Code ▾

Lab 8 - Introduction to Linear Regression

George Cruz

2020-12-07

Hide

```r
library(tidyverse)
library(openintro)
library(statsr)
```

```
## Warning: package 'statsr' was built under R version 4.0.3
```

Exercise 1

What are the dimensions of the dataset?

Hide

```r
dim(hfi)
```

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

The Dataset has 1458 rows and 123 columns.

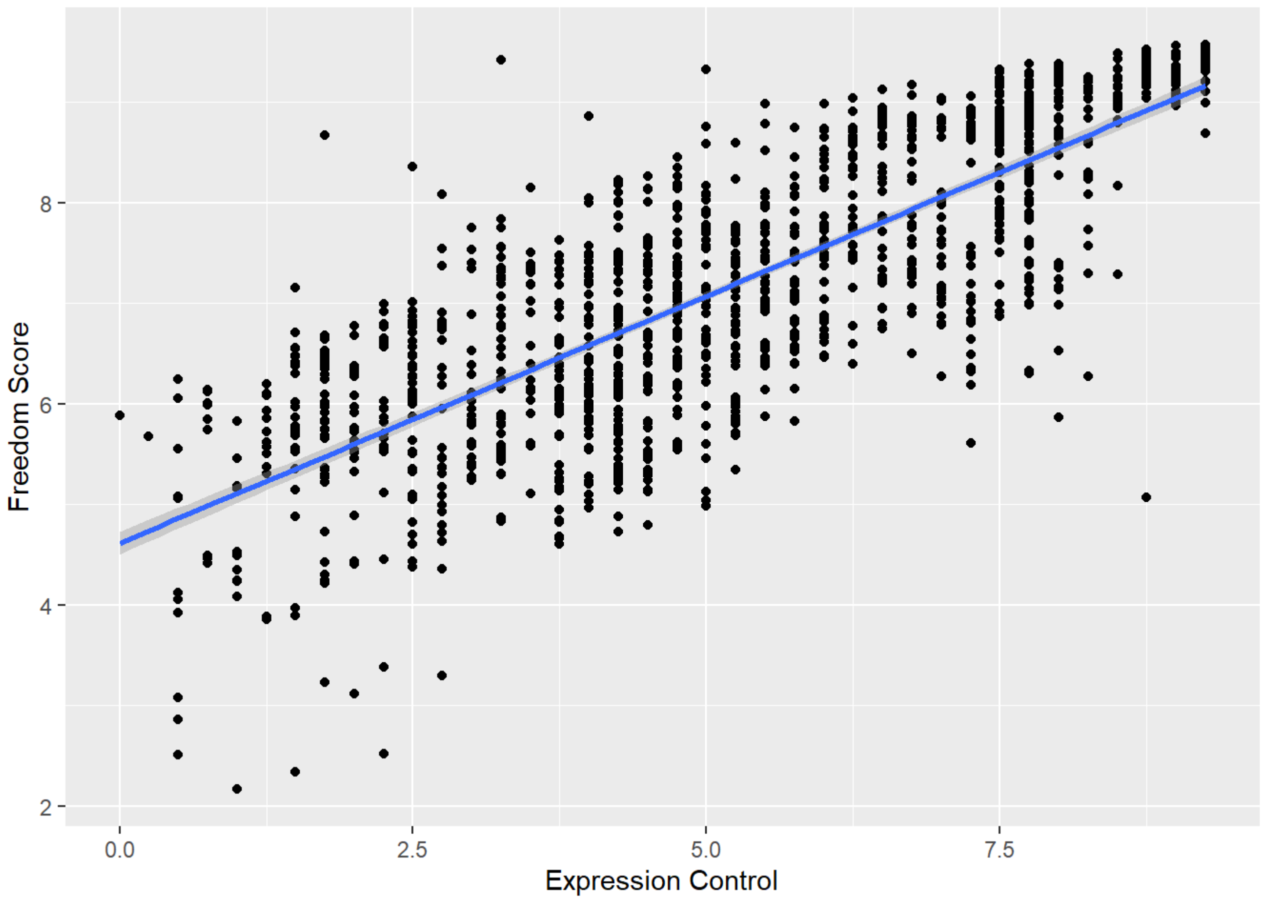
Exercise 2

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 is particularly useful.

Hide

```
ggplot(hfi, aes(x=pf_expression_control, y=pf_score)) +
  geom_point() +
  geom_smooth(method=lm) +
  xlab("Expression Control") +
  ylab("Freedom Score")
```



Because the relationship looks linear, we quantify the strength of the relationship with the correlation coefficient.

Hide

```
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

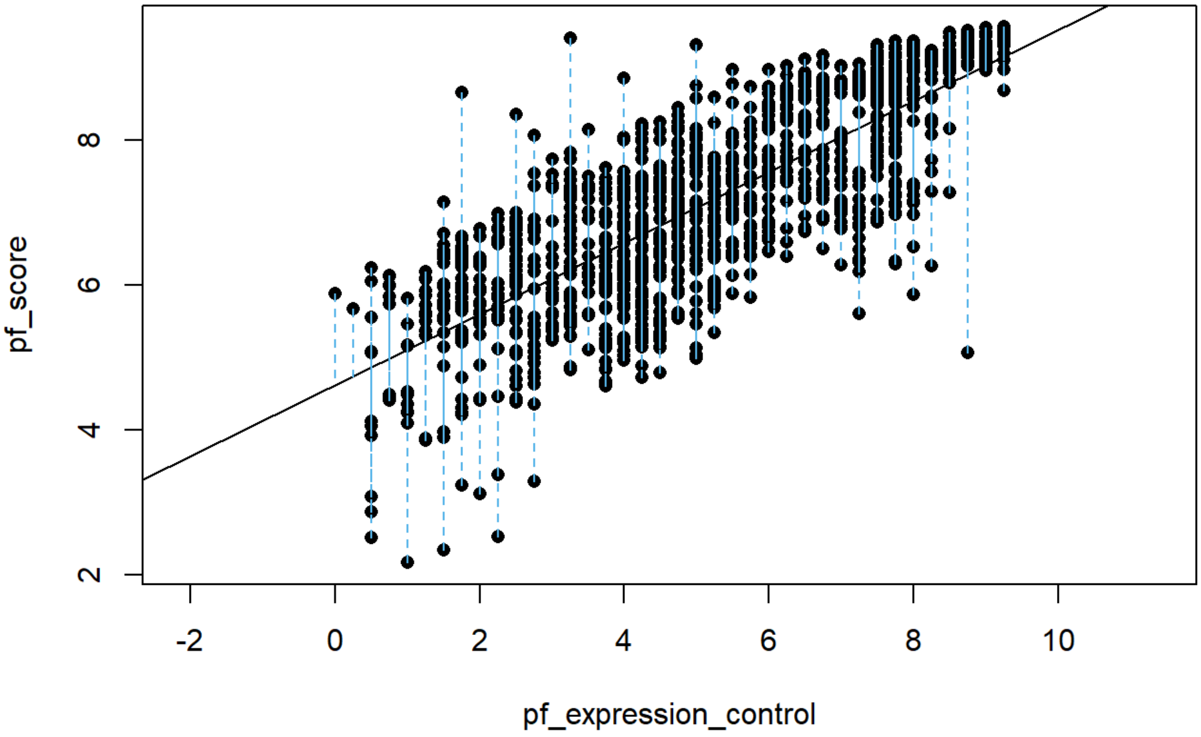
Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

The relationship appears to have a positive correlation, with the Freedom Score increasing as the Expression Control increases.

Hide

```
#this isn't working, wondering if we should just remove rows with NA

hf2 <- hfi %>% drop_na(pf_score) %>% drop_na(pf_expression_control)
plot_ss(x = pf_expression_control, y = pf_score, data = hf2)
```



```
## 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

I could not get the plot_ss to run interactively.

Hide

```
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
```

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?

Hide

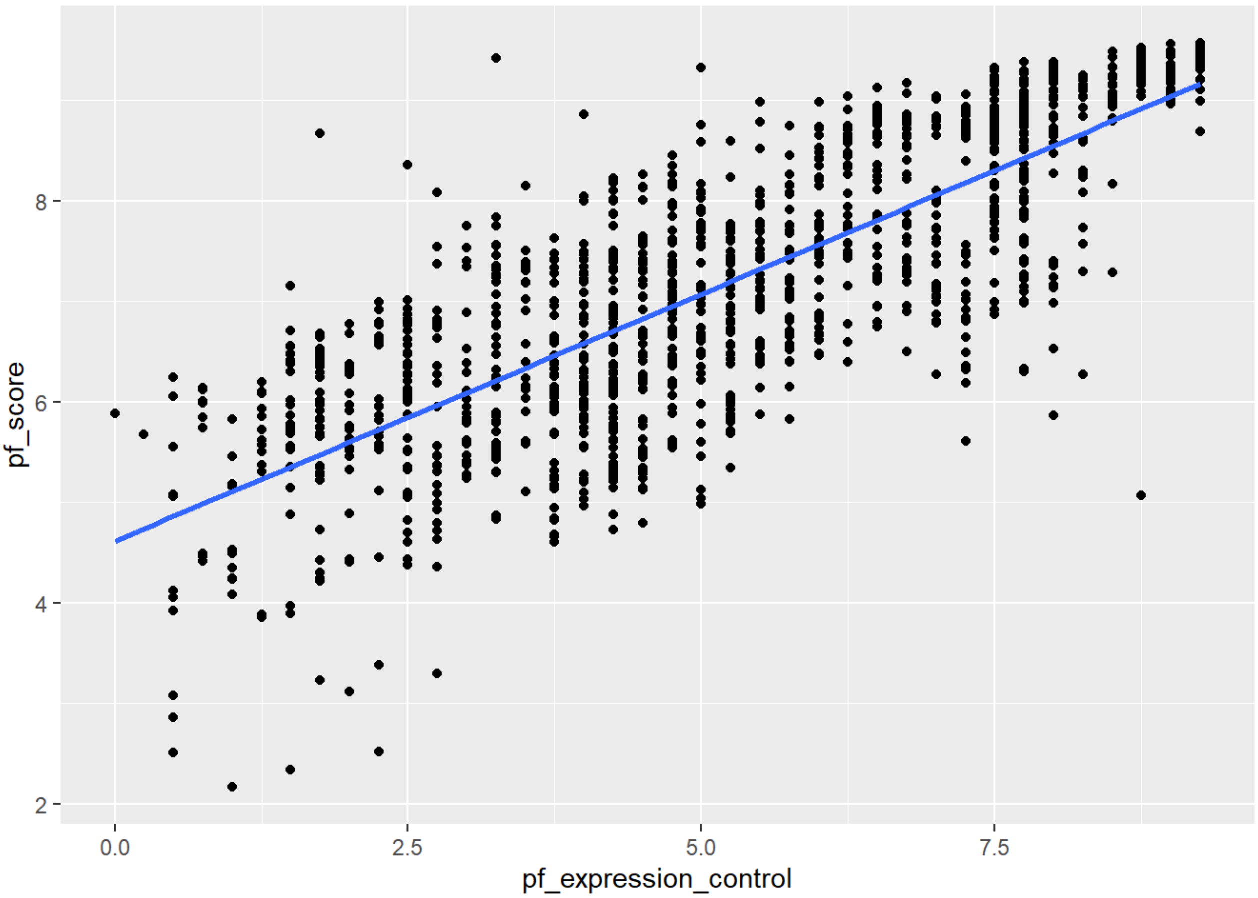
```r
m2 <- lm(hf_score ~ pf_expression_control, hfi)
summary(m2)
```

```
##
## 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
```

The slope tells us that human freedom increases 0.349 for every 1 point increase in pf_expression_control.

Hide

```r
ggplot(data = hfi, aes(x = pf_expression_control, y = pf_score)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```



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?

Hide

```
4.61707 + 0.49143 * 6.7
```

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

They would predict the personal freedom score to be 7.91

If we look at the actual value:

Hide

```
values <- hfi %>%
  select(pf_score, pf_expression_control) %>%
  filter(pf_expression_control >= 6.7 ) %>%
  filter(pf_expression_control <= 6.8 ) %>%
  summarise(avg = mean(pf_score))

values
```

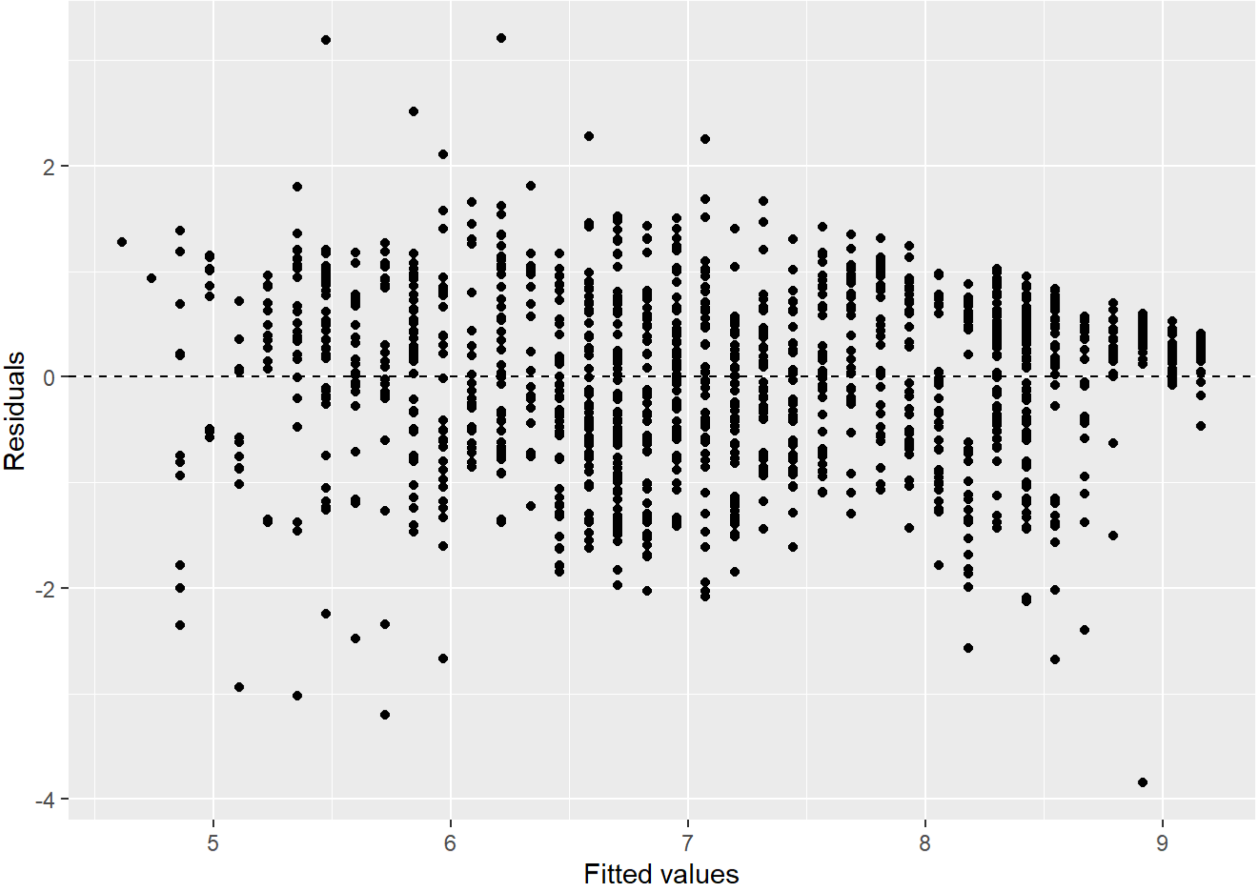
```
## # A tibble: 1 x 1
##     avg
##   <dbl>
## 1  8.01
```

I could not find a pf_expression_control of 6.70 so I averaged the 6.75 values.

I would say the prediction will be underestimated.

Hide

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



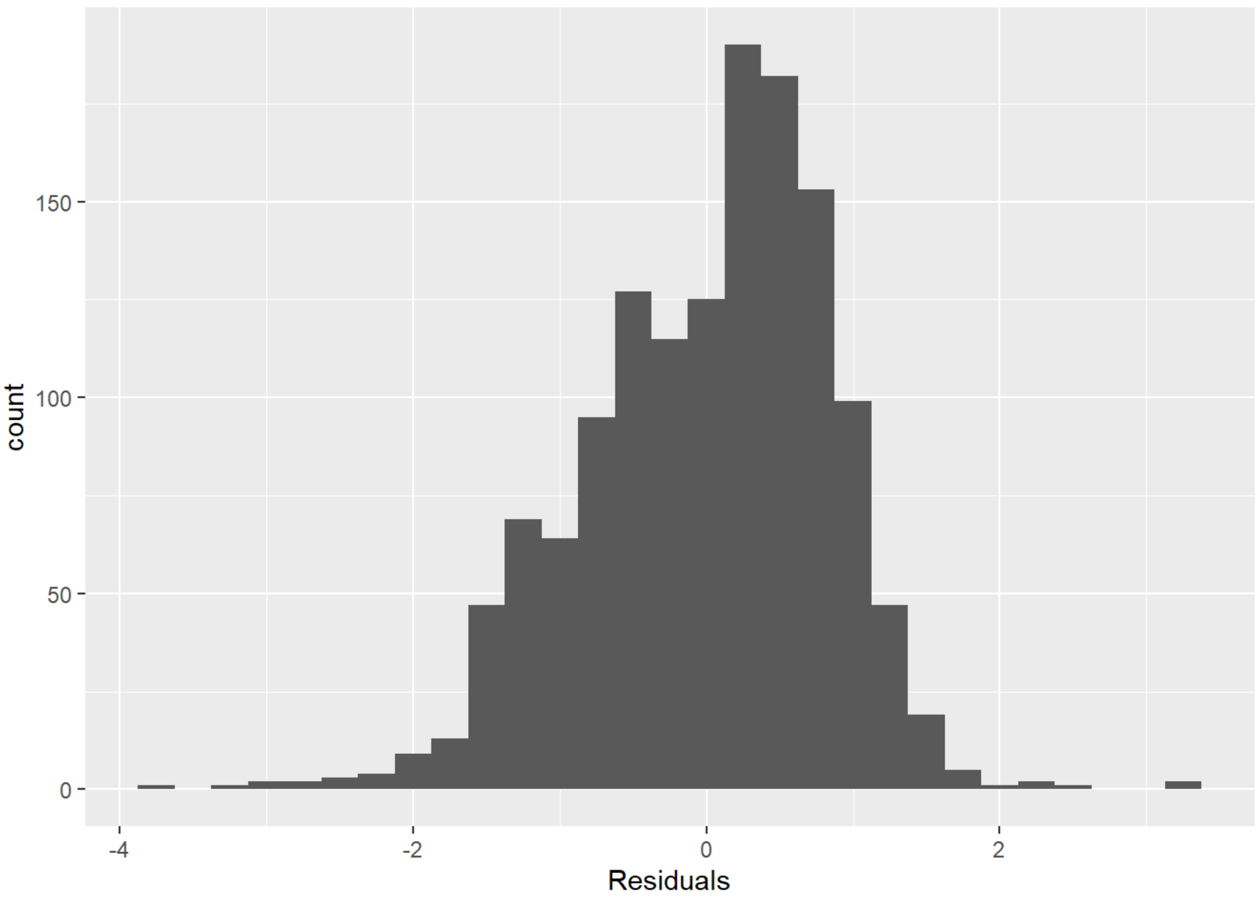
Exercise 7

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

There is not an apparent pattern in the residuals plot although they appear to be distributely almost equally above and belowe the line.

Hide

```
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(binwidth = 0.25) +
  xlab("Residuals")
```



When using binwidth 25 as specified in the lab, the histogram is distorted. With a bin width 0f 0.25 we get a nearly normal distribution.

Exercise 8

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

Yes.

Exercise 9

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

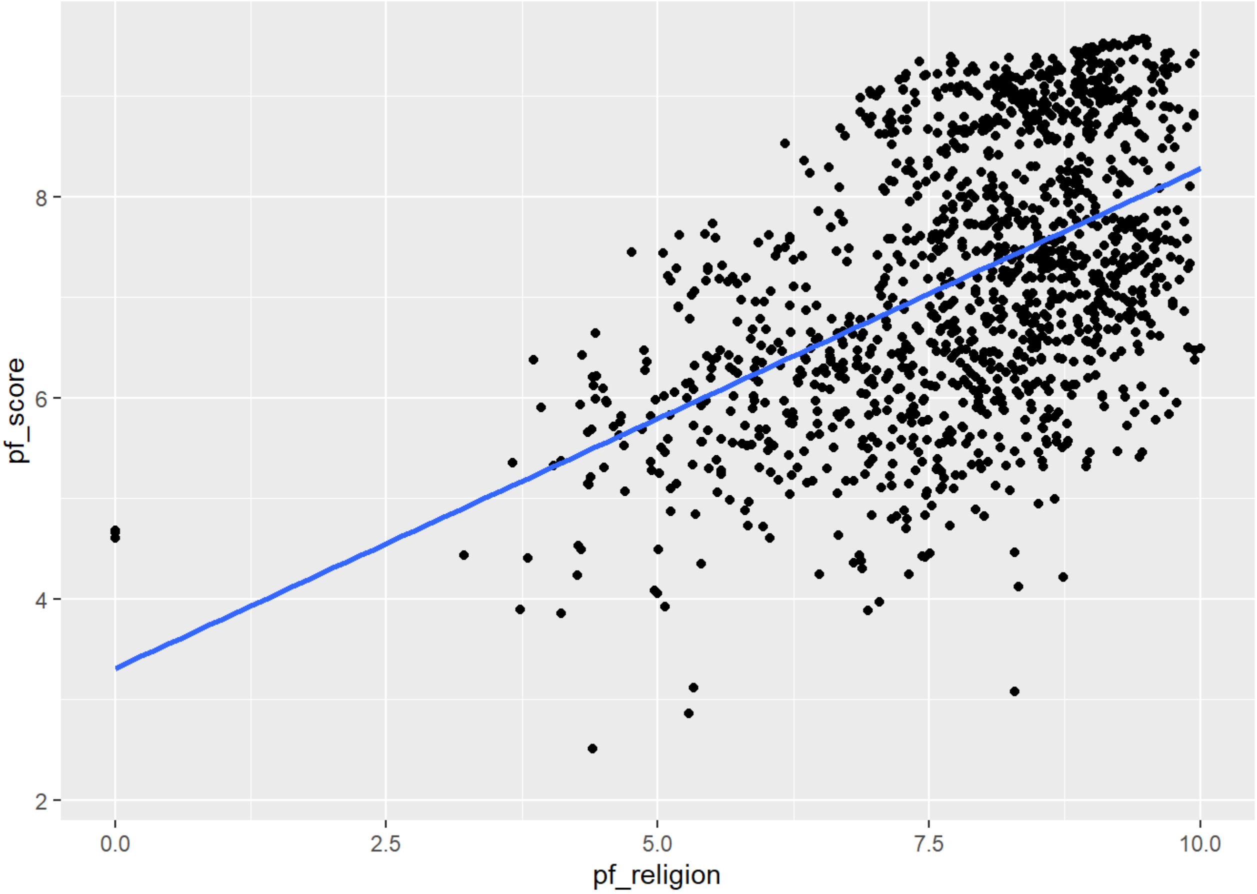
Yes.

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?

Hide

```
ggplot(data = hfi, aes(x = pf_religion, y = pf_score)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```



There appears to be a positive correlation especially on the countries with high degrees of religious freedom.

Exercise 11

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?

Hide

```
m3 <- lm(pf_score ~ pf_religion, hfi)
summary(m3)
```

```
##
## Call:
## lm(formula = pf_score ~ pf_religion, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3515 -0.8728 -0.0744  1.0532  2.3520
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.30981    0.18744   17.66   <2e-16 ***
## pf_religion  0.49689    0.02346   21.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.17 on 1366 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.2472, Adjusted R-squared:  0.2467
## F-statistic: 448.6 on 1 and 1366 DF,  p-value: < 2.2e-16
```

The R^2 of this relationship is 24.67% vs the 57.72% of the pf_expression_control.

Exercise 12

I was surprised that the relationship between religious freedom and pf_score was not, at least, as strong as the relationship between pf_expression_control and pf_score.

…