Introduction to linear regression

```
library(tidyverse)
library(openintro)
data('hfi', package='openintro')
```

Exercise 1

What are the dimensions of the dataset?

```
dim(hfi)
```

```
## [1] 1458  123
```

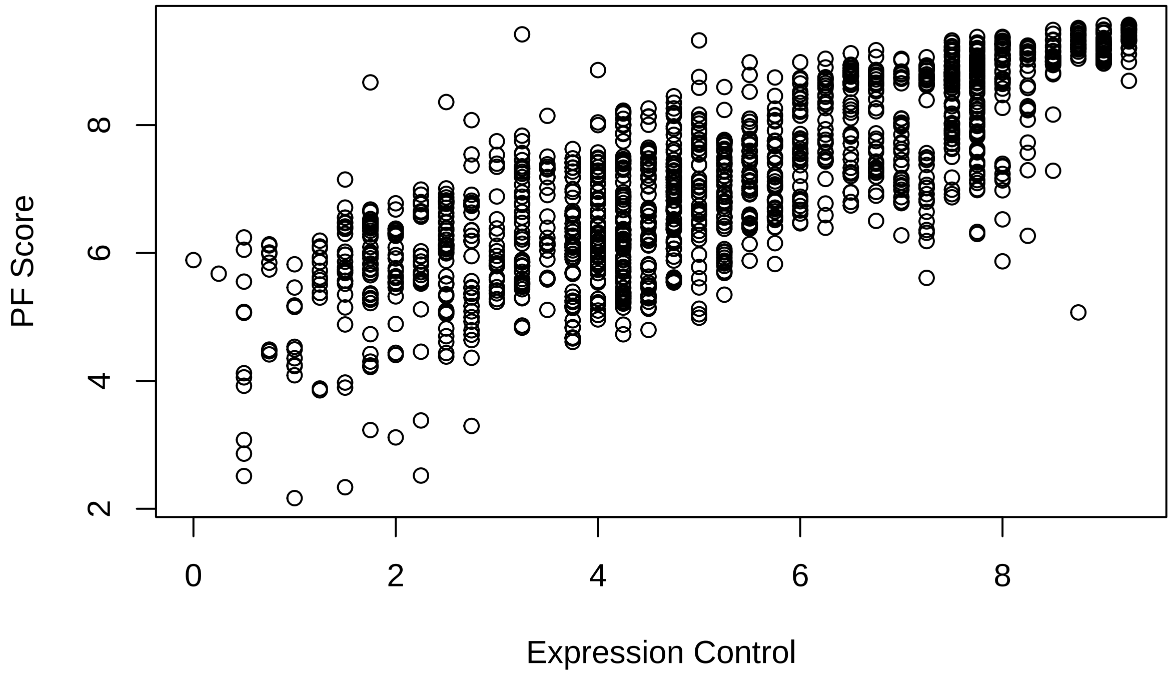
> There are 123 columns (dimensions) of the dataset – at a high level, each row represents one country for one year and there are scores based on different levels of freedom for various types (political, religious, economical, personal).

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 scatterplot would be a good way to display the relationship between two numerical variables.

```
plot(hfi$pf_score ~ hfi$pf_expression_control,
     xlab = "Expression Control", ylab = "PF Score")
```



The relationship does look linear in nature. Given the two values seem correlated in some way, I would be comfortable using a linear model to predict the personal freedom score given `pf_expresion_control`.

Exercise 3

```
hfi %>%
  summarise(cor(pf_expression_control, pf_score, use = "complete.obs"))
```

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.

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

The relationship between these two variables is moderately positive as the correlation between them is 0.796; as one goes up, so does the other. There are a few outliers at the lower ends of the spectrum, but otherwise it seems to be a solid relationship.

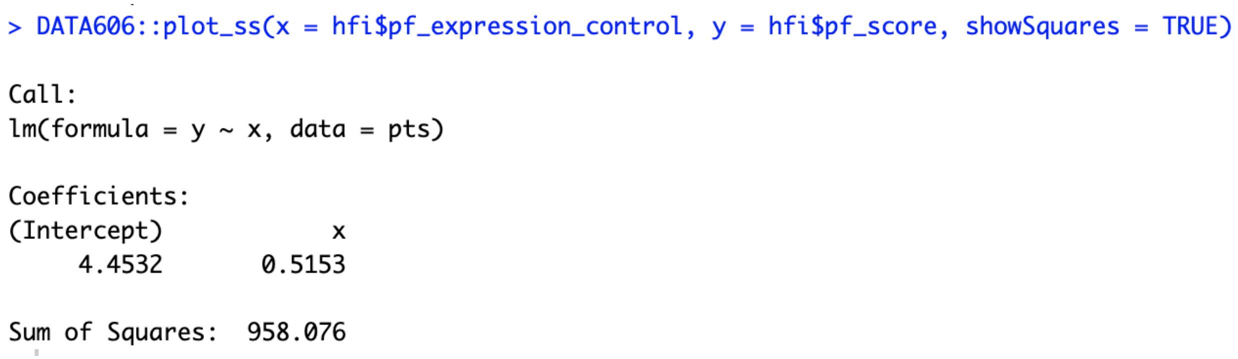
Exericse 4

```
# This will only work interactively (i.e. will not show in the knitted document)
hfi <- hfi %>% filter(complete.cases(pf_expression_control, pf_score))
DATA606::plot_ss(x = hfi$pf_expression_control, y = hfi$pf_score)
```

```
DATA606::plot_ss(x = hfi$pf_expression_control, y = hfi$pf_score, showSquares = TRUE)
```

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?

> 958.076.



```
> DATA606::plot_ss(x = hfi$pf_expression_control, y = hfi$pf_score, showSquares = TRUE)

Call:
lm(formula = y ~ x, data = pts)

Coefficients:
(Intercept)            x
     4.4532       0.5153

Sum of Squares:  958.076
```

Figure 1: Lowest Sum of Squares at 958.076

Exercise 5

```
m1 <- lm(pf_score ~ pf_expression_control, data = hfi)
m2 <- lm(hf_score ~ pf_expression_control, data = hfi)
```

```
summary(m2)
```

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?

```
##
## 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
```

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

> This slope tells us that human freedom is less effected by the political pressure on media content than `pf_score`, but there is still an effect. Specifically, for each additional amount of political pressure on media content score, the human freedom score increases by 0.3499

Exercise 6

```r


pf_expression_control <- 6.7

pf_prediction <- 4.61707 + 0.49143 * pf_expression_control
pf_prediction
```

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?

```
## [1] 7.909651
```

> We would expect 7.909651 here, but how does it compare to those with an actual control value of 6.7? There are none, so let's look at those with at 6.75.

```r
library(dplyr)
expected <- hfi |> select(countries, pf_expression_control, pf_score) |>
  filter(pf_expression_control == 6.75) |>
  arrange(pf_score)

head(expected, 20)
```

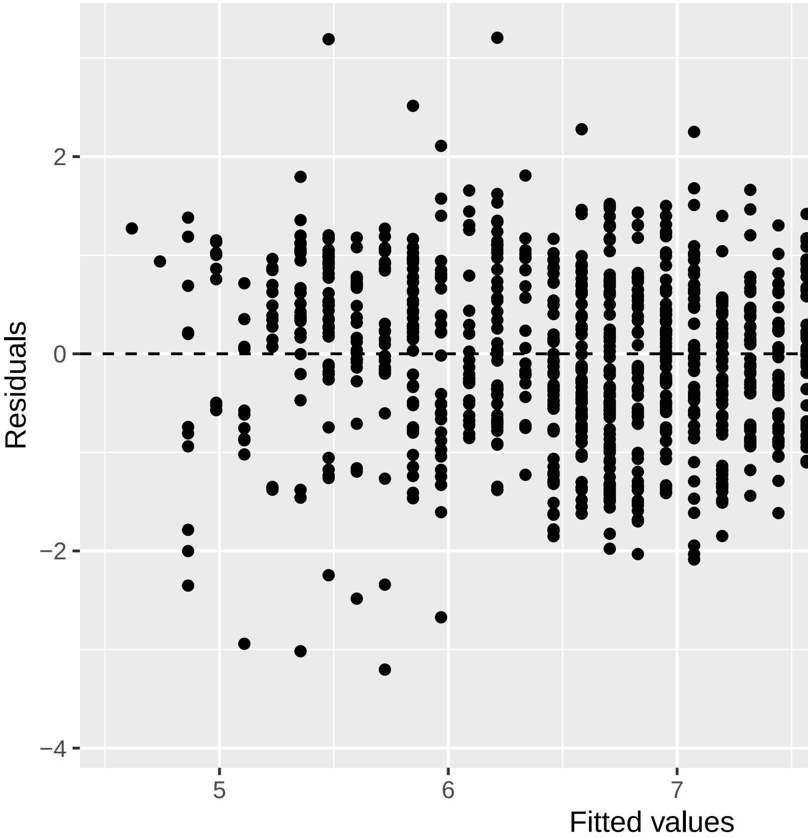
```
## # A tibble: 20 x 3
##    countries       pf_expression_control pf_score
##    <chr>                           <dbl>    <dbl>
##  1 Myanmar                          6.75     6.50
##  2 Pap. New Guinea                  6.75     6.90
##  3 Guyana                           6.75     6.95
##  4 Pap. New Guinea                  6.75     7.19
##  5 Guyana                           6.75     7.19
##  6 Pap. New Guinea                  6.75     7.25
##  7 Suriname                         6.75     7.29
##  8 Burkina Faso                     6.75     7.31
##  9 Namibia                          6.75     7.32
## 10 Guyana                           6.75     7.34
## 11 Namibia                          6.75     7.39
## 12 Belize                           6.75     7.43
## 13 Mongolia                         6.75     7.58
## 14 Guyana                           6.75     7.63
## 15 Suriname                         6.75     7.75
## 16 Suriname                         6.75     7.79
## 17 Ghana                            6.75     7.87
## 18 Chile                            6.75     8.22
## 19 Chile                            6.75     8.27
## 20 Bulgaria                         6.75     8.41
```

For the most part, this estimate is relatively accurate – the closest one to our estimate would be Ghana at 7.87. The residual for this would be 7.87 - 7.909651 which is -0.039651, which would technically be an overestimate; given our value was 6.7 though and this one was 6.5, it's actually quite close.

Exercise 7

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

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

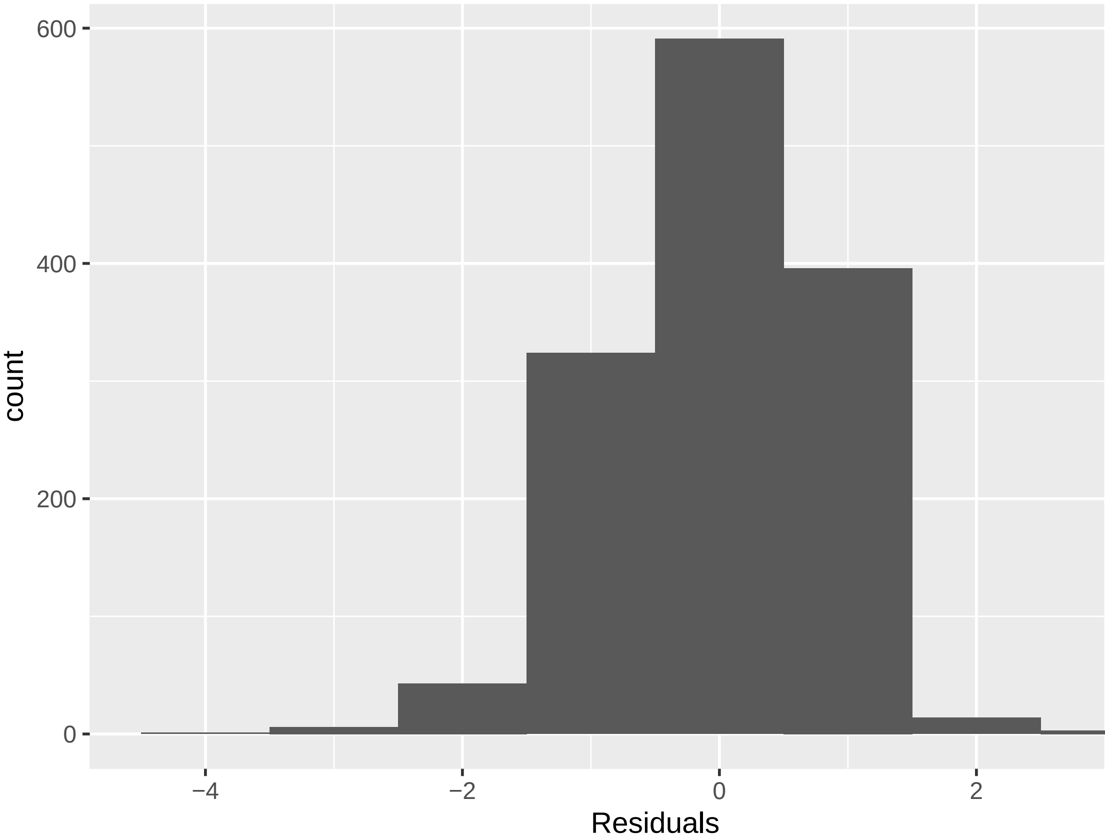


> There is no apparent pattern in the residuals plot, indicating that there is a linear relationship between the two variables given they lie around 0.

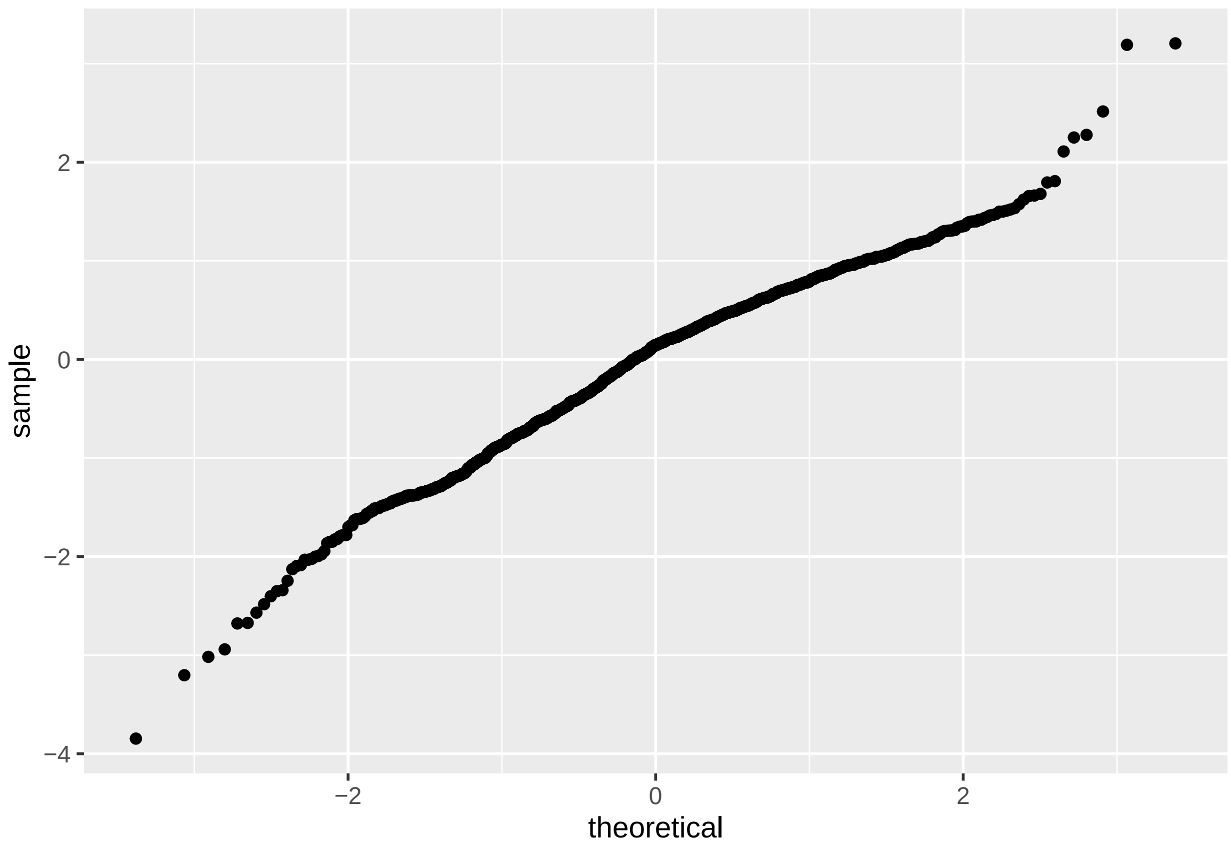
Exercise 8

```
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  xlab("Residuals")
```

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



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



It does look like the condition appears to be met given the data looks normal in both plots.

Exercise 9

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

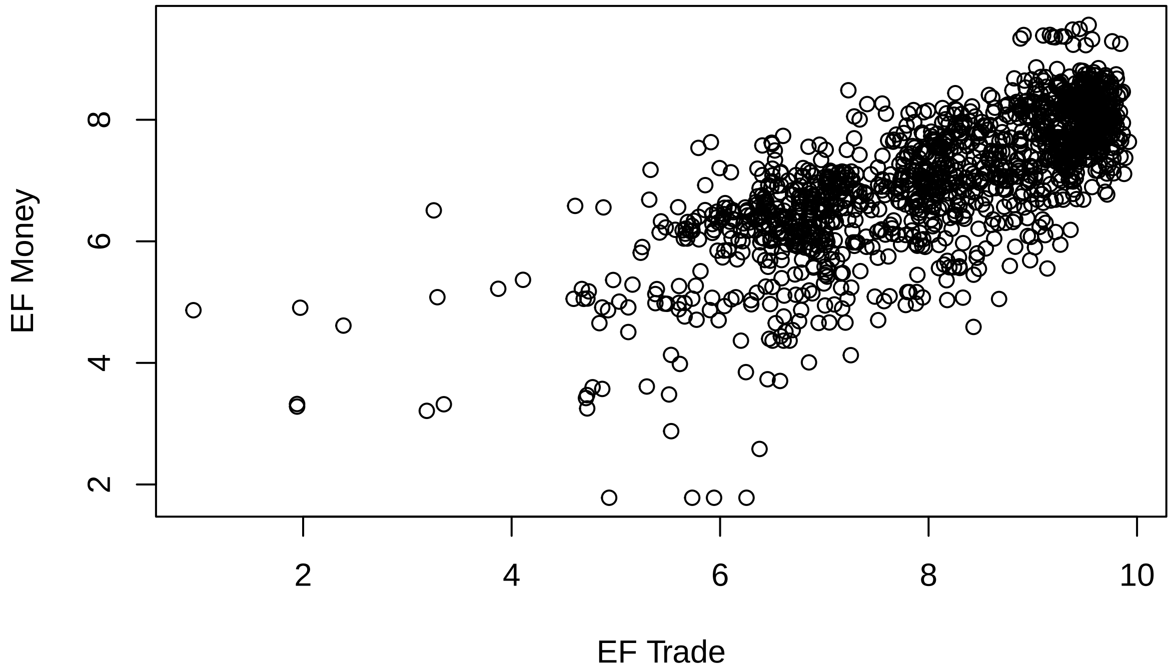
The constant variability condition does appear to be met as the variability of points around the least squares line looks roughly constant.

Exercise 10

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 this, I'll choose `ef_trade` and `ef_money`; I believe they'd be positively correlated as the more trade, the sounder money would be.

```r
plot(hfi$ef_trade ~ hfi$ef_money,
     xlab = "EF Trade", ylab = "EF Money")
```



```
m3 <- lm(ef_trade ~ ef_money, data = hfi)
m3
```

```
##
## Call:
## lm(formula = ef_trade ~ ef_money, data = hfi)
##
## Coefficients:
## (Intercept)     ef_money
##      2.1712       0.6016
```

At first glance, there does seem to be a linear relationship.

Exercise 11

```


summary(m2)
```

How does this relationship compare to the relationship between `pf_expression_control` and `pf_score`? Use the R^2 values from the two model summaries to compare. Does your independent variable seem to predict your dependent one better? Why or why not?

```
##
```

```
## 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
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = ef_trade ~ ef_money, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1515 -0.3694  0.0639  0.4912  2.3809
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.17118    0.12058   18.01   <2e-16 ***
## ef_money     0.60161    0.01463   41.11   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7464 on 1373 degrees of freedom
##   (83 observations deleted due to missingness)
## Multiple R-squared:  0.5517, Adjusted R-squared:  0.5514
## F-statistic:  1690 on 1 and 1373 DF,  p-value: < 2.2e-16
```

> Compared to the previous relationship, my R^2 is 0.5514 vs. 0.5772. This would indicate that the other model can explain more variability (~2% more) than my model.

Exercise 12

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

> I was surprised about `pf_religion` and `pf_religion_restrictions` – inherently, you’d think these would potentially be negatively correlated (the more restrictions, the less freedom), however it seems pretty positive and generally a good fit (~59% $Rˆ2%).

```
m5 <- lm(pf_religion ~ pf_religion_restrictions, data = hfi)
summary(m5)
```

```
##
## Call:
## lm(formula = pf_religion ~ pf_religion_restrictions, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9093 -0.5054  0.1795  0.5982  1.9138
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.62655    0.09810   36.97   <2e-16 ***
## pf_religion_restrictions  0.58473    0.01309   44.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8268 on 1362 degrees of freedom
##   (94 observations deleted due to missingness)
## Multiple R-squared:  0.5944, Adjusted R-squared:  0.5941
## F-statistic:  1996 on 1 and 1362 DF,  p-value: < 2.2e-16
```