Problem Set 3

Philip Kruger
18328699
20/11/22

Question 1

Part 1

The following code yielded the following output:

```
summary(lm(dat$voteshare~dat$difflog))
```

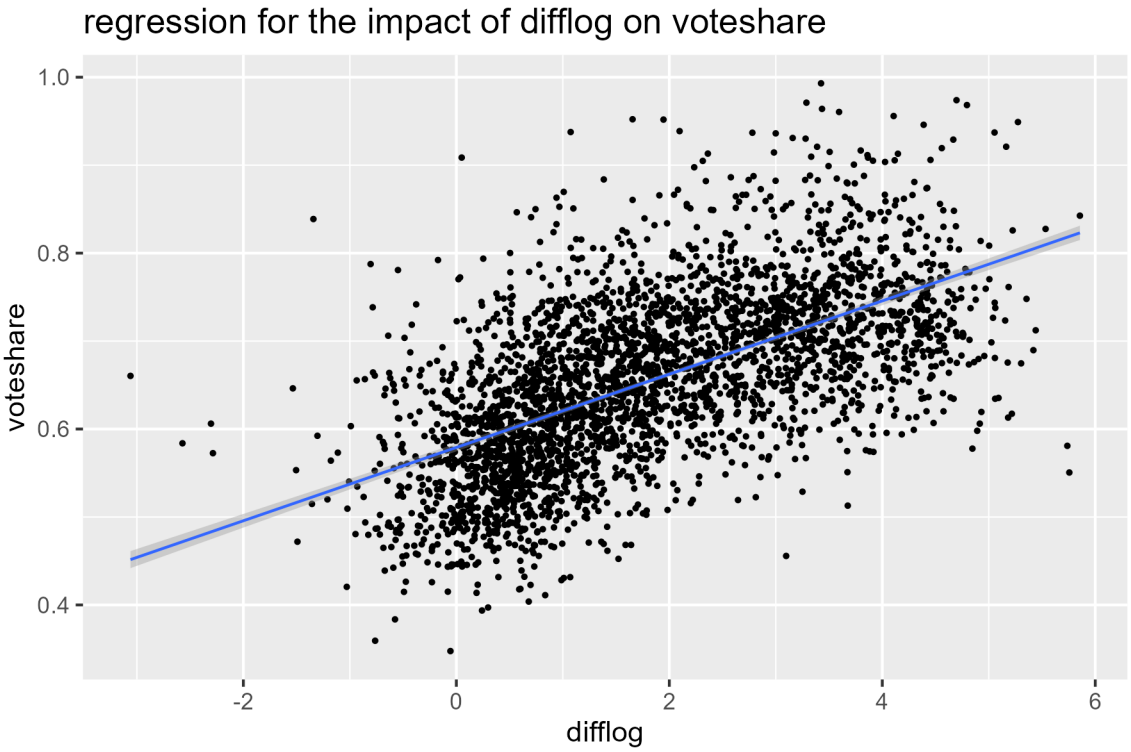
Output:

```
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 ***
dat$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
```

Part 2

The following code was used to yield the graph below:

```
ggplot(dat, mapping = aes(difflog,voteshare))+  #graph
geom_point(size = 0.5) + geom_smooth(method='lm', formula= y~x, size = 0.5)+
ggtitle("regression for the impact of difflog on voteshare")
```



Part 3

The following code is used to save the residuals:

```
res1 <- summary(lm(dat$voteshare~dat$difflog))$residual
```

Part 4

As can be seen from the regression in Part 1 the prediction equation is:

$$0.579031 + 0.041666X_1 + \epsilon$$

Question 2

Part 1

The following code yielded the following output:

```
summary(lm(dat$presvote~dat$difflog))
```

