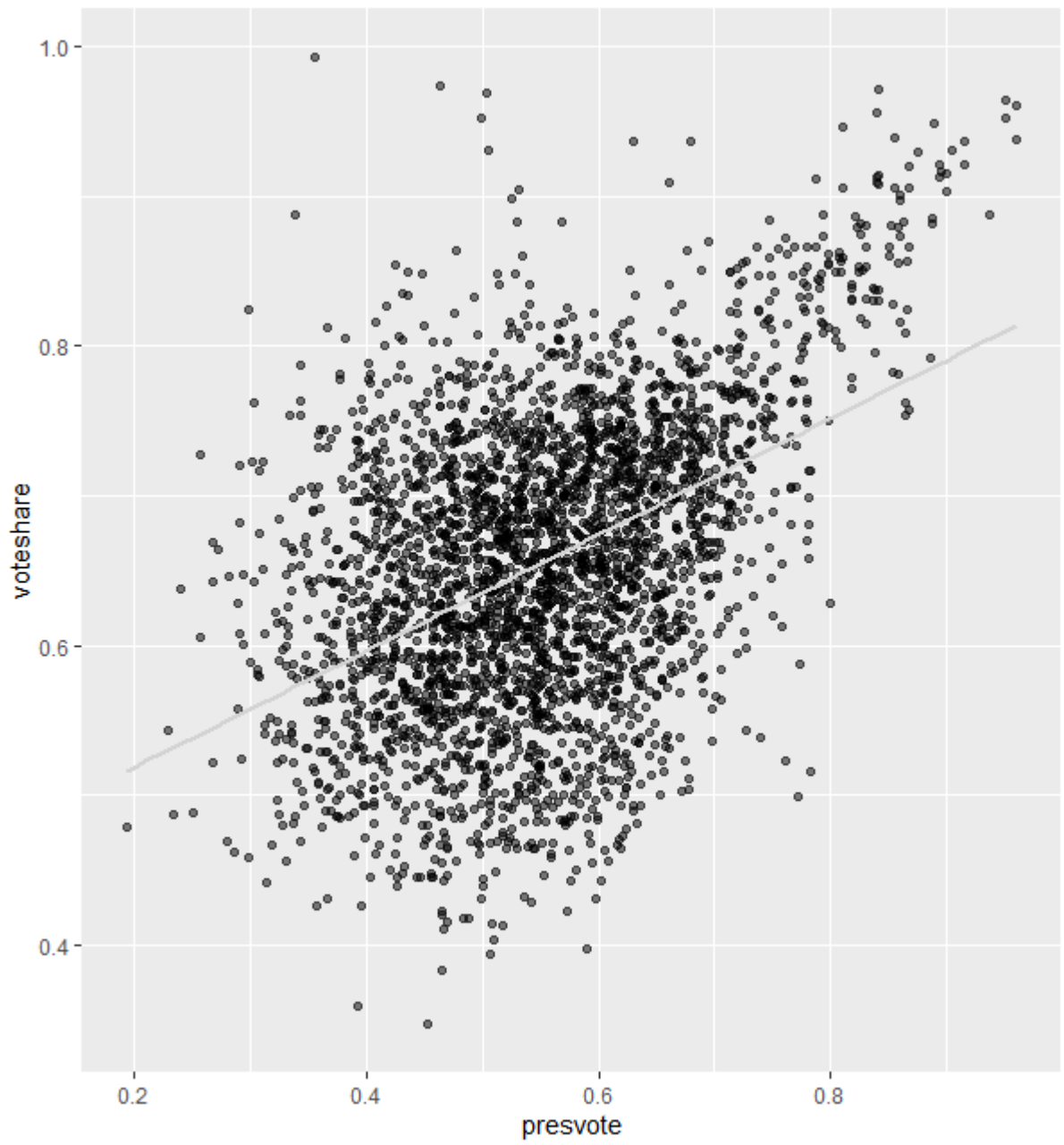
Residual standard error: 0.08815 on 3191 degrees of freedom

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.

```
1 plot3<- ggplot(incumbents_subset, aes(x=presvote, y=voteshare))+
2   geom_point(alpha=I(0.5))+
3   geom_smooth(method = "lm", se=F, color = "lightgrey")
4 plot3
```



3. Write the prediction equation.

$$y = 0.388x + 0.441$$

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.

```
1 fitmodel4 <- lm(residuals1 ~ residuals2 , data = incumbents_subset)
2 summary(fitmodel4)
```

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
residuals2	2.569e-01	1.176e-02	21.84	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

