Data 606 - Lab 8

Cameron Smith

Contents

Main Exercises

Exercise 1

What are the dimensions of the dataset?

The 'hfi' dataset has 1,458 observations and 123 variables.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro)
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
library(statsr)
```

```
##
## Attaching package: 'statsr'
```

```
## The following objects are masked from 'package:openintro':
##
##     calc_streak, evals, nycflights, present
```

```
data(hfi)
```

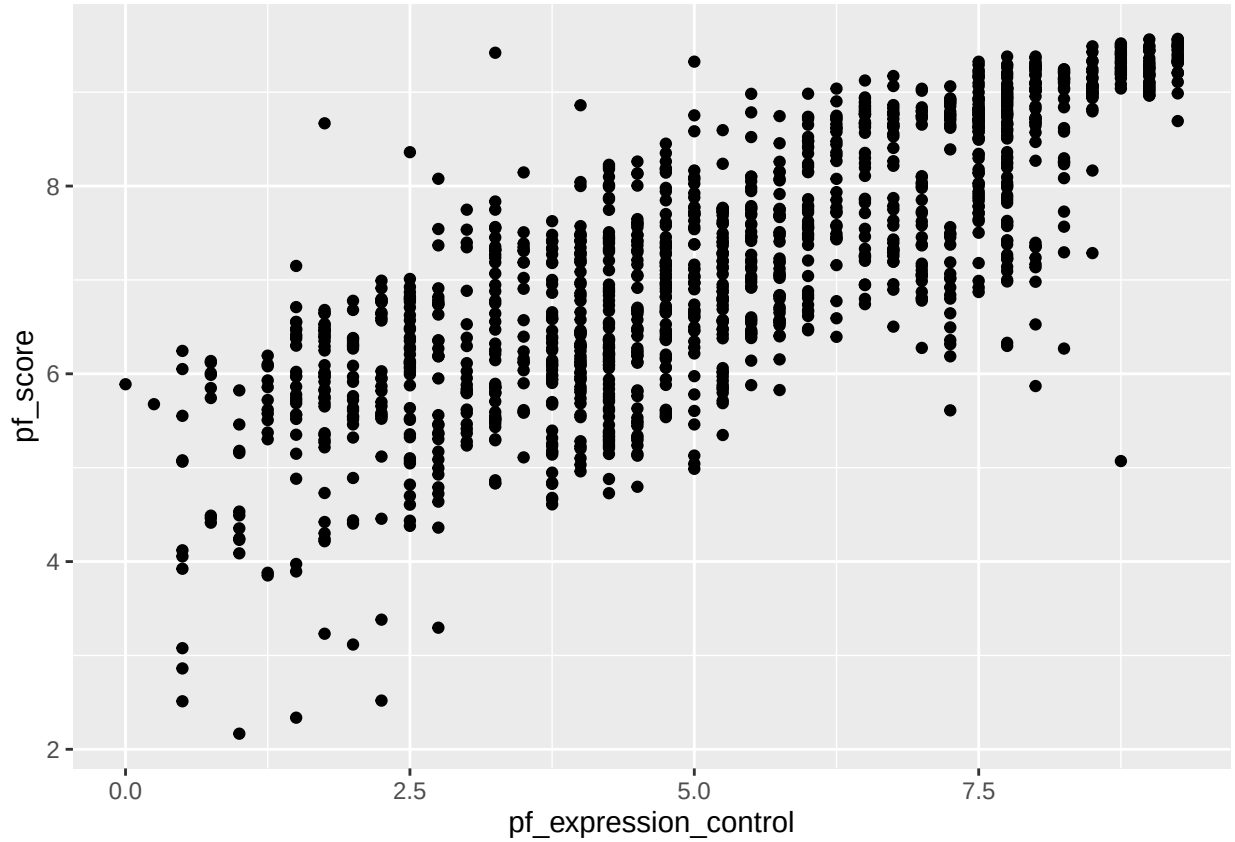
Exercise 2

What type of plot would you use to display the relationship between the personal freedom score, pf_score, and one of the other numerical variables? Plot this relationship using the variable pf_expression_control as the predictor. Does the relationship look linear? If you knew a country's pf_expression_control, or its score out of 10, with 0 being the most, of political pressures and controls on media content, would you be comfortable using a linear model to predict the personal freedom score?

A scatter plot can be used to quickly visualize the relationship (or lack thereof) between two variables, as follows. Yes there appears to be a positive correlation between pf_expression control and pf_score.

