Problem Set 3

Applied Stats/Quant Methods 1

Due: November 12, 2021

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 in `.pdf` form.
- This problem set is due before class on Friday November 12, 2021. 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`.

   The Y intercept is 0.579, meaning that when difflog (the x value) is at a value of 0, the voteshare is 0.579.

   In this case, the coefficient estimate for difflog is 0.0416, meaning that for every 1 unit increase in difflog, there is a 0.0416 unit increase in voteshare. Below on the next page you can see the regression I ran to obtain this information.

```
reg1 <- lm(data = incumbents, voteshare ~ difflog)

```

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

   Here is the code I used to make my scatterplot:

```
ggplot(incumbents, aes(x=difflog, y=voteshare )) +
geom_point(alpha = 0.5) +
geom_smooth(method = lm, formula = y~x) +
ggtitle("Relationship Between Voteshare and Difflog") +
labs(y="Voteshare") +
labs(x="Difflog")

```

   Here is my Scatterplot:

Figure 1:

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

```
resids1 <- resid(reg1)

```

4. Write the prediction equation.

   The prediction equation follows the formula y = mx + c.
   Given the coefficients I obtained above in part 1 by running regression 1, my prediction equation is therefore:

```
predict_1 <- 0.04167*difflog + 0.57903
```

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`.

```
reg2 <- lm(data = incumbents, presvote ~ difflog)

```

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

   My Scatterplot code was as follows:

```
ggplot(incumbents, aes(x=difflog, y=presvote)) +
geom_point(alpha = 0.5) +
geom_smooth(method = "lm", formula = y~x) +
ggtitle("Relationship Between Presvote and Difflog") +
labs(y="Presvote") +
labs(x="Difflog")

```

Here is my scatterplot:

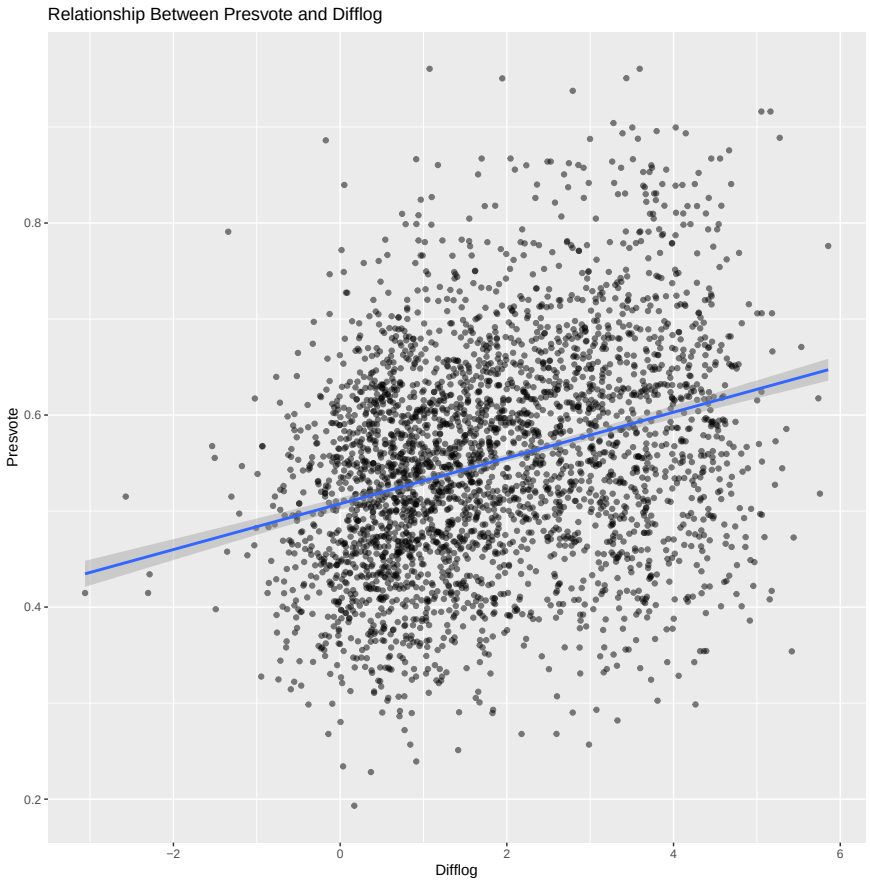


Figure 2:

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

```
resids2 <- resid(reg2)

```

4. Write the prediction equation.

   Prediction equation:
   y= mx + c

   My values for this prediction equation have been taken from the coefficients I received as an output when I printed my "reg2" regression in R.

```
predict_2 <- 0.02384*difflog + 0.50758

```

   In this case, the coefficient estimate for difflog is 0.02384, meaning that for every 1 unit increase in difflog, there is a 0.02384 unit increase in presvote. 0.50758 is the y-intercept, the value of y (the outcome variable) when the x value is equal to zero.