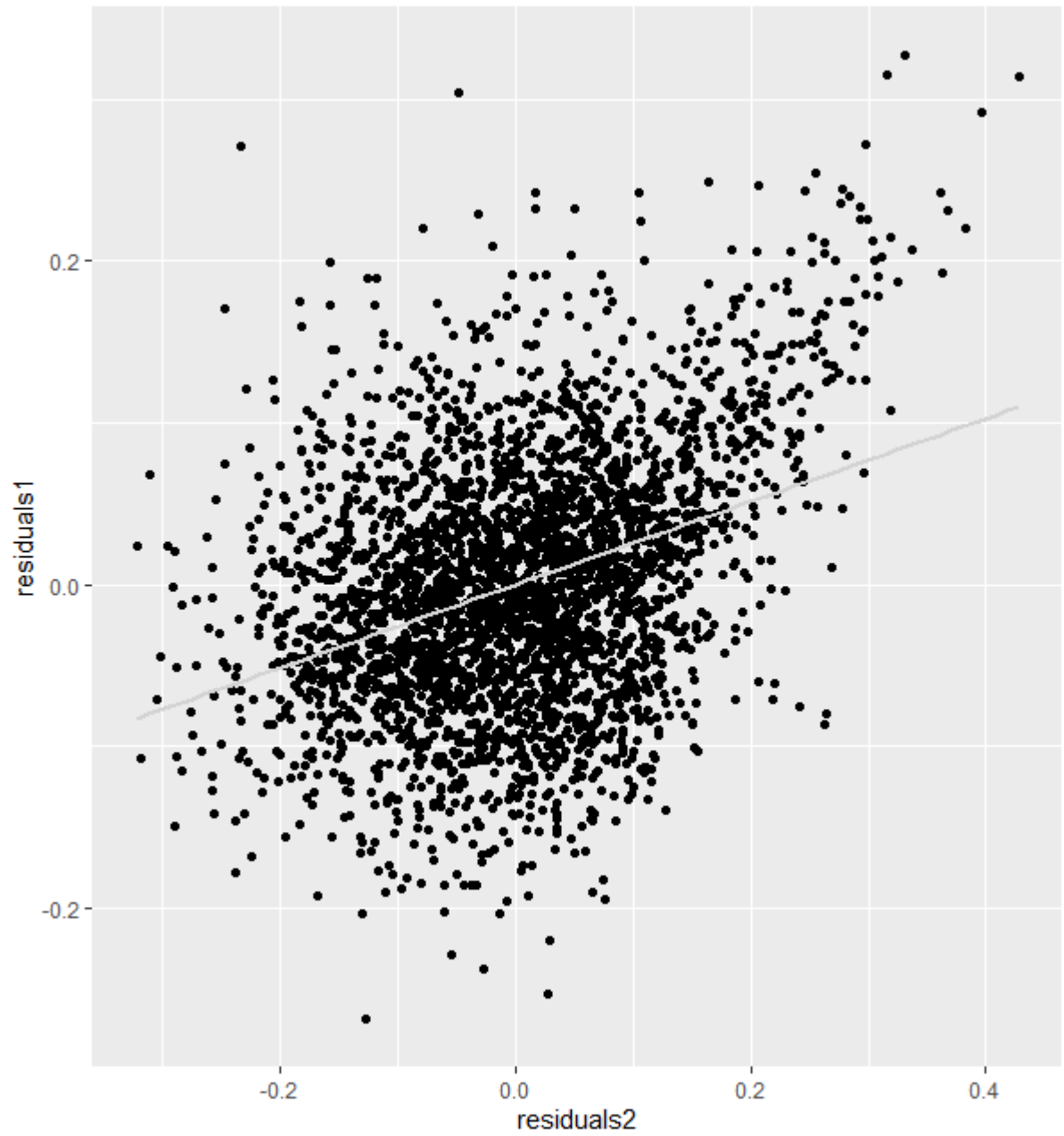
Residual standard error: 0.07338 on 3191 degrees of freedom

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.

```
1 plot4 <- ggplot(incumbents_subset , aes(x=residuals2 , y=residuals1))+
2   geom_point()+
3   geom_smooth(method = "lm" , se=F, color = "lightgrey")
4 plot4
```



3. Write the prediction equation.

$$y = -4.86 * 10^{-18}x + 0.257$$

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`.

```
1 multiplefit <- lm(voteshare ~ difflog + presvote, data = incumbents_subset)
2 summary(multiplefit)
```

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.

$$y = 0.0355x_1 + 0.2569x_2 + 0.4486$$

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