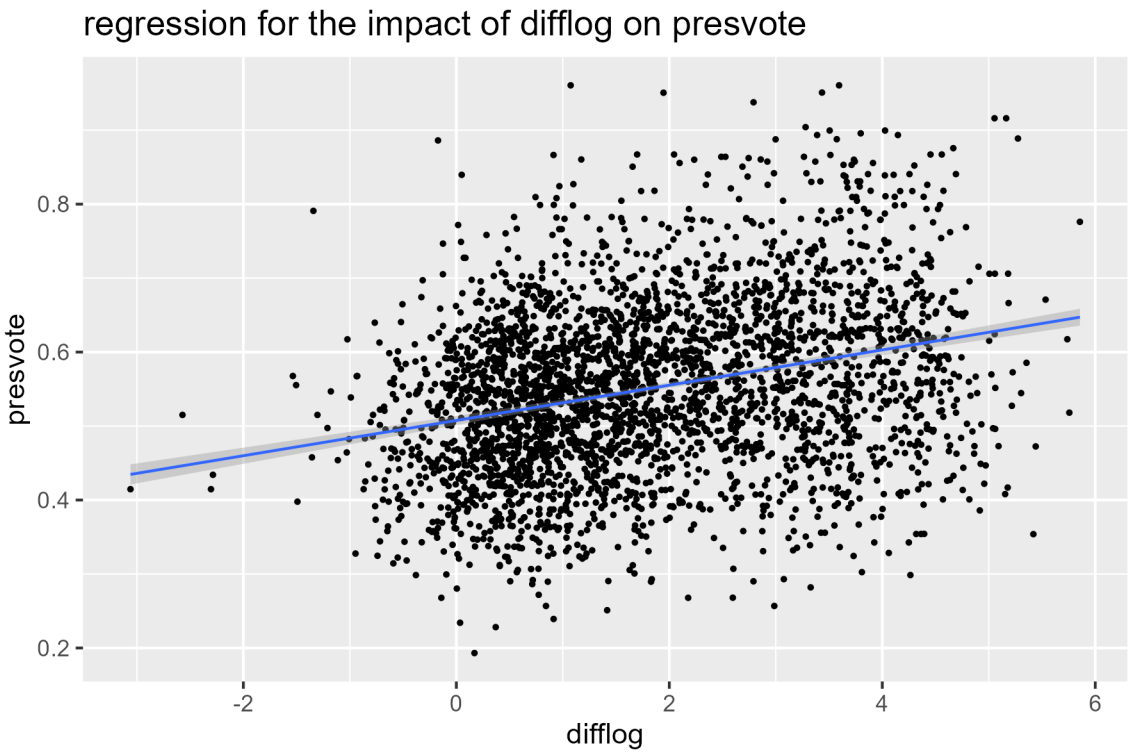
Output:

```
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 ***
dat$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
```

Part 2

The following code was used to yield the graph below:

```
ggplot(dat, mapping = aes(difflog,presvote))+  #graph
geom_point(size = 0.5) + geom_smooth(method='lm', formula= y~x, size = 0.5)+
ggtitle("regression for the impact of difflog on presvote")
```



Part 3

The following code is used to save the residuals:

```
res2 <- summary(lm(dat$presvote~dat$difflog))$residual
```

Part 4

As can be seen from the regression in Part 1 the prediction equation is:

$$0.507583 + 0.023837X_1 + \epsilon$$

Question 3

Part 1

The following code yielded the following output:

```
summary(lm(dat$voteshare~dat$presvote))
```

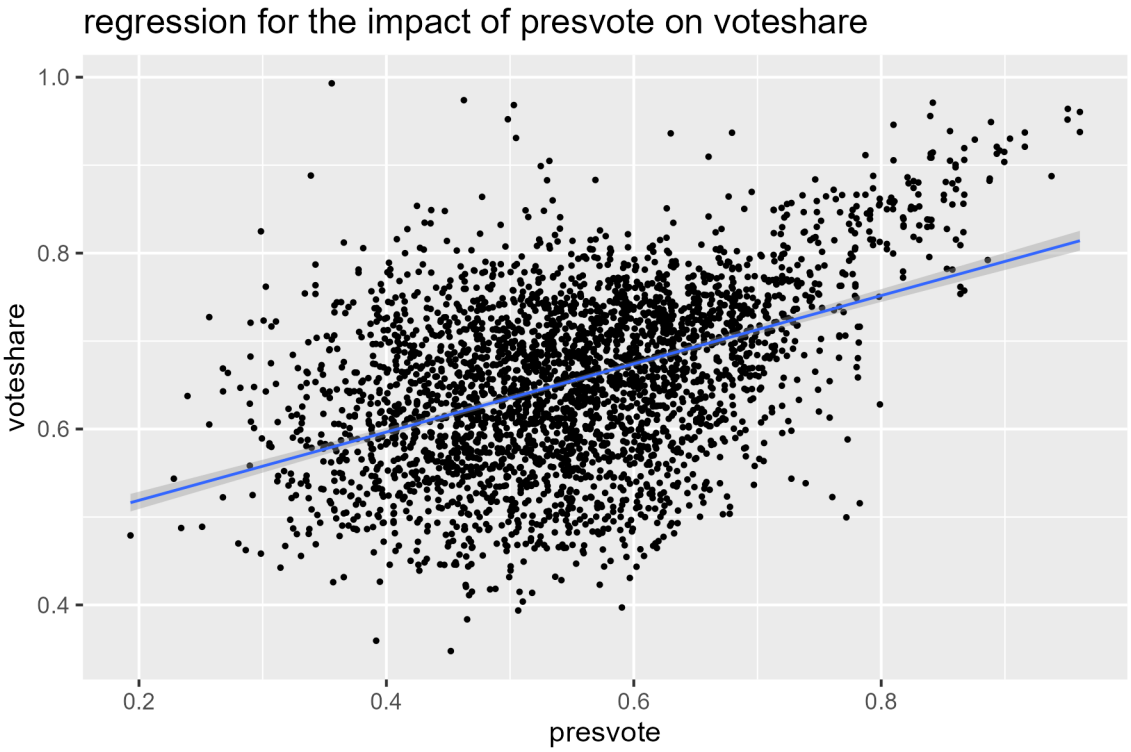
Output:

```
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 ***
dat$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
```