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`.

```
reg3 <- lm(data = incumbents, voteshare ~ presvote)

```

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

   Here is the code I created my scatterplot with:

```
ggplot(incumbents, aes(x=presvote, y=voteshare)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y~x) +
  ggtitle("Relationship Between Presvote and Voteshare") +
  labs(y="Voteshare") +
  labs(x="Presvote")

```

   Here is my scatterplot:

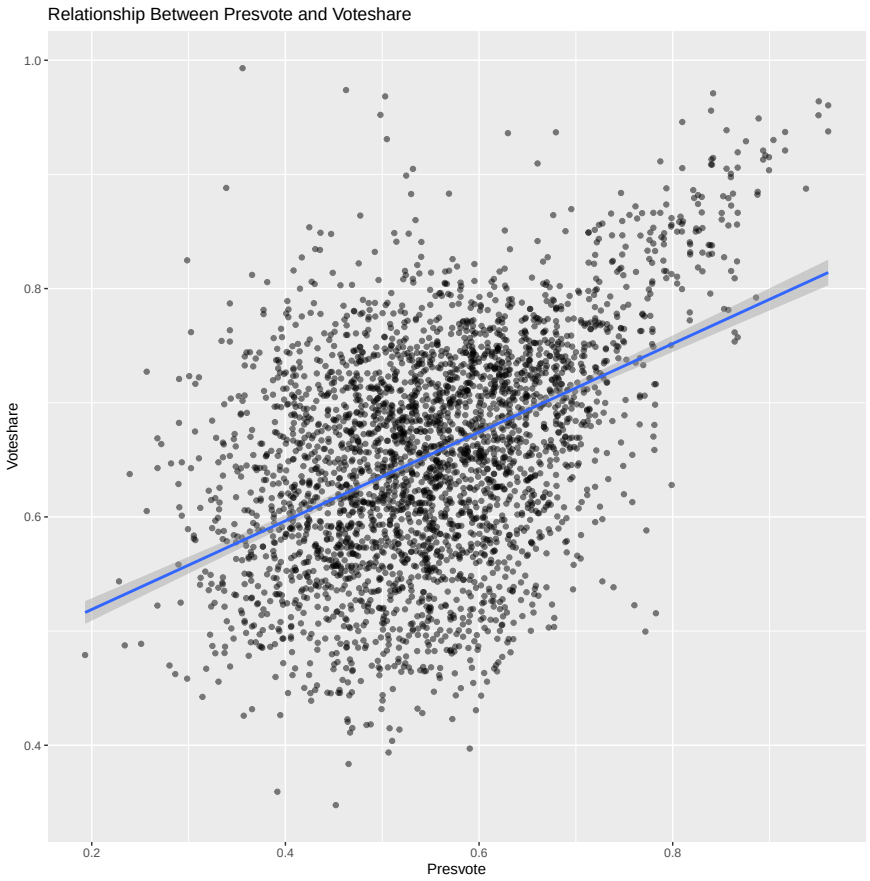


Figure 3:

3. Write the prediction equation.
   prediction equation:
   y = mx + c

   My values for this prediction equation have been taken from the coefficients I received as an output when I printed my "reg3" regression in R.

```
predict_3 <- 0.3880*presvote + 0.4413

```

   In this case, the coefficient estimate for presvote is 0.3880, meaning that for every 1 unit increase in presvote, there is a 0.3880 unit increase in voteshare. 0.4413 is the y-intercept, the value of y (the outcome variable) when the x value is equal to zero.

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.

```
reg4 <- lm(data = incumbents, resids1 ~ resids2)

```

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

   The code for my scatterplot was as follows:

```
ggplot(incumbents, aes(x=resids2, y=resids1)) +
geom_point(alpha = 0.5) + #add a scatterplot
geom_smooth(method = "lm", formula = y~x) +
ggtitle("Relationship Between Resids1 and Resids2") +
labs(y="Resids1") +
labs(x="Resids2")

```

   My scatterplot can be seen at the end of Question 4.