```
hfi %>%  ggplot(aes(x = pf_expression_control, y = pf_score)) +
    geom_point()
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



```r
# Quantify the strength of the relationship w/ the correlation coefficient
hfi %>%
  summarise(cor(pf_expression_control, pf_score, use = "complete.obs"))
```

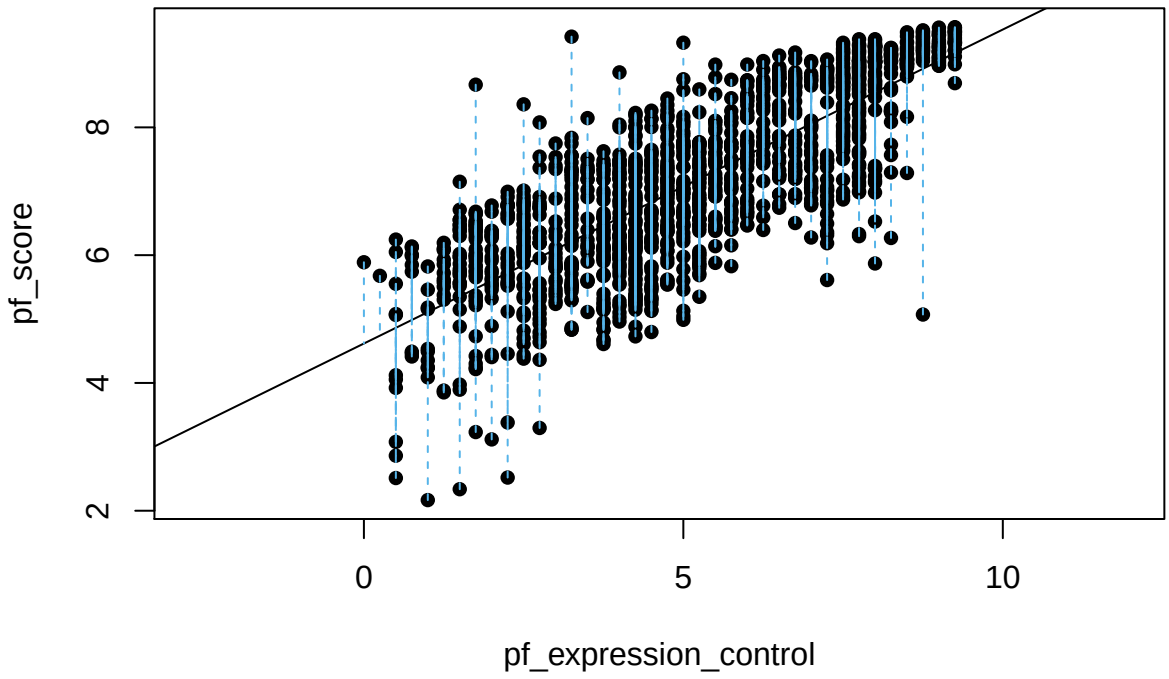
```
## # A tibble: 1 x 1
##   `cor(pf_expression_control, pf_score, use = "complete.obs")`
##                                                          <dbl>
## 1                                                        0.796
```

Exercise 3

Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

There is a strong upward linear trend (i.e. positive linear correlation) between the variables.

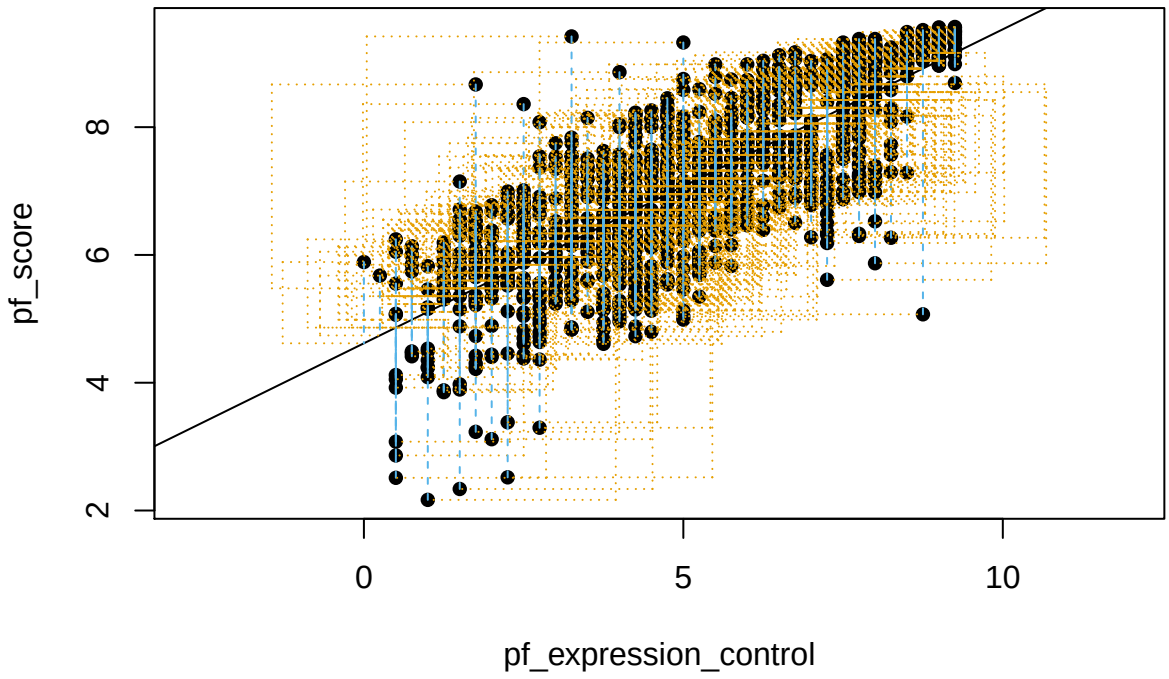
```r
newdf <- hfi %>% select(pf_expression_control, pf_score) %>% drop_na()

plot_ss(x = pf_expression_control, y = pf_score, data = newdf)
```




```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##      4.6171       0.4914
```

```
##
## Sum of Squares:  952.153
```

```
# Run again, but showing the squared residuals
plot_ss(x = pf_expression_control, y = pf_score, data = newdf, showSquares = TRUE)
```




```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##      4.6171       0.4914
##
## Sum of Squares:  952.153
```