Part 2

The following code was used to yield the graph below:

```
ggplot(dat, mapping = aes(presvote,voteshare))+  #graph
geom_point(size = 0.5) + geom_smooth(method='lm', formula= y~x, size = 0.5)+
ggtitle("regression for the impact of presvote on voteshare")
```



Part 3

As can be seen from the regression in Part 1 the prediction equation is:

$$0.441330 + 0.388018X_1 + \epsilon$$

Question 4

Part 1

The following code yielded the following output:

```
summary(lm(res1~res2))
```

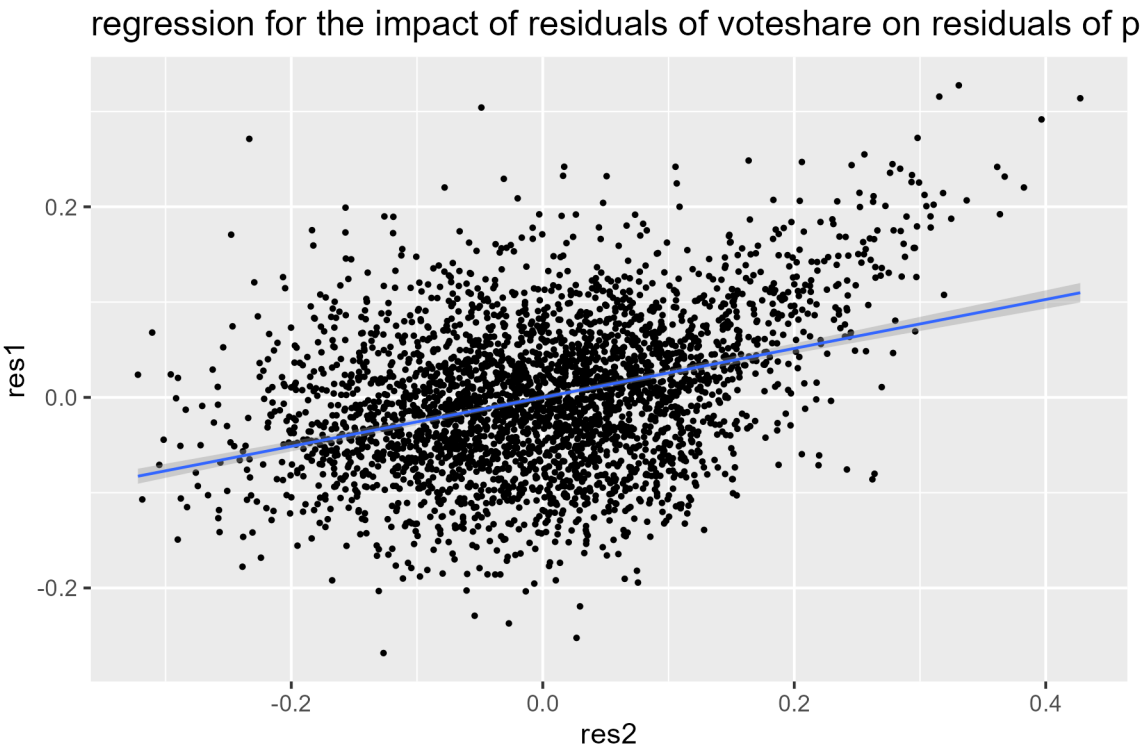
Output:

```
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) -4.860e-18  1.299e-03    0.00        1
res2         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
```

Part 2

The following code was used to yield the graph below:

```
ggplot(dat, mapping = aes(res2,res1))+  #graph
geom_point(size = 0.5) + geom_smooth(method='lm', formula= y~x, size = 0.5)+
ggtitle("regression for the impact of residuals of voteshare on residuals of presvote")
```



Part 3

As can be seen from the regression in Part 1 the prediction equation is:

$$-4.86 * 10^{-18} + 0.2569X_1 + \epsilon$$

Question 5

Part 1

The following code yielded the following output:

```
summary(lm(dat$voteshare~cbind(dat$difflog,dat$presvote)))
```

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 ***
cbind(dat$difflog, dat$presvote)1 0.0355431  0.0009455   37.59   <2e-16 ***
cbind(dat$difflog, dat$presvote)2 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
```

Part 2

As can be seen from the regression in Part 1 the prediction equation is:

$$0.44864 + 0.03554X_1 + 0.25688X_2 + \epsilon$$

Part 3

As can be seen in the regressions in Question 4 Part 1 and Question 5 Part 1 the residuals are identical. A residual is the difference between the actual value and the value predicted by the model. As such it makes sense that these are identical as in both cases the residual is simply measuring the inherent variability + the effect of other variables not accounted for in the model. The inherent variability of the dataset will be identical for both regressions as they are done on the same data set and since question 4 and 5 both account for the same factors (effect on difflog and presvote on voteshare) it makes sense that the effect of variables not accounted for on the residuals are the same. As such the residuals are identical.