3. Write the prediction equation.

The prediction equation I wrote following the formula of y = mx + c using the coefficients obtained through inputting the function print(reg4) was as follows:

```
predict_4 <- 0.2569e-01*resids2 + -4.860e-18

```

In this case, the coefficient estimate for resids2 is 0.2569e-01, meaning that for every 1 unit increase in resids2, there is a 0.2569e-01 unit increase in resids1. -4.860e-18 is the y-intercept, the value of y (the outcome variable) when the x value is equal to zero.

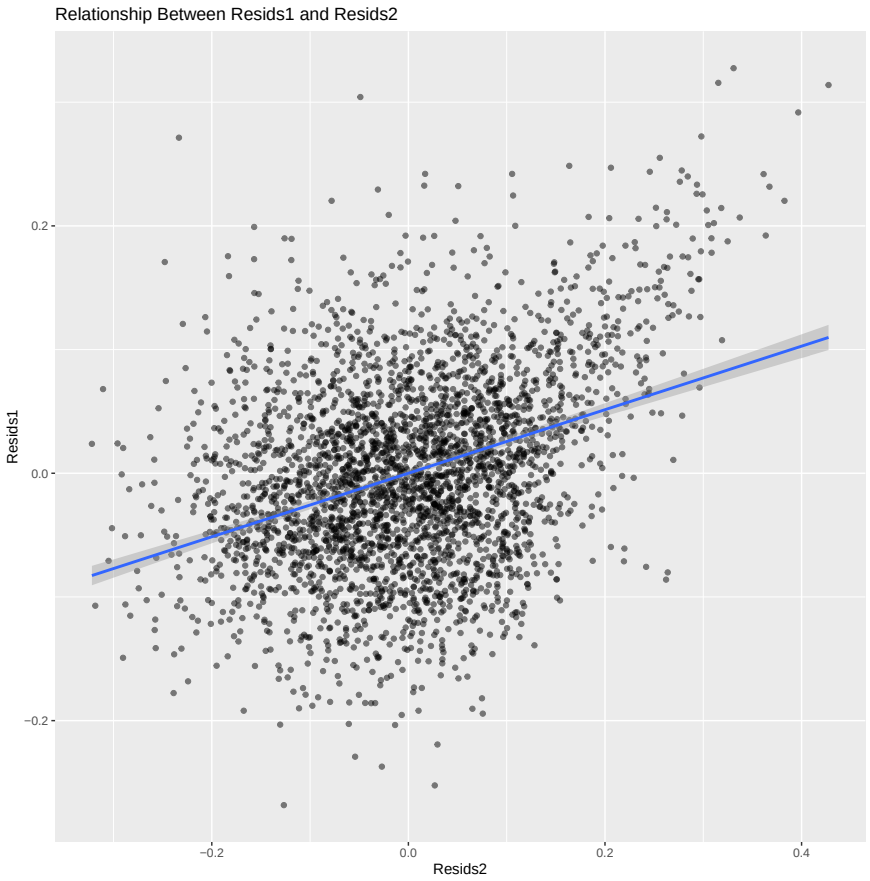


Figure 4:

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`.

   The regression I ran was as follows:

```
reg5 <- lm(voteshare ~ difflog + presvote, data = incumbents)

```

2. Write the prediction equation.

   The prediction equation I wrote this time followed the formula y = mx1 + mx2 + c, because this was a multivariate regression.I used the coefficients outputted by R after I ran the function print(reg5).

```
predict_5 <- 0.03554*difflog + 0.25688*presvote + 0.44864

```

   In this case where there are 2 x-values, the coefficient estimate for difflog is 0.3554, meaning that for every 1 unit increase in difflog, there is a 0.3554 unit increase in voteshare. Meanwhile, the coefficient estimate for presvote is 0.25688, meaning that for every 1 unit increase in presvote, there is a 0.25688 unit increase in voteshare. 0.44864 is the y-intercept, the value of y (the outcome variable) when the x-values are equal to zero.