Exercise 4

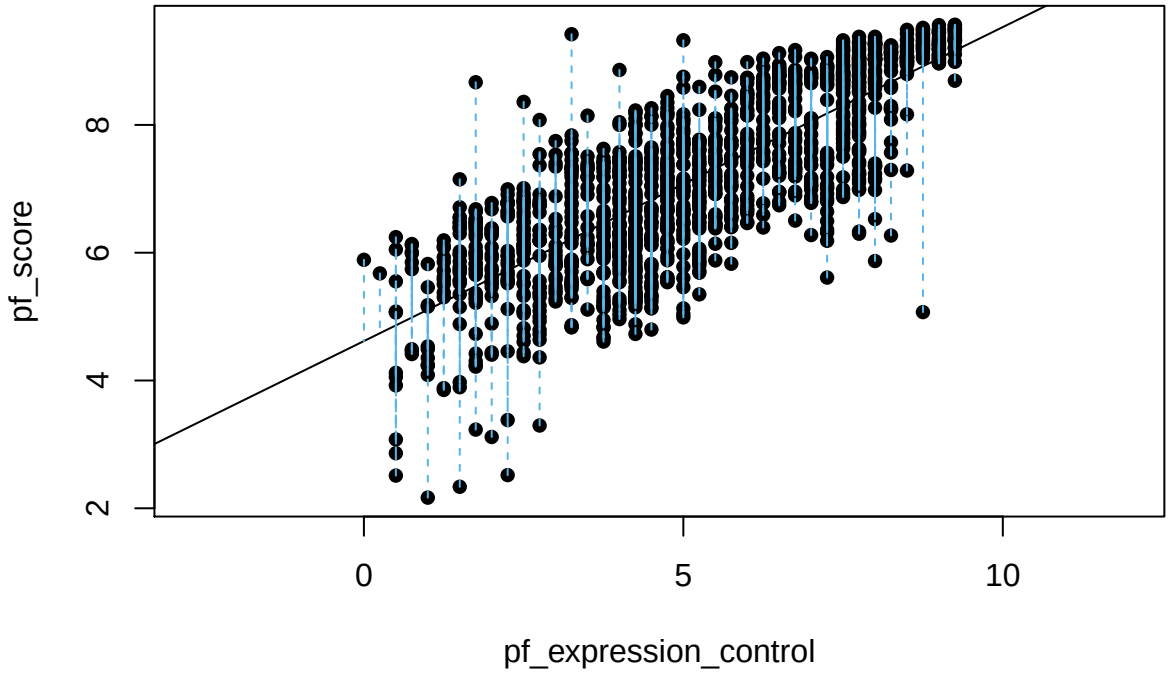
Using plot_ss, choose a line that does a good job of minimizing the sum of squares. Run the function several times. What was the smallest sum of squares that you got? How does it compare to your neighbors?

I ran the plot_ss code below 5 times with the below results - the smallest being 976. My (virtual) neighbors had similar results.

- 1154
- 976
- 989
- 1040

- 1090

```
plot_ss(x = pf_expression_control, y = pf_score, data = newdf)
```




```
## Click two points to make a line.
## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##      4.6171       0.4914
##
## Sum of Squares:  952.153
```

Exercise 5

Fit a new model that uses pf_expression_control to predict hf_score, or the total human freedom score. Using the estimates from the R output, write the equation of the regression line. What does the slope tell us in the context of the relationship between human freedom and the amount of political pressure on media content?

Based on the output for the below model (m2), the equation of the regression line is as follows:

$$\hat{y} = 5.153687 + 0.349862 \times pf\_expression\_control$$

The slope tells us that there is a positive correlation between the two variables.

```r
# Examples from paragraph above exercise
m1 <- lm(pf_score ~ pf_expression_control, data = hfi)
summary(m1)
```

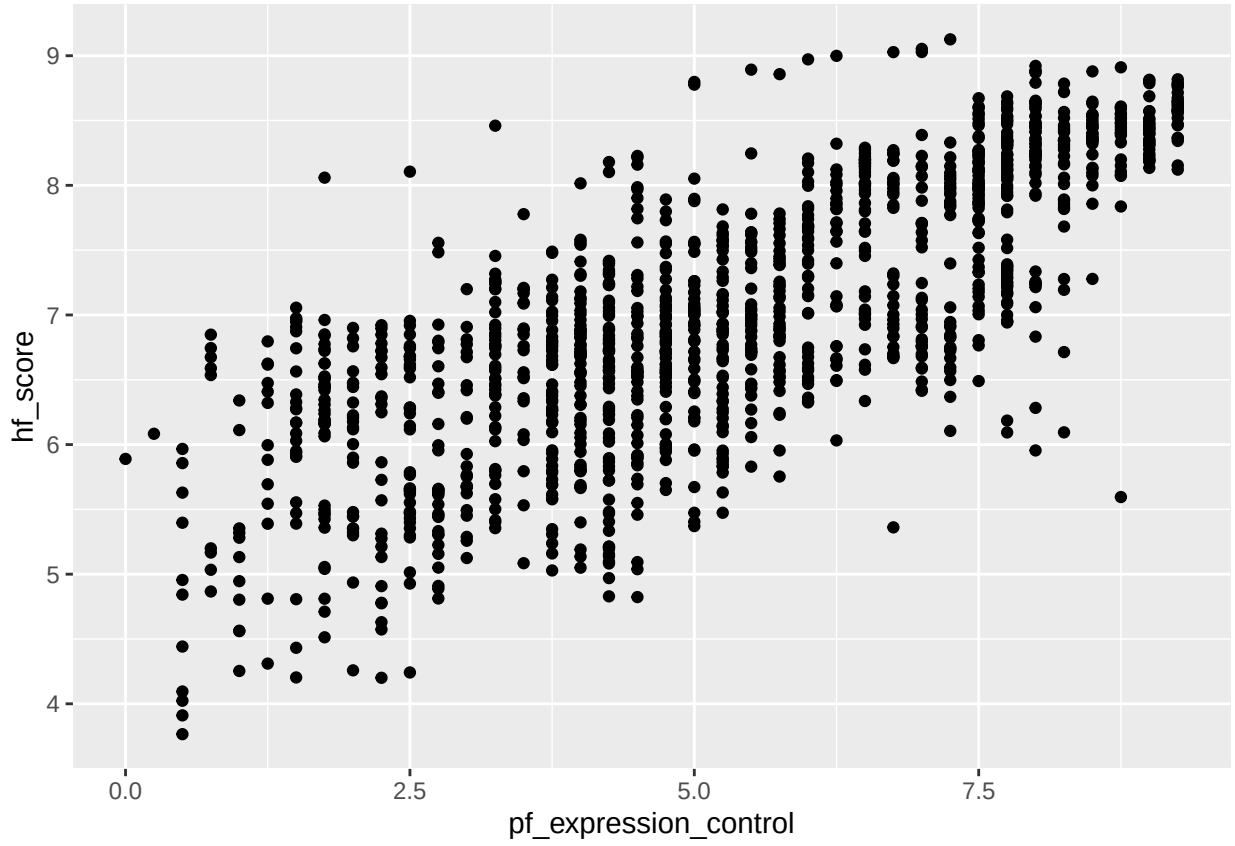
```
##
## Call:
## lm(formula = pf_score ~ pf_expression_control, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8467 -0.5704  0.1452  0.6066  3.2060
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.61707    0.05745   80.36   <2e-16 ***
## pf_expression_control  0.49143    0.01006   48.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8318 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.6342, Adjusted R-squared:  0.634
## F-statistic:  2386 on 1 and 1376 DF,  p-value: < 2.2e-16
```

```r
# New model with hf_score predicted from expression control
m2 <- lm(hf_score ~ pf_expression_control, data = hfi)
summary(m2)
```

```
##
## Call:
## lm(formula = hf_score ~ pf_expression_control, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6198 -0.4908  0.1031  0.4703  2.2933
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.153687   0.046070  111.87   <2e-16 ***
## pf_expression_control 0.349862   0.008067   43.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.667 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.5772
## F-statistic:  1881 on 1 and 1376 DF,  p-value: < 2.2e-16
```

```r
# Plot the data used for the new model
hfi %>%  ggplot(aes(x = pf_expression_control, y = hf_score)) +
    geom_point()
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



Exercise 6

If someone saw the least squares regression line and not the actual data, how would they predict a country's personal freedom school for one with a 6.7 rating for pf_expression_control? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

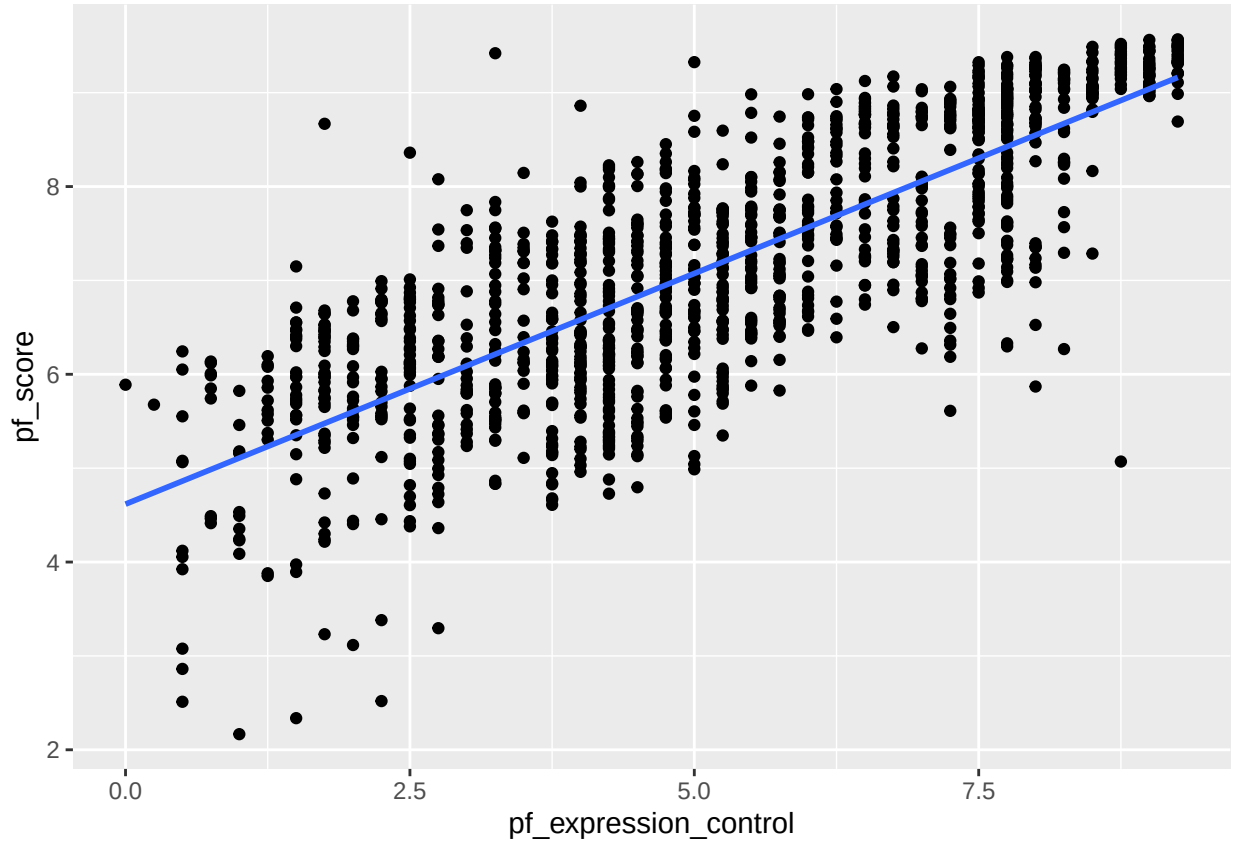
Based on 'eye balling' the data, I would predict a pf_score of just under 8. This would in fact likely be an overestimate, as the majority of similar data points fall under that point of the regression line. There would not actually be a residual for the prediction. Residuals only apply to actual data points, indicating the space between the point and the line. A prediction would be right on the line.

```
# Examples fom paragraph above exercise
ggplot(data = hfi, aes(x = pf_expression_control, y = pf_score)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 80 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```

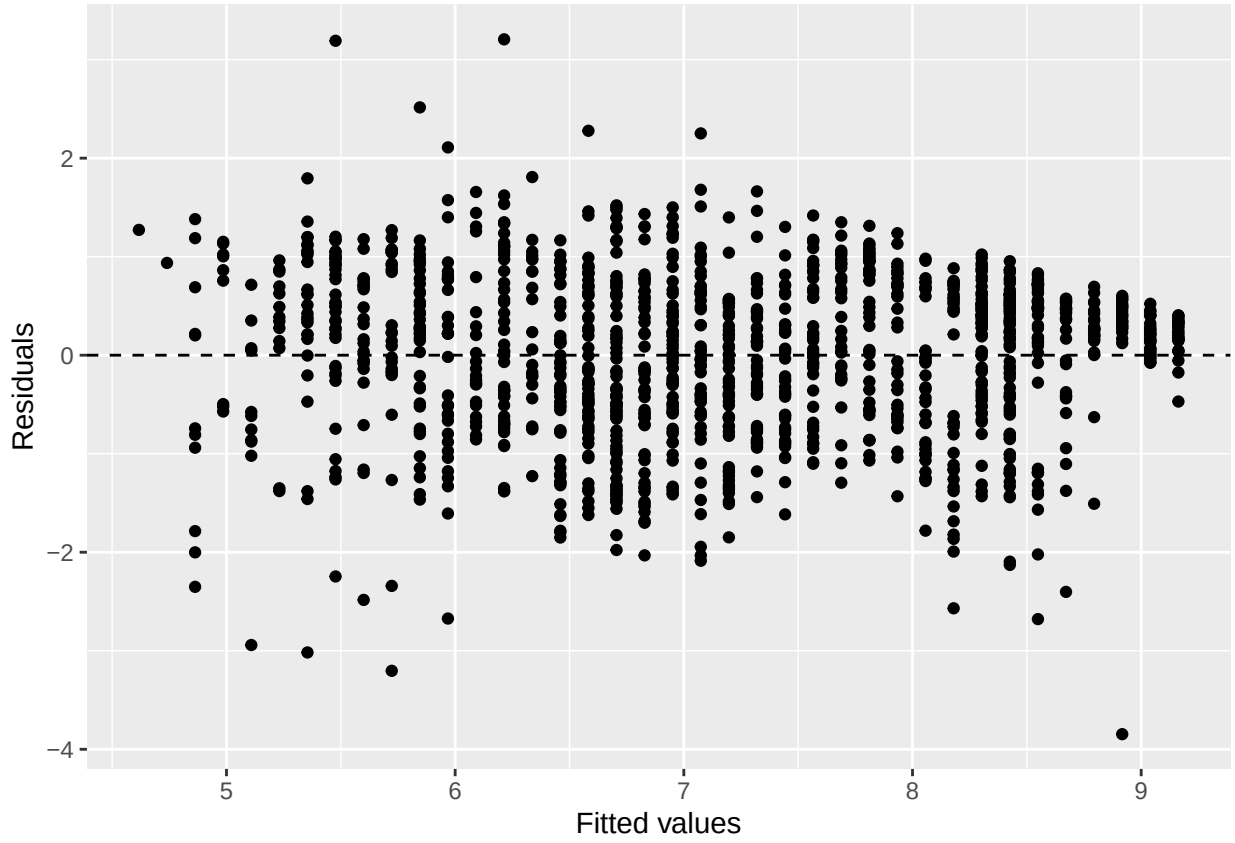


Exercise 7

Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between the two variables?

The residuals plot seems to indicate linearity, as the data points (roughly) follow the line at y = 0. As with the scatter plot above, the plot indicatives a positive linear relationship.

```
ggplot(data = m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```

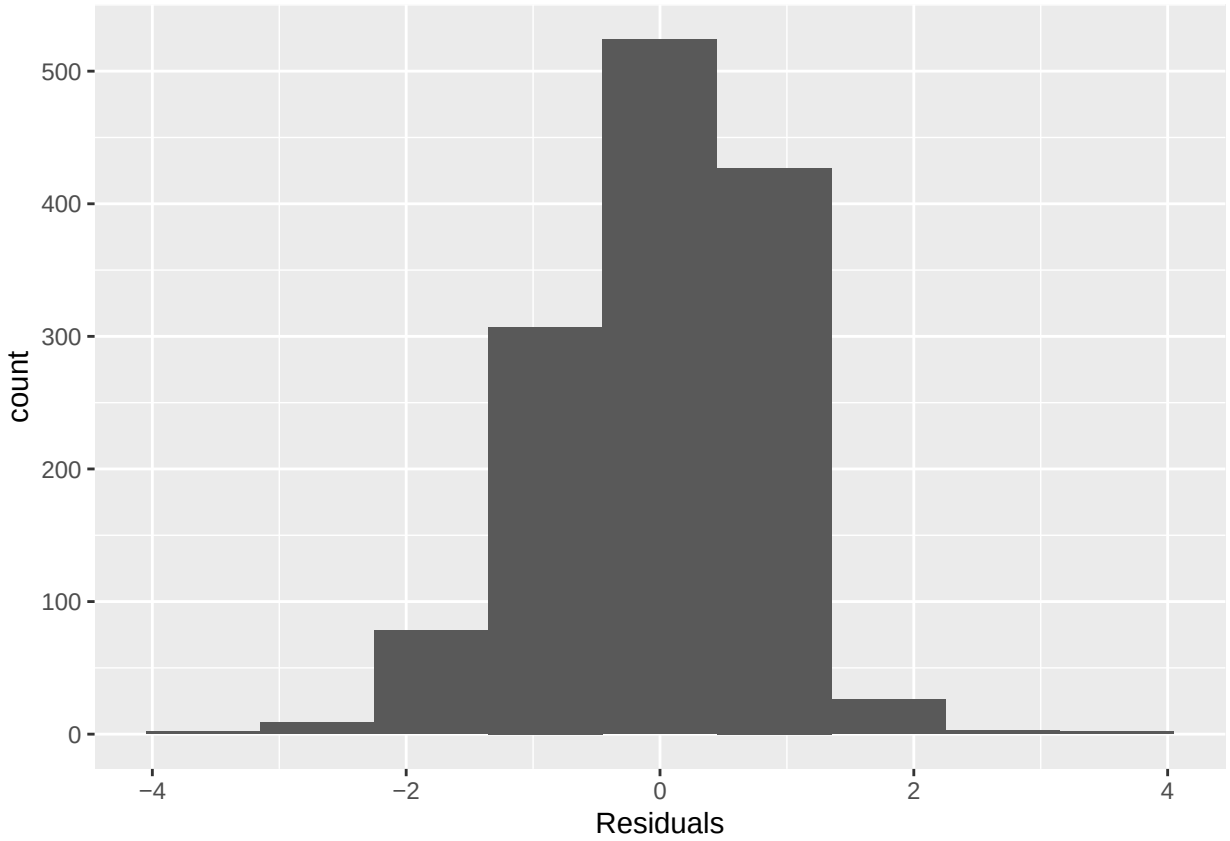


Exercise 8

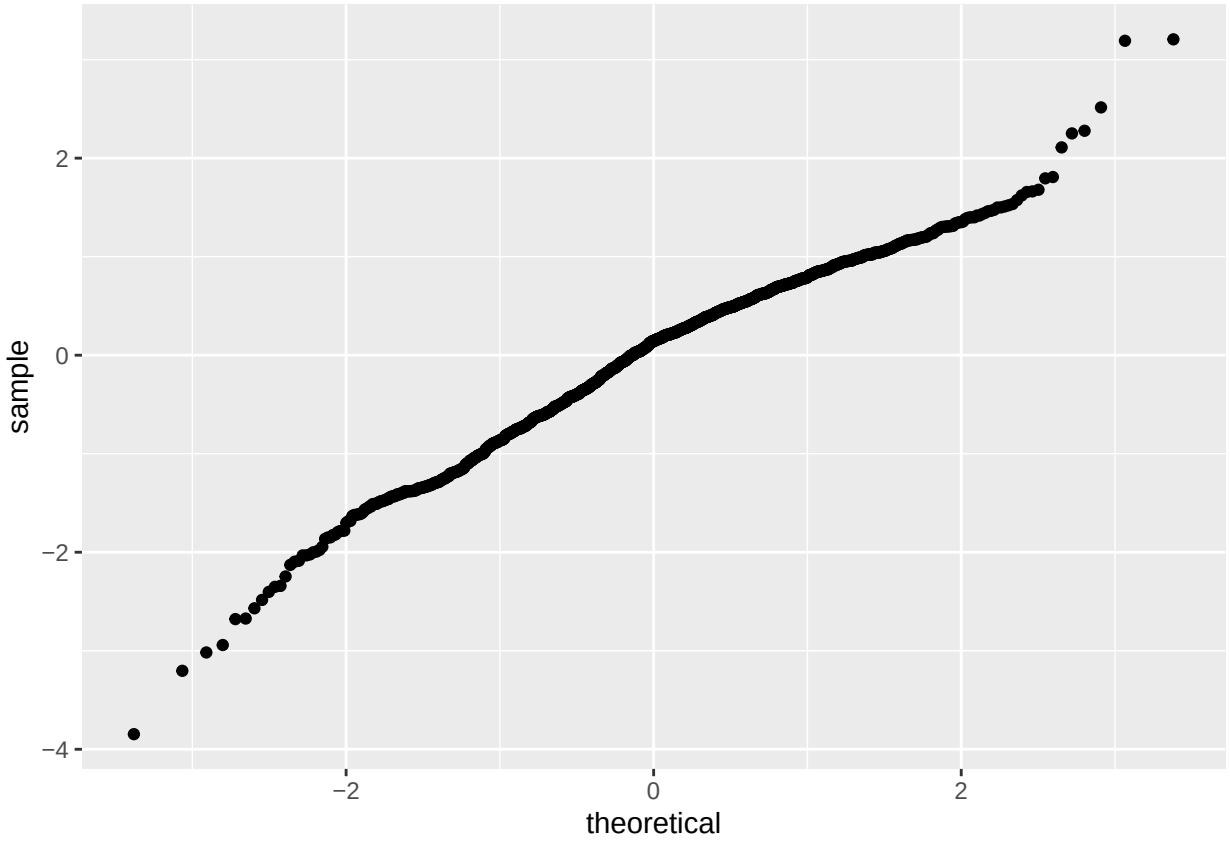
**Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?*

Yes, the nearly normal residuals conditions appears to be met. The normality plot ('QQ plot') in particular indicates a nearly normal distribution, with a very strong relationship between the sample values (residuals) and the theoretical values. The histogram did not indicate anything useful with the default values (bin width of 25), but when I changed it to a lower value it also indicates a normal distribution.

```
# Nearly normal residuals
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(binwidth = .9) +
  xlab("Residuals")
```




```
# Normality plot
ggplot(data = m1, aes(sample = .resid)) +
  stat_qq()
```



Exercise 9

Based on the residuals vs. fitted plot, does the constant variability condition appear to be met?

Yes, it seems to although as tHe fitted values approach 8 or so a different pattern appears to be emerging. However, constant variability seems to apply to the majority of the data.

More Practice

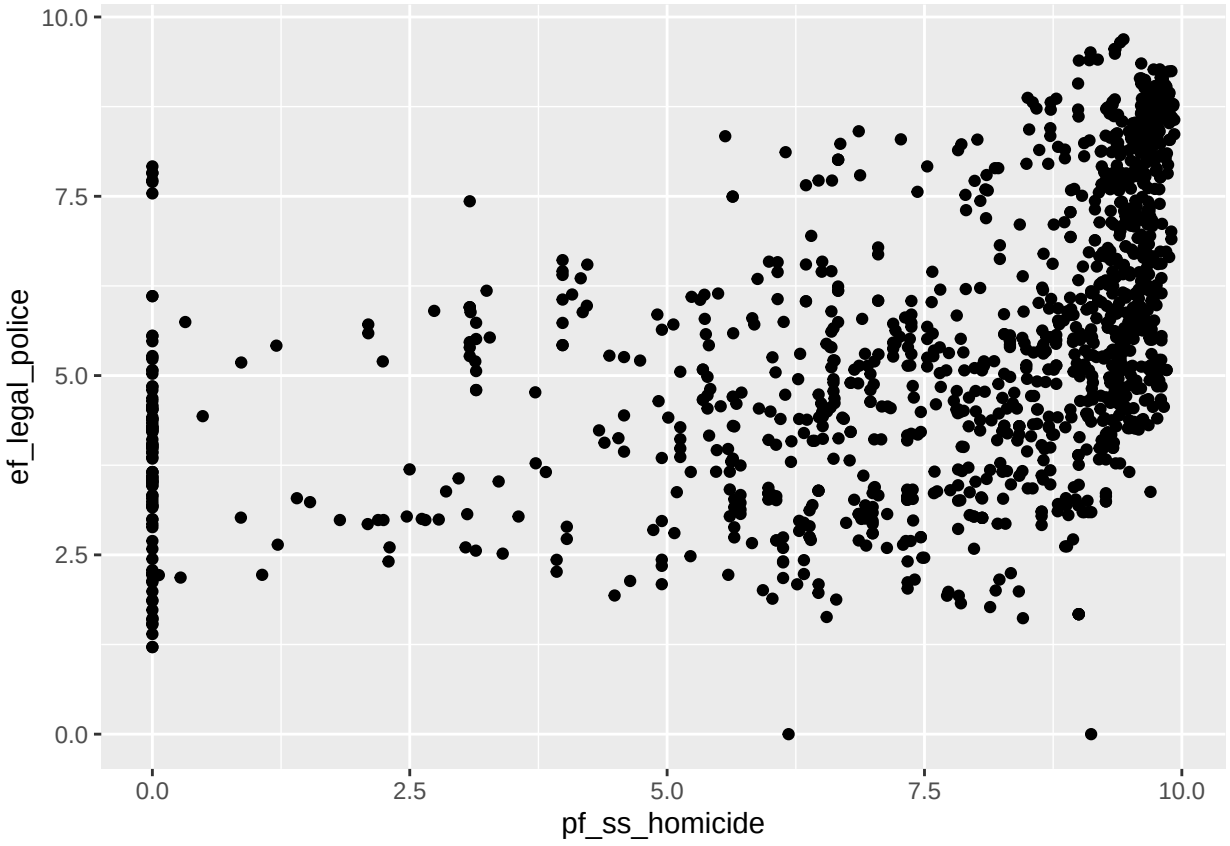
Additonal Example 1

Choose another freedom variable and a variable you think would strongly correlate with it.. Produce a scatterplot of the two variables and fit a linear model. At a glance, does there seem to be a linear relationship?

For the below example I chose pf_ss_homicide (Homicide) and ef_legal_police (Reliability of police), the general hypothesis being that the number of homicides will increase as the reliability of police decreases. Based on the scatterplot it does look there is a linear relationship but it is difficult to tell whether it is linear.

```
# Scatter plot
hfi %>% ggplot(aes(x = pf_ss_homicide, y = ef_legal_police)) +
  geom_point()
```

```
## Warning: Removed 169 rows containing missing values (geom_point).
```



Additional Example 2

How does this relationship compare to the relationship between pf_expression_control and pf_score? Use the R2 values from the two model summaries to compare. Does your independent variable seem to predict your dependent one better? Why or why not?

The r squared values from the pf_expression control vs pf_score model are: 0.5775, 0.5772 (adjusted) The r squared values from the pf_ss_homicide vs ef_legal_police model are: 0.2219, 0.2213 (adjusted)

Based on the above the first model seems to have a much stronger correlation, as nearly 58% of the variability of explained by the independent variable versus only about 22% in the second model.

```
# Model based on two new variables
m3 <- lm(pf_ss_homicide ~ ef_legal_police, data = hfi)

# View output / summary of model
summary(m3)
```

```
##
## Call:
## lm(formula = pf_ss_homicide ~ ef_legal_police, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1278 -0.4915  0.4980  1.6145  5.2981
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      3.82074    0.20492   18.64   <2e-16 ***
## ef_legal_police  0.67039    0.03499   19.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.471 on 1287 degrees of freedom
##   (169 observations deleted due to missingness)
## Multiple R-squared:  0.2219, Adjusted R-squared:  0.2213
## F-statistic: 367.1 on 1 and 1287 DF,  p-value: < 2.2e-16
```

Additional Example 3

What's one freedom relationship you were most surprised about and why? Display the model diagnostics for the regression model analyzing this relationship.

In addition to the model I mentioned above (murders vs reliability of police), which I found quite surprising, I found it interesting that there seems to be little if no relationship between ef_legal_integrity (Integrity of the legal system) and ef_government (Size of government), with an r squared of only 12.9.

```
# Model based on two new variables
m4 <- lm(ef_legal_integrity ~ ef_government, data = hfi)

# View output / summary of model
summary(m4)
```

```
##
## Call:
## lm(formula = ef_legal_integrity ~ ef_government, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2572 -1.5800  0.2034  1.6414  3.9208
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.02089    0.29628   33.82   <2e-16 ***
## ef_government -0.59598    0.04503  -13.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.013 on 1179 degrees of freedom
##   (277 observations deleted due to missingness)
## Multiple R-squared:  0.1293, Adjusted R-squared:  0.1286
## F-statistic: 175.1 on 1 and 1179 DF,  p-value: < 2.2e-16
```