Lab 8 - Linear Regressions

Keith Colella

2023-04-23

Setup

Config

```
library(tidyverse)
library(openintro)
```

Data

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

Exercise 1

Question

What are the dimensions of the dataset?

Response

```
nrow(hfi)
```

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

```
ncol(hfi)
```

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

The dataset has 123 variables (columns) with 1458 observations (rows).

Exercise 2

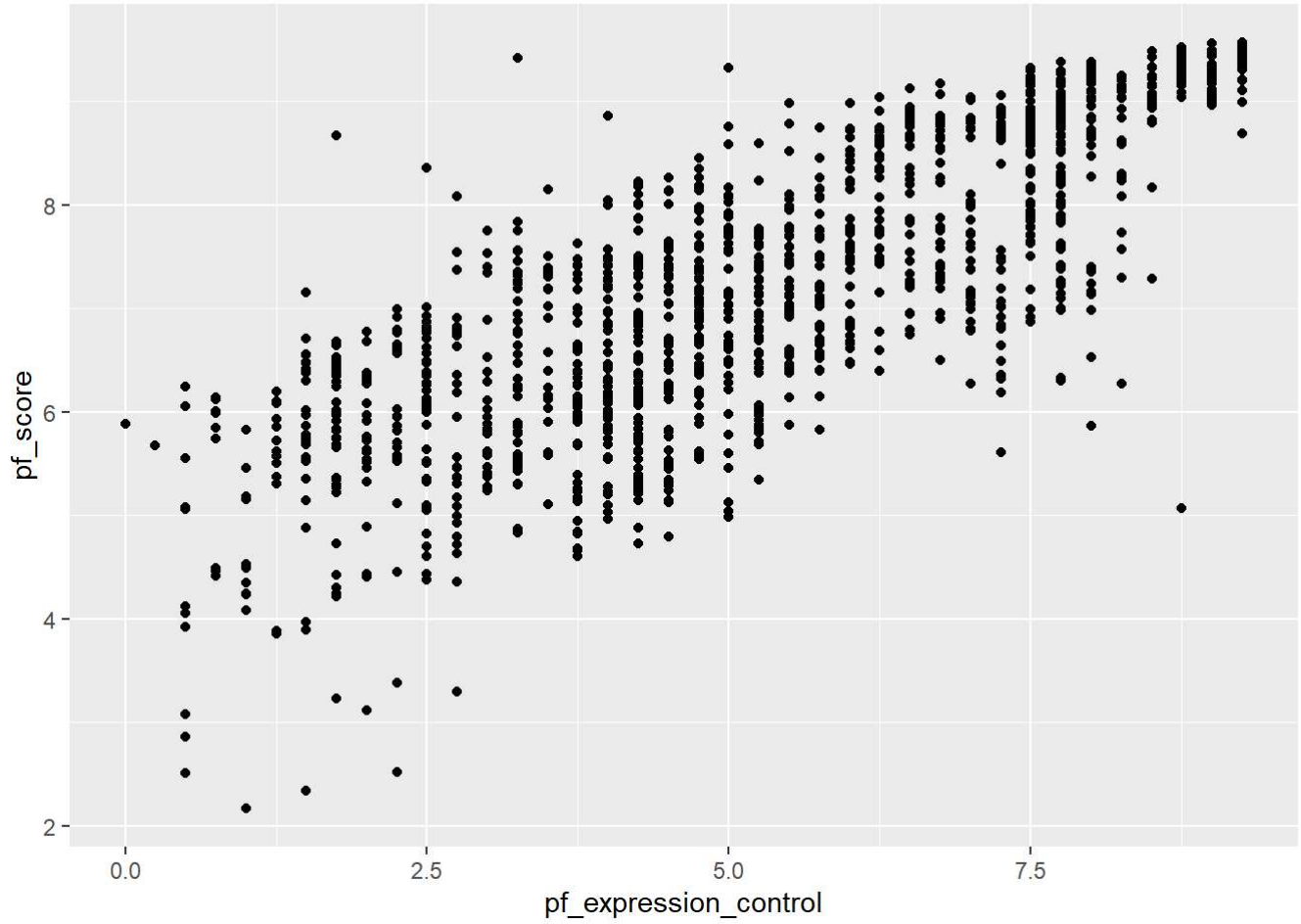
Question

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?

Response

```
ggplot(hfi, aes(pf_expression_control, pf_score)) +
  geom_point()
```

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



A scatter plot is a helpful way to explore potential linear relationships between two variables. Typically, the response variable takes the y axis, and the predictor takes the x axis. There does appear to be a reasonably linear, positive relationship between the personal freedom score and the expression control score. So, a linear model may be appropriate to predict `pf_score` based on `pf_expression_control`.

Exercise 3

Question

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.

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

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

Response

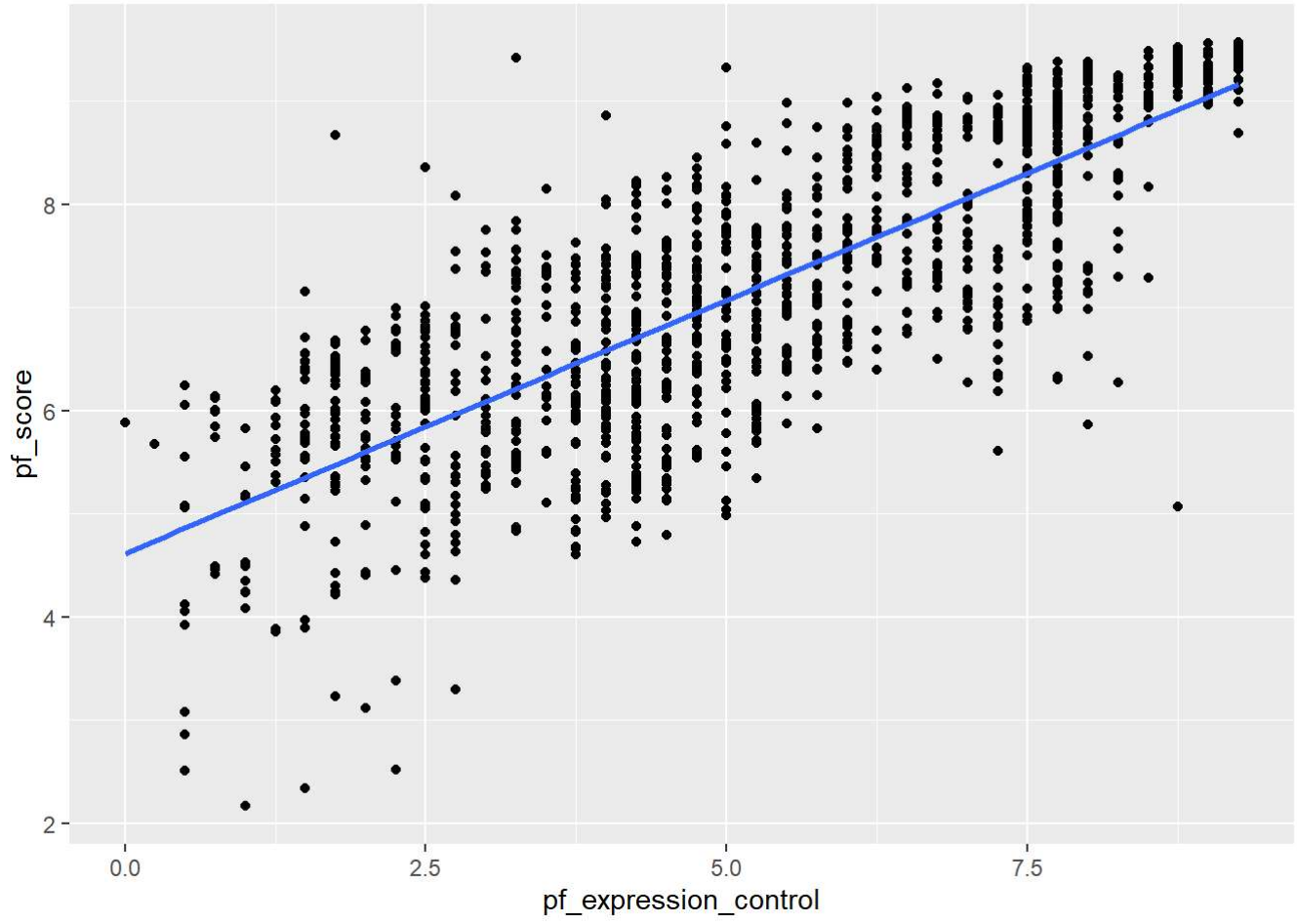
The relationship appears positive. That is, as `pf_expression_control` increases, so does `pf_score`. The relationship appears relatively strong, as indicated by the correlation coefficient of ~0.80. There is, however, significant spread around any best fit line. As such, I would expect a linear regression to have a relative low R^2.

```
ggplot(hfi, aes(pf_expression_control, pf_score)) +
  geom_point() +
  geom_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()`).
```

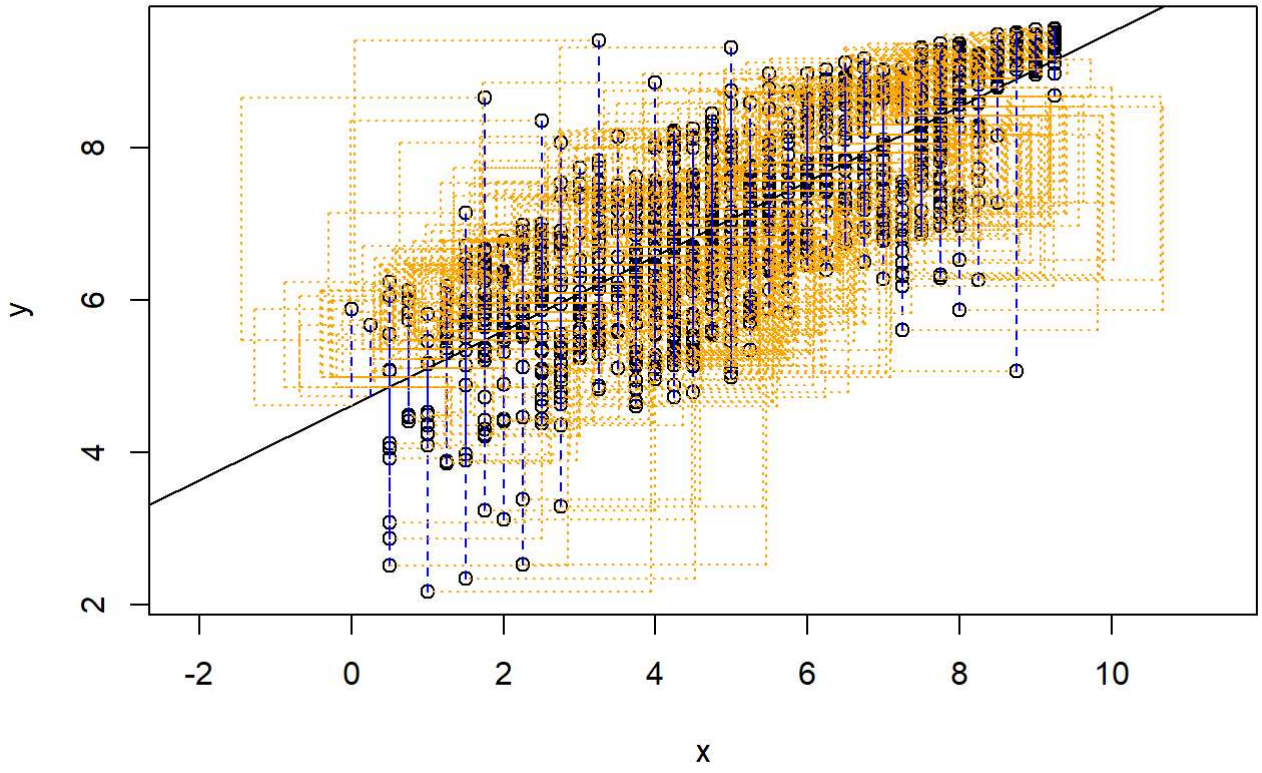


Exercise 4

Question

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?

```
hfi_clean <- hfi %>% filter(!is.na(pf_score))

DATA606::plot_ss(x = hfi_clean$pf_expression_control,
                 y = hfi_clean$pf_score,
                 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
```

Response

The lowest sum of squares I got when manually choosing two points was ~5068, with ann intercept of 4.9872 and a β of 0.4861. As indicated above, however, the true minimized sum of squares with the best fit line is much lower.

Exercise 5

Question

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?

Response

```
model_fit <- lm(hf_score ~ pf_expression_control, data = hfi)

summary(model_fit)
```

```
##
## 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 positive slope indicates a positive correlation between `pf_expression_control` and `hf_score`. In other words, as `pf_expression_control` increases, so does `hf_score`. So, countries with less political pressure on media content tend to have high total human freedom scores.

Exercise 6

Question

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?

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

Response

```
predict(m1, newdata = data.frame(pf_expression_control = 6.7))
```

```
##        1
## 7.909663
```

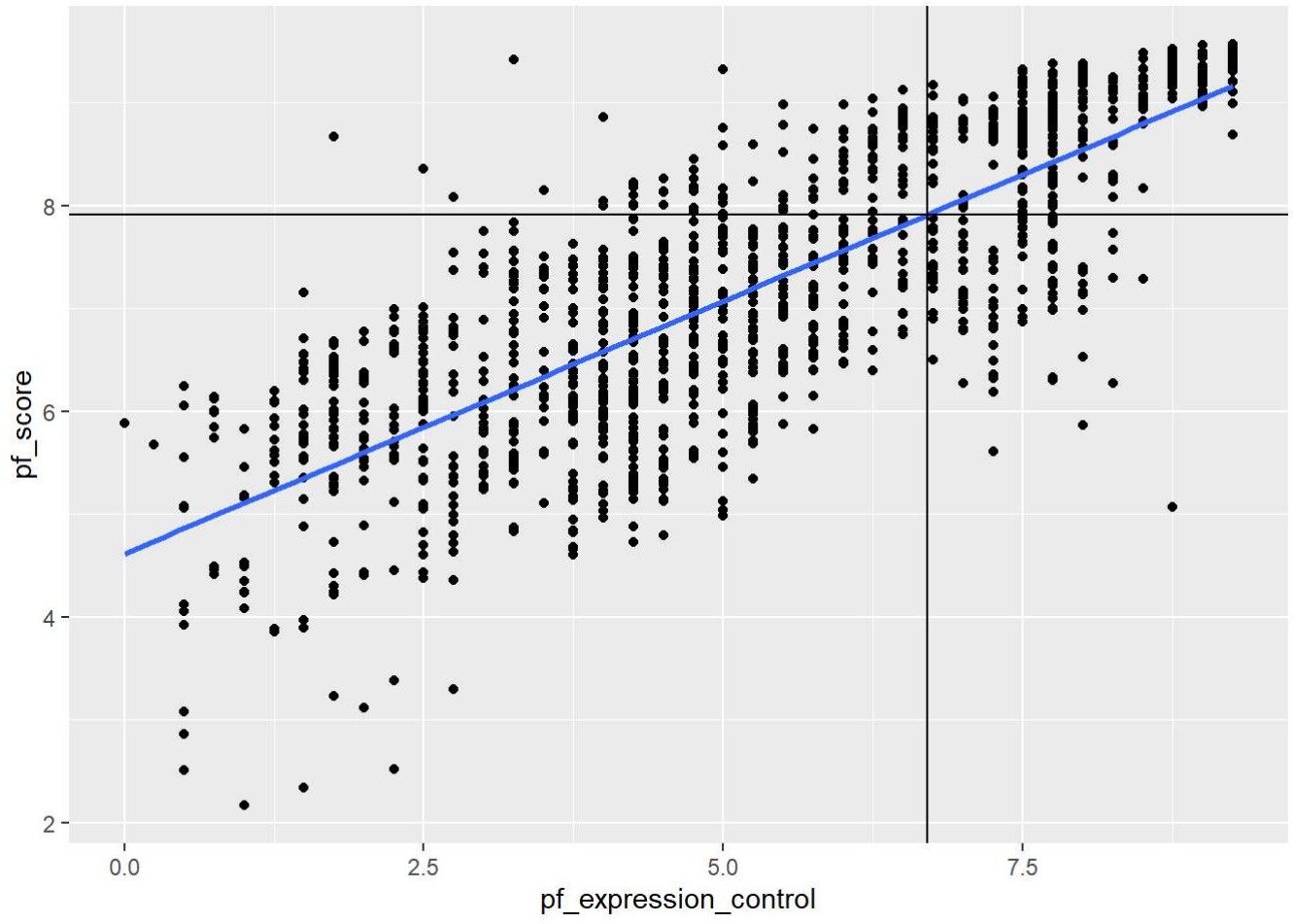
The model would predict a `pf_score` of ~7.91. We can pinpoint this location on the plot below.

```
ggplot(hfi, aes(pf_expression_control, pf_score)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  geom_hline(yintercept = 7.91) +
  geom_vline(xintercept = 6.70)
```

```
## `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()`).
```



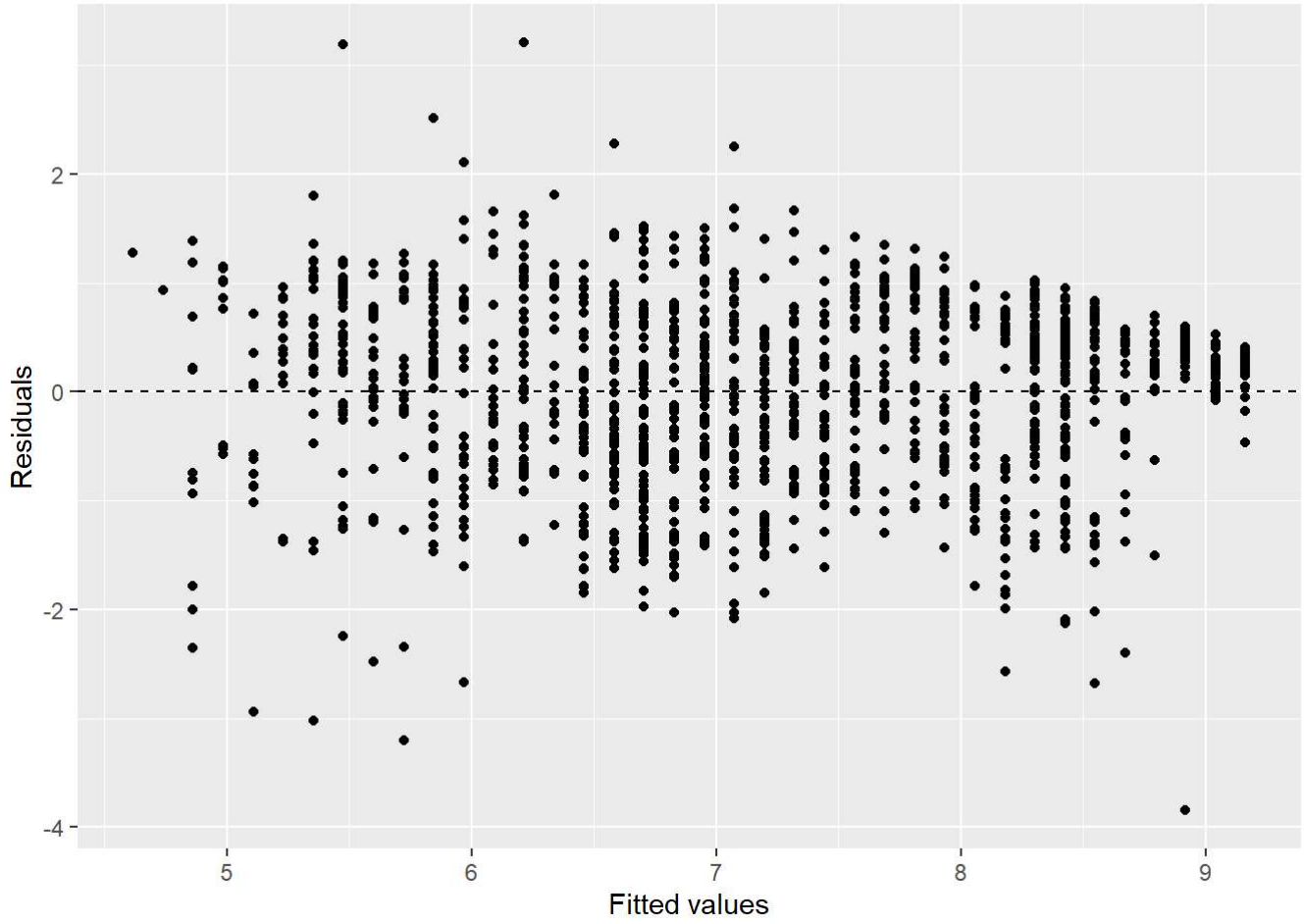
It appears there are observations both above and below the line at this point. So, while there is some error associated with this prediction, we cannot confidently say whether it is an over or underprediction.

Exercise 7

Question

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

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



Response

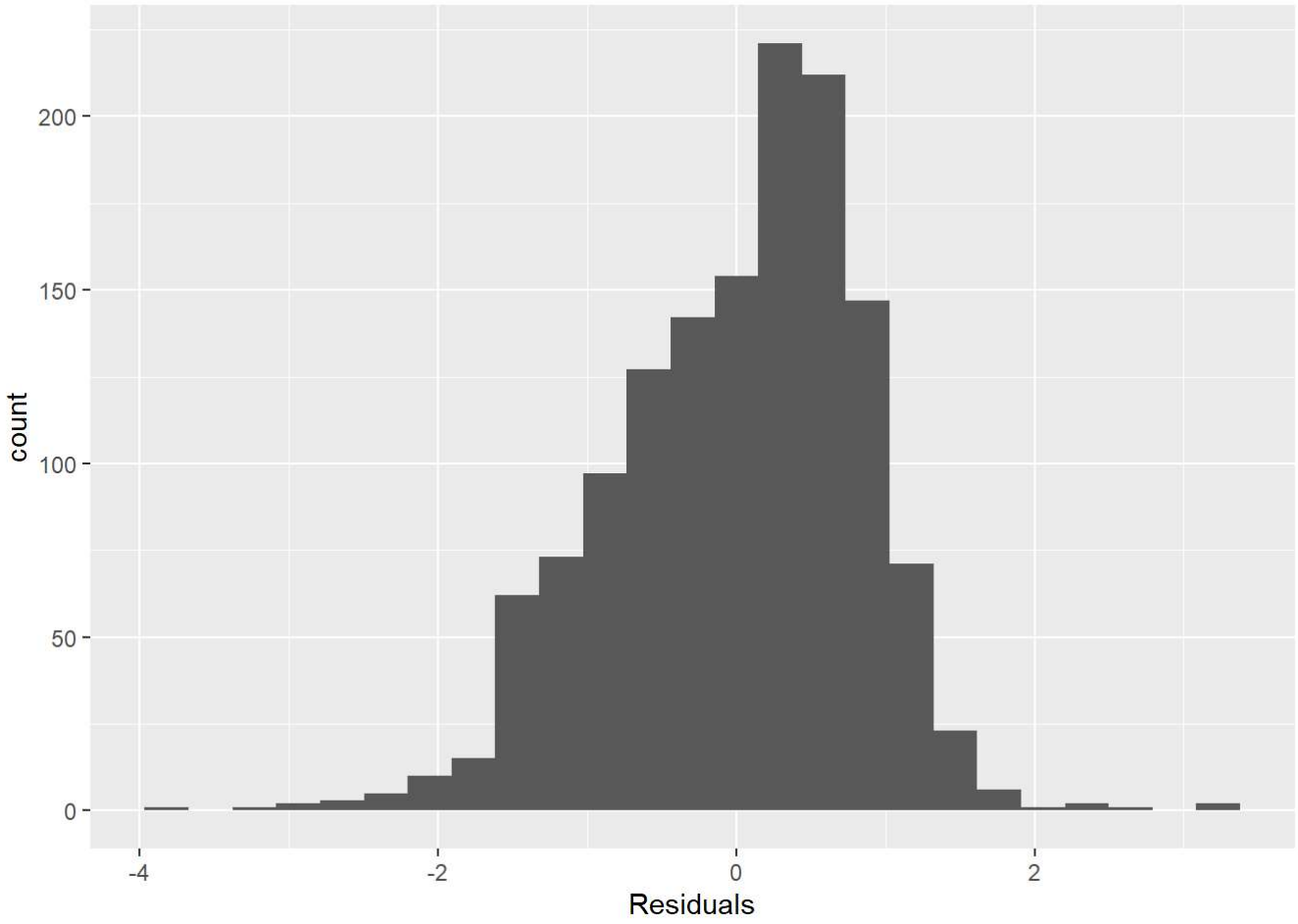
The residuals appear evenly distributed above and below zero, with no apparent pattern. This lends support to the notion that the relationship between the two variables is indeed linear.

Exercise 8

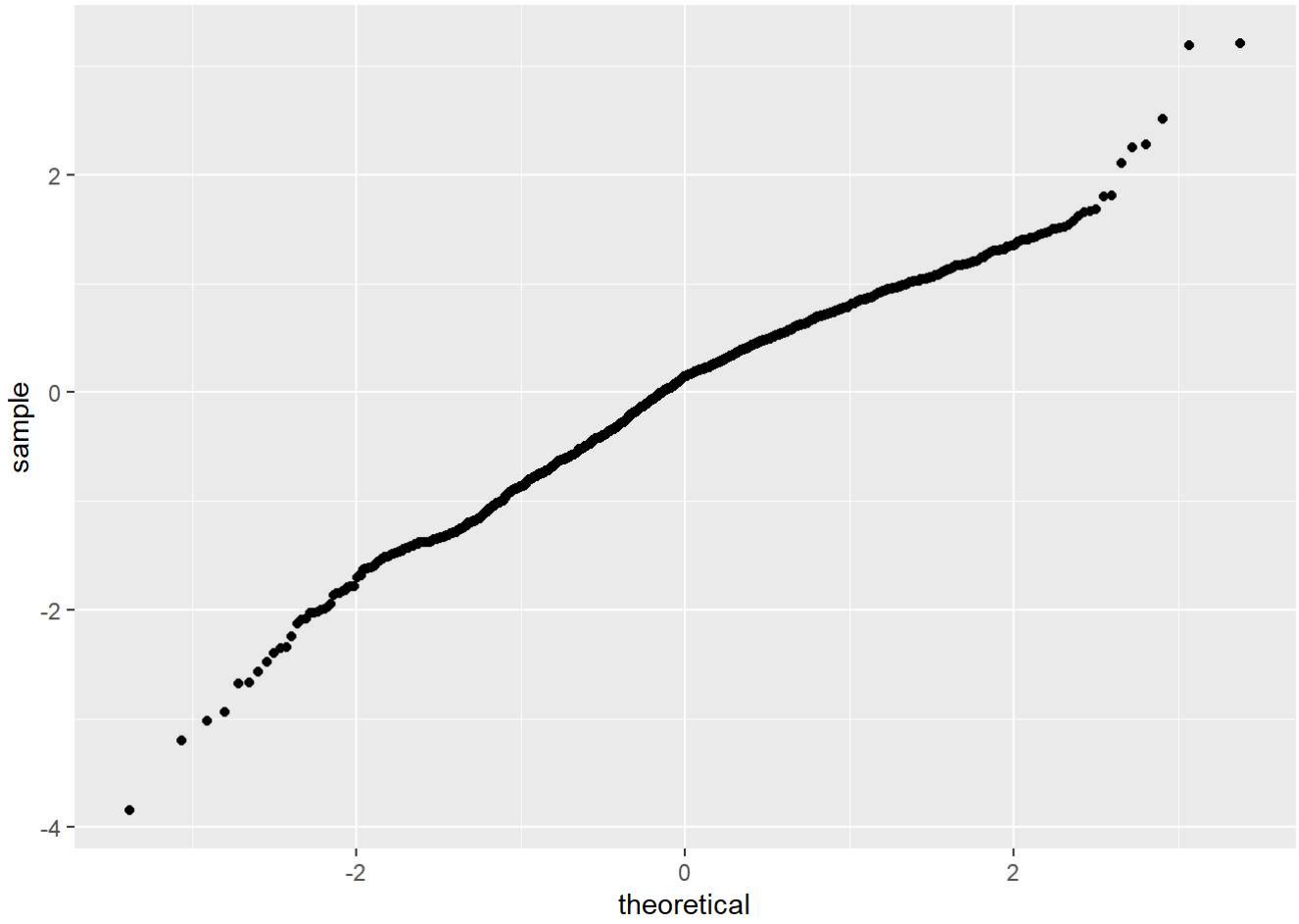
Question

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

```
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(bins = 25) +
  xlab("Residuals")
```




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



Response

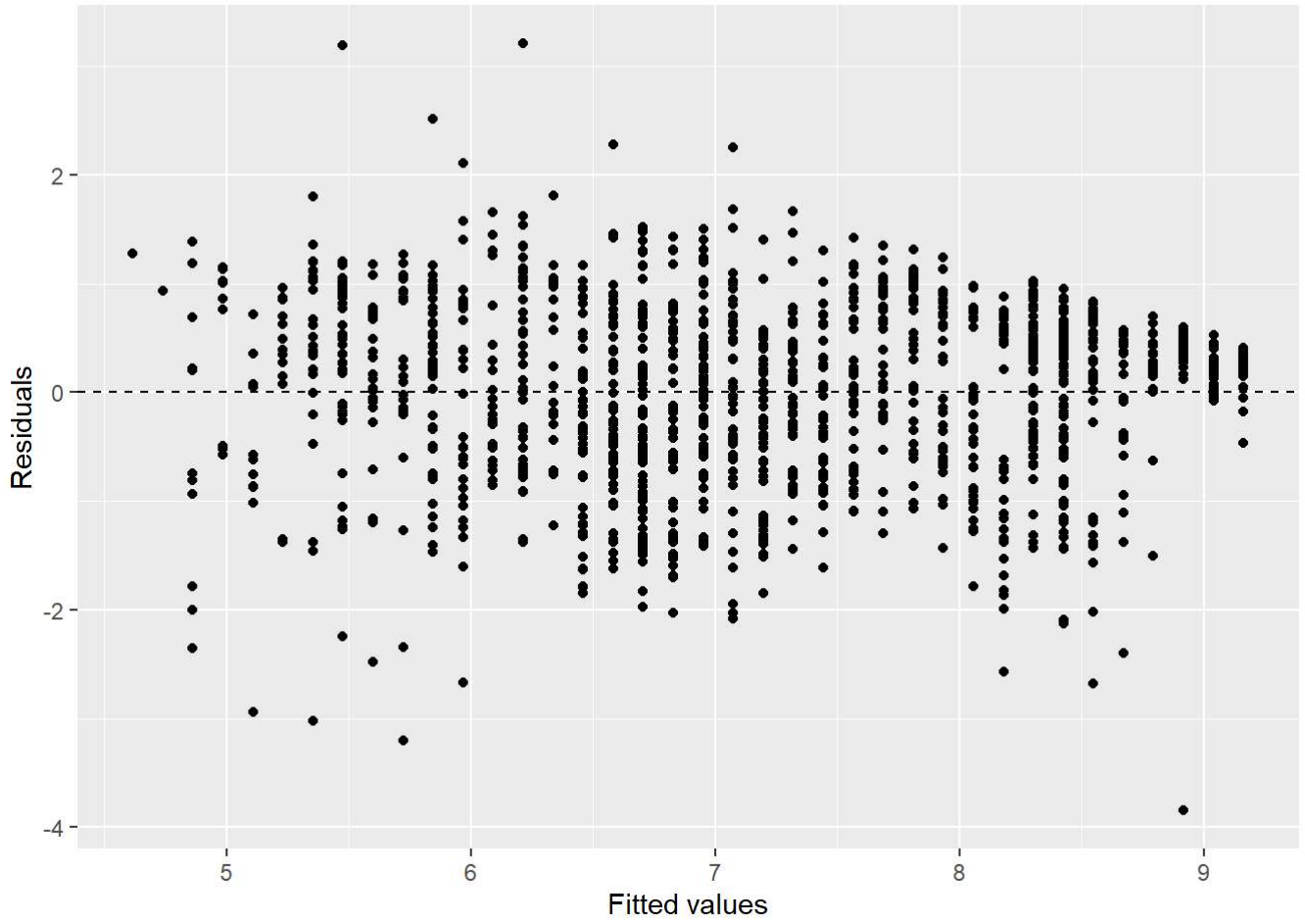
The histogram appears almost normal and centered on zero, though there is a slight negative skew, and some potential outliers in the right tail. This is confirmed in the qq plot. Both plots, however, indicate sufficient normality in terms of residual analysis.

Exercise 9

Question

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

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



Response

Variance of residuals appears sufficiently constant to meet the condition.

Exercise 10

Question

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?

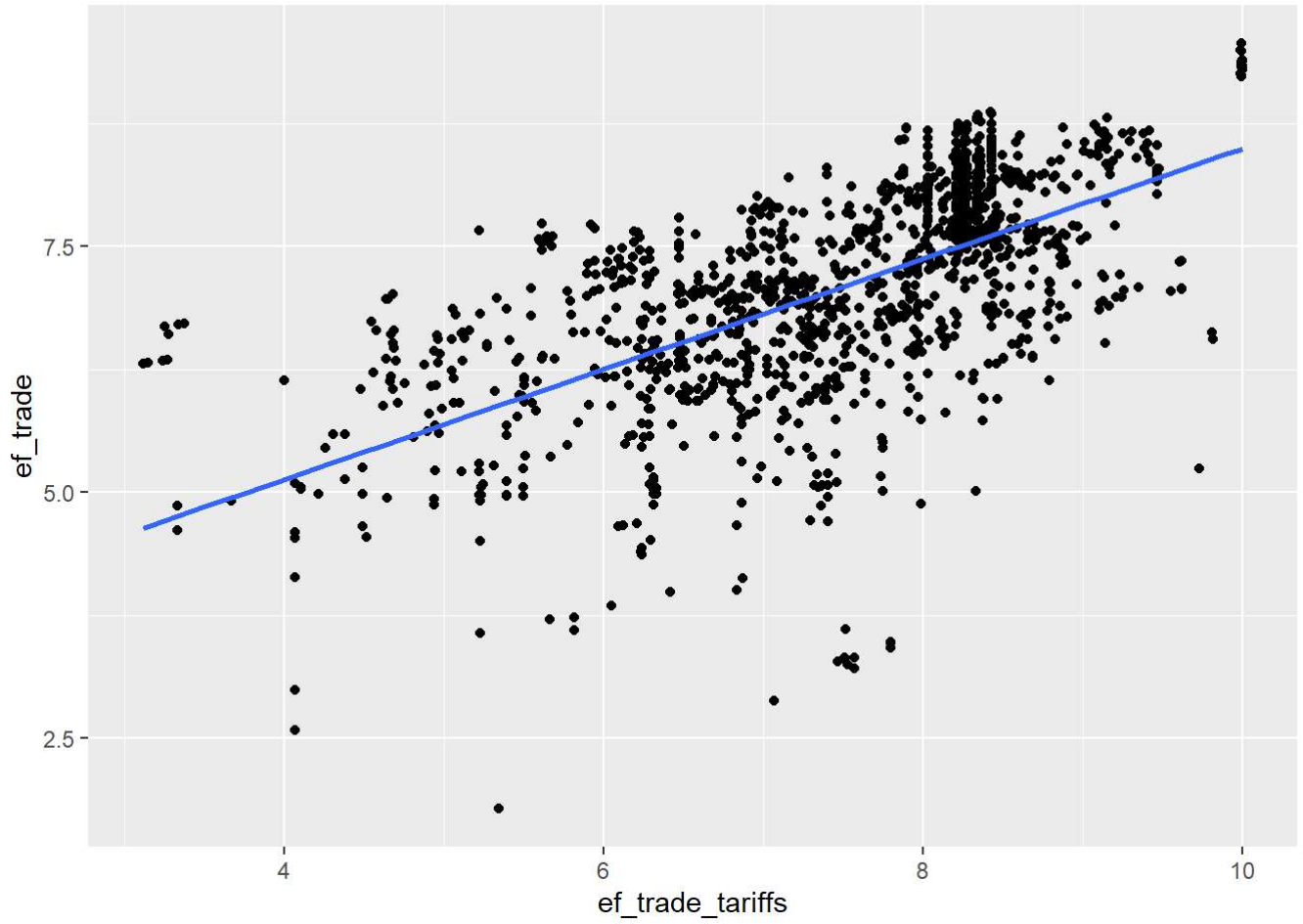
Response

```
ggplot(hfi, aes(ef_trade_tariffs, ef_trade)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```

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

```
## Warning: Removed 85 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 85 rows containing missing values (`geom_point()`).
```



There does indeed appear to be a linear relationship between a country's economic freedom trade score and its tariffs score.

Exercise 11

Question

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?

Response

```
ef_model <- lm(ef_trade ~ ef_trade_tariffs, data = hfi)

cat('EF model results',
    '\nR-squared: ', summary(ef_model)$r.squared,
    '\nAdjusted R-squared: ', summary(ef_model)$adj.r.squared,
    '\nPF model results',
    '\nR-squared: ', summary(m1)$r.squared,
    '\nAdjusted R-squared: ', summary(m1)$adj.r.squared
)
```

```
## EF model results
## R-squared:  0.3813893
## Adjusted R-squared:  0.3809381
## PF model results
## R-squared:  0.6342361
## Adjusted R-squared:  0.6339702
```

The `pf_score` model has a significantly higher R^2, indicating that `pf_expression_control` is a better linear predictor of `pf_score` than `ef_trade_tariffs` is for `ef_trade`.

Exercise 12

Question

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

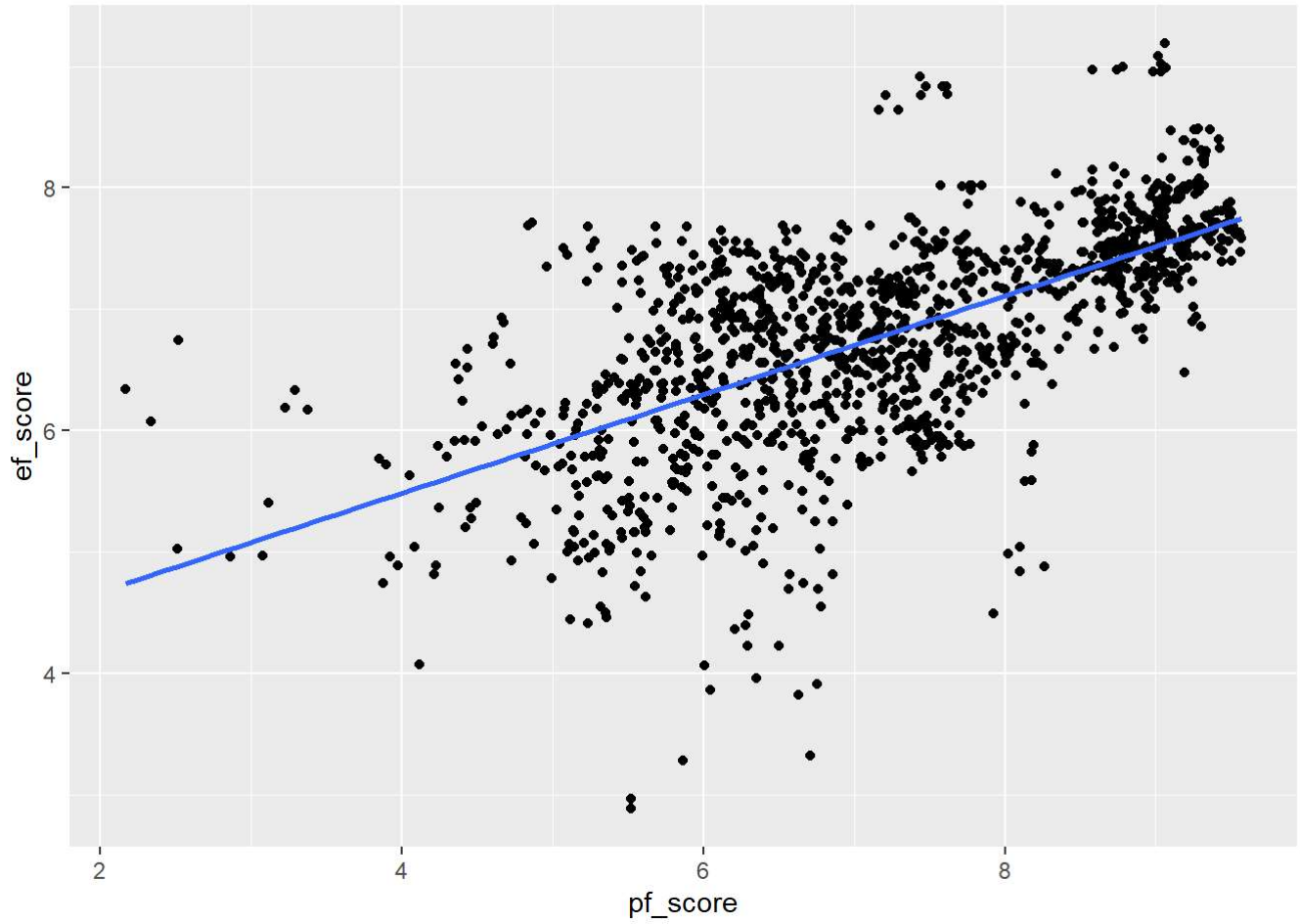
Response

```
ggplot(hfi, aes(pf_score, ef_score)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```

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

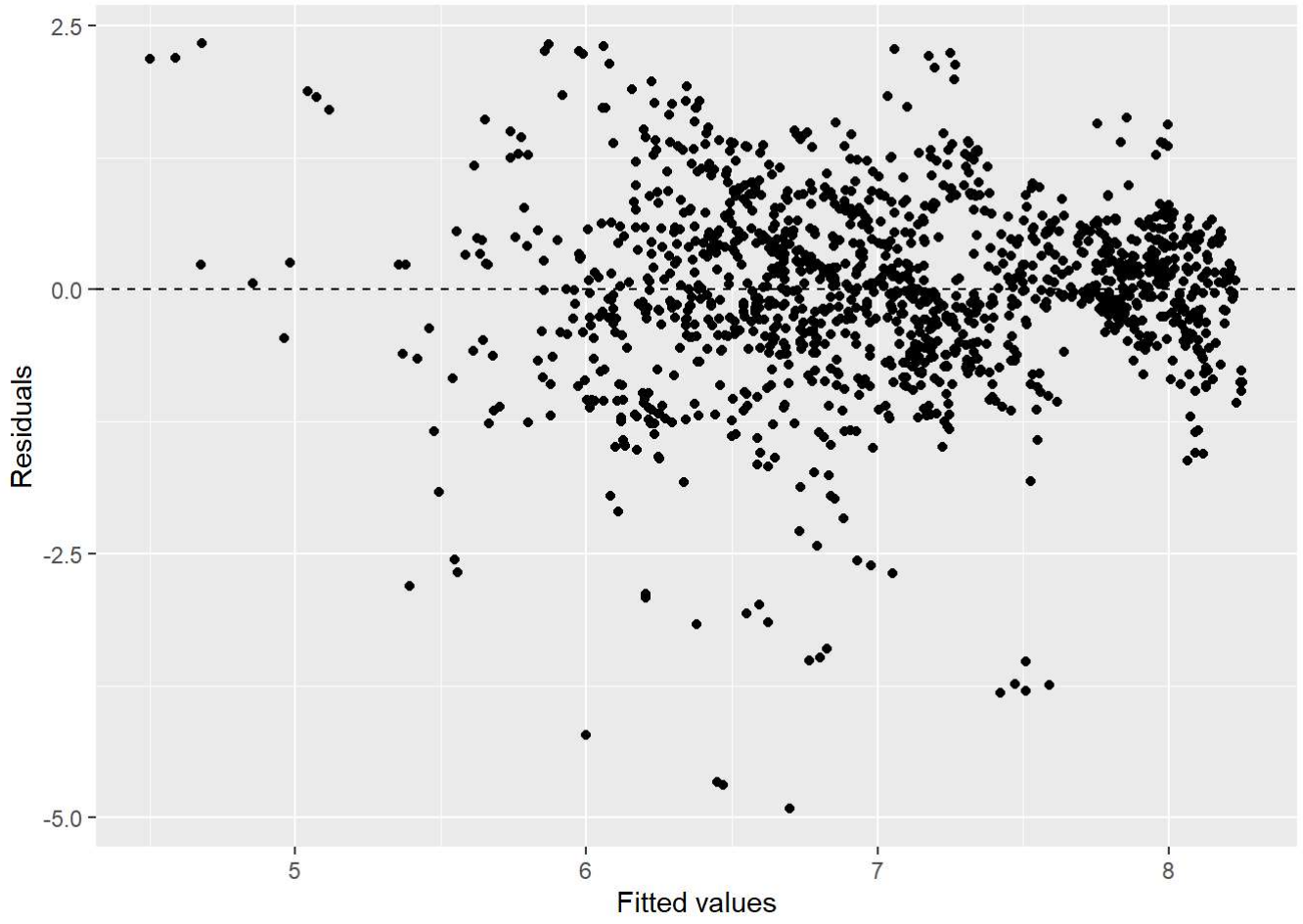
```
## Warning: Removed 80 rows containing non-finite values (`stat_smooth()`).
```

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

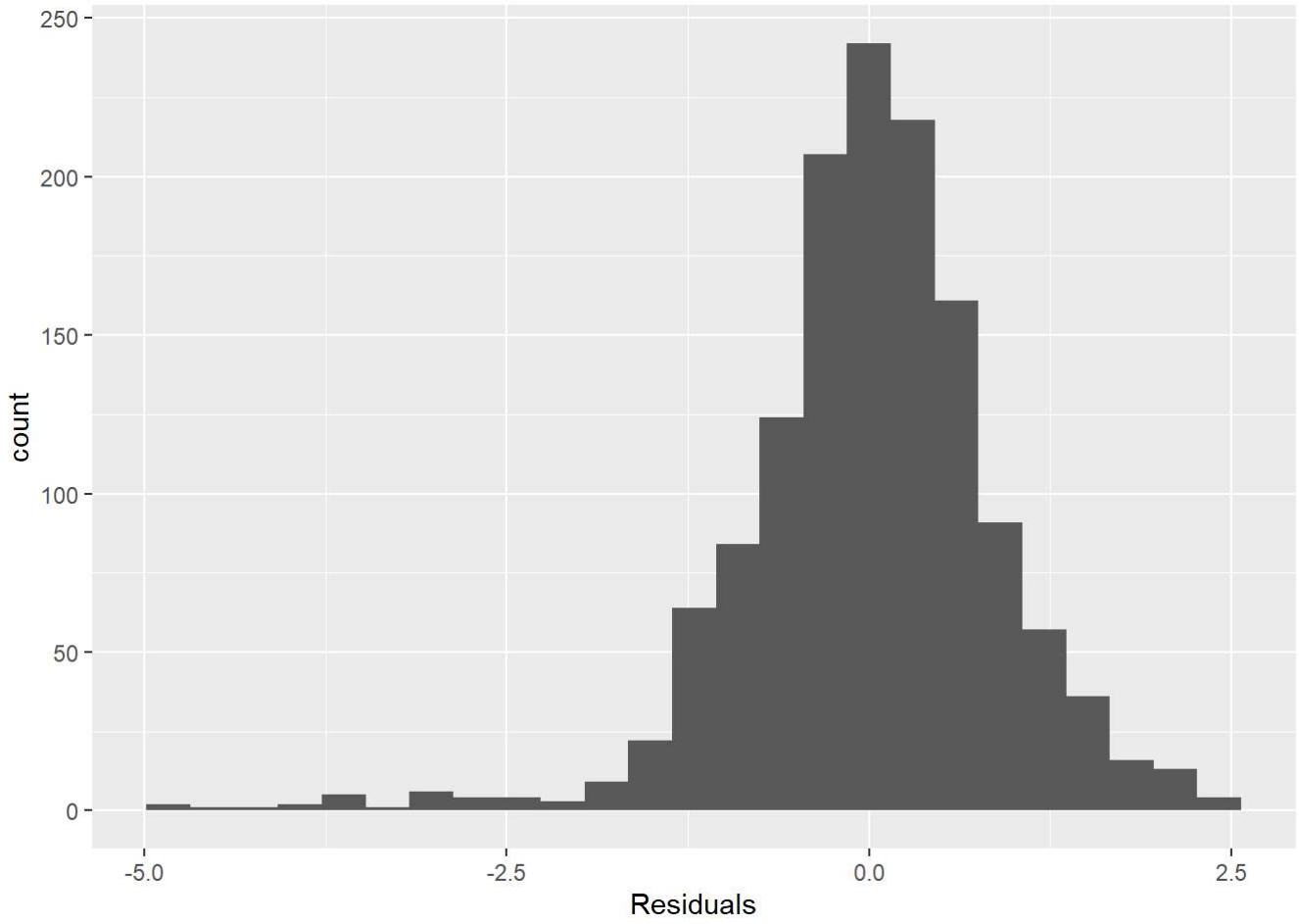


I was surprised that the EF and PF scores were not more tightly correlated. Yes, there is a clear linear relationship between the two variables, as shown by the diagnostics below.

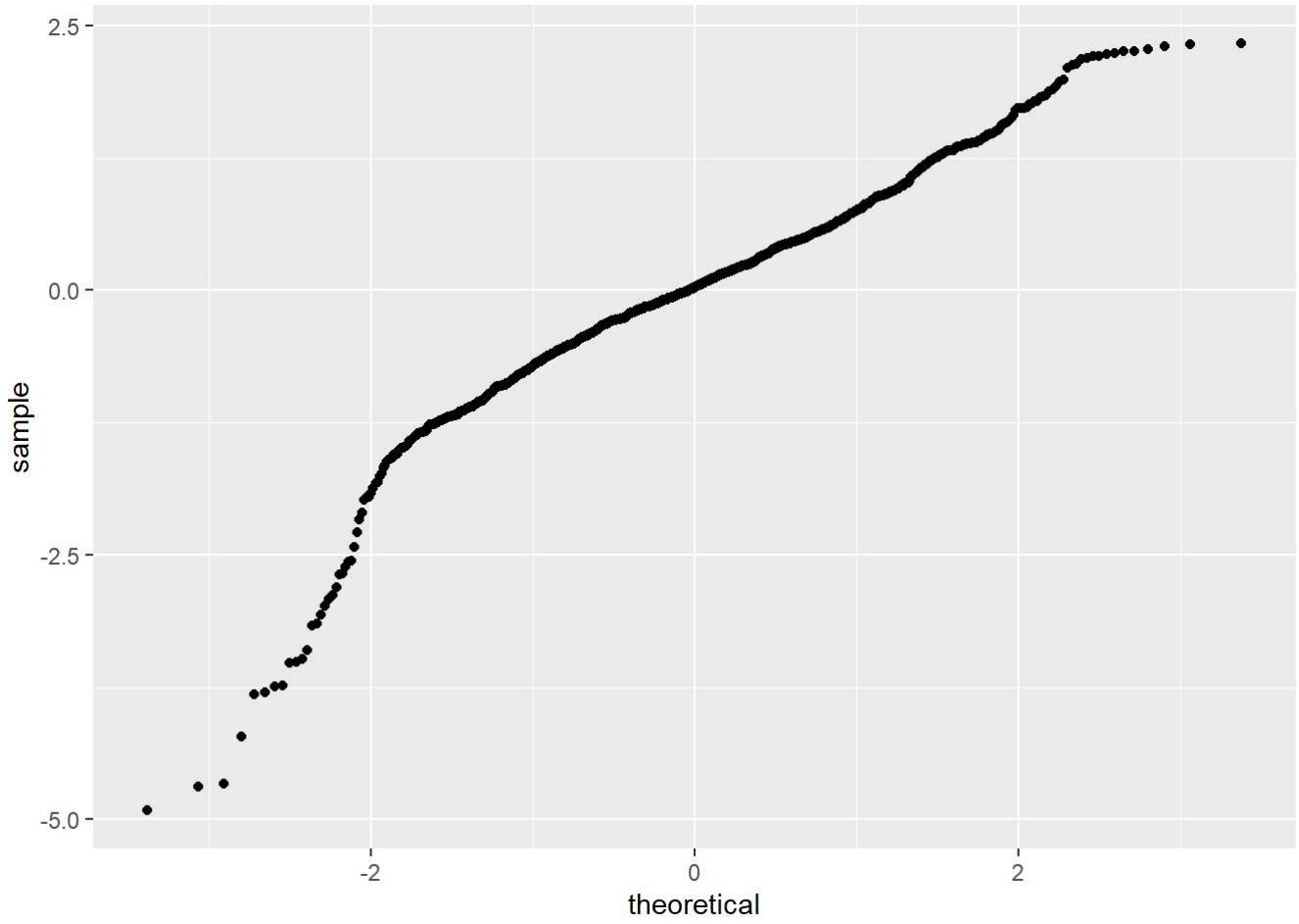
```
ef_pf_model <- lm(ef_trade ~ pf_score, data = hfi)

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




```
ggplot(data = ef_pf_model, aes(x = .resid)) +
  geom_histogram(bins = 25) +
  xlab("Residuals")
```




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



Residuals appear relatively normal and centered on zero, despite some outliers in the left tail. Variability also appears sufficiently constant, and no pattern emerges in the residual plot to indicate a non-linear relationship. So, we can conclude a linear relationship exists. What's surprising is the relatively poor performance of the model, as indicated by the low R^2.

```
summary(ef_pf_model)
```

```
##
## Call:
## lm(formula = ef_trade ~ pf_score, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9179 -0.4320  0.0259  0.4906  2.3310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.40417    0.12621   26.97   <2e-16 ***
## pf_score     0.50679    0.01722   29.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8783 on 1375 degrees of freedom
##   (81 observations deleted due to missingness)
## Multiple R-squared:  0.3866, Adjusted R-squared:  0.3862
## F-statistic: 866.6 on 1 and 1375 DF,  p-value: < 2.2e-16
```

I expected the two scores to be very tightly correlated, but the low R^2 indicates otherwise. I suspect the relationship is complicated by states that have worked to create pro-business environments while still enforcing tight restrictions on the political freedoms of its citizens (e.g. Gulf states, Hong Kong, Singapore).

```
hfi %>%
  mutate(pf_ef_diff = pf_score - ef_score) %>%
  filter(year == 2016) %>%
  arrange(pf_ef_diff) %>%
  select(countries, pf_score, ef_score, pf_ef_diff) %>%
  head(10)
```

```
## # A tibble: 10 × 4
##    countries            pf_score ef_score pf_ef_diff
##    <chr>                   <dbl>    <dbl>      <dbl>
##  1 Yemen, Rep.              2.17     6.34      -4.17
##  2 Syria                    2.51     5.02      -2.51
##  3 United Arab Emirates     5.07     7.5       -2.43
##  4 Iraq                     3.12     5.4       -2.28
##  5 Brunei Darussalam        4.66     6.93      -2.27
##  6 Saudi Arabia             4.44     6.52      -2.08
##  7 Gambia, The              5.30     7.34      -2.04
##  8 Qatar                    5.53     7.49      -1.96
##  9 Egypt                    3.89     5.72      -1.83
## 10 Burundi                  4.41     5.92      -1.51
```

My above prediction appears partially true, but there appear to be more prominent factors driving discrepancies between political and economic freedoms scores. Conflict appears to be key, as evidenced by the fact that Yemen, Syria and Iraq (all of which experienced conflict in 2016) have the greatest discrepancies. As I predicted, we do also see some states that have a heavy pro-business lean coupled with tight cultural restrictions (e.g. UAE, Saudi, Qatar).