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

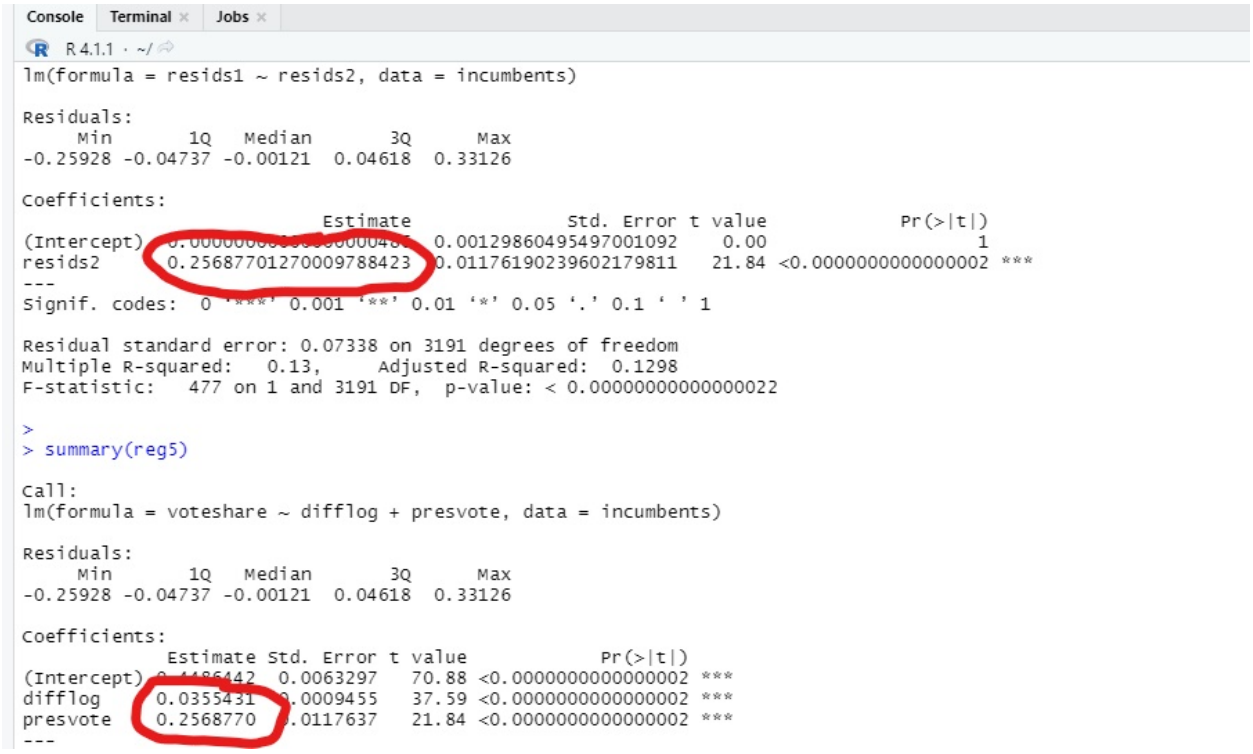
   When I ran the code summary(reg4) and summary(reg5) and compared the outputs, I found that the "Estimate" values of variance for resids2 in the output for regression 4 and presvote in the output for regression 5 were the same value, 0.2568770.

   As far as my understanding goes, this means that regression 4 and regression 5 are serving similar functions and explaining the same amount of variance in the same way, and the variance that is not explained by those regressions is also the same.

   The estimates are the same because the residuals account for unexplained variables, so when those variables are plotted together, they will be the same as the residuals plotted together.

   The point behind this seems to be that it is much easier to achieve the same ends by running a multivariate regression rather than running many different regressions to get the same answer.

   On the next page you can see where I have observed the values that are the same.


Console
Terminal
Jobs
R 4.1.1 · ~/
lm(formula = resids1 ~ resids2, data = incumbents)
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) 0.0000000000000000048 0.00129860495497001092 0.00 1
resids2 0.25687701270009788423 0.01176190239602179811 21.84 <0.0000000000000002 ***

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: < 0.00000000000000022
>
> summary(reg5)
Call:
lm(formula = voteshare ~ difflog + presvote, data = incumbents)
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 <0.0000000000000002 ***
difflog 0.0355431 0.0009455 37.59 <0.0000000000000002 ***
presvote 0.2568770 0.0117637 21.84 <0.0000000000000002 ***

Figure 5: