Introduction to linear regression

Souleymane Doumbia

The Human Freedom Index is a report that attempts to summarize the idea of "freedom" through a bunch of different variables for many countries around the globe. It serves as a rough objective measure for the relationships between the different types of freedom - whether it's political, religious, economical or personal freedom - and other social and economic circumstances. The Human Freedom Index is an annually co-published report by the Cato Institute, the Fraser Institute, and the Liberales Institut at the Friedrich Naumann Foundation for Freedom.

In this lab, you'll be analyzing data from Human Freedom Index reports from 2008-2016. Your aim will be to summarize a few of the relationships within the data both graphically and numerically in order to find which variables can help tell a story about freedom.

Getting Started

Load packages

In this lab, you will explore and visualize the data using the **tidyverse** suite of packages. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

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

The data

The data we're working with is in the openintro package and it's called `hfi`, short for Human Freedom Index.

1. What are the dimensions of the dataset?

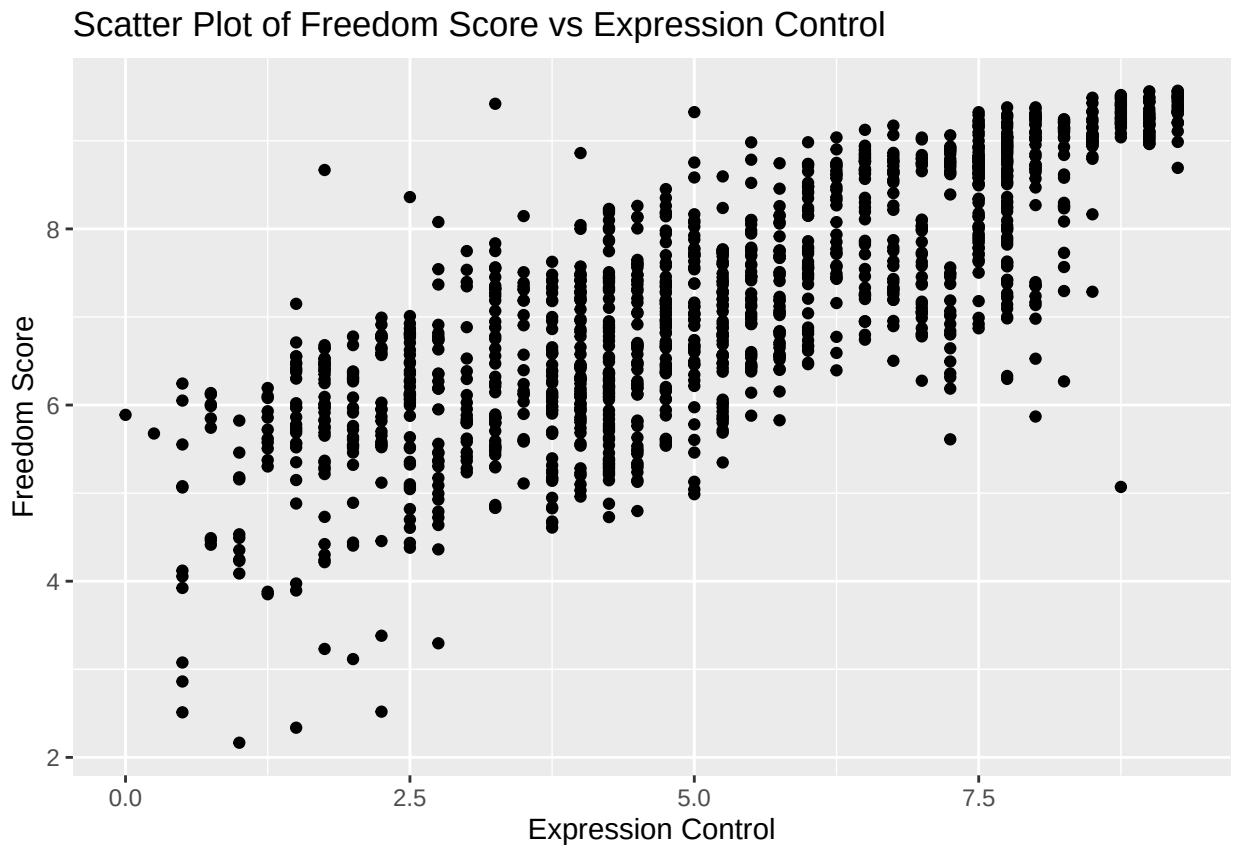
```
dim(hfi)
```

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

There are 1458 observations/rows and 123 variables/columns in the dataset

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?

```
ggplot(hfi, aes(x=pf_expression_control, y=pf_score)) +
  geom_point() +
  labs(x = "Expression Control", y = "Freedom Score") +
  ggtitle("Scatter Plot of Freedom Score vs Expression Control")
```



To visualize the connection between the personal freedom score (pf_score) and pf_expression_control, a scatter plot is ideal. This type of plot is frequently employed to illustrate how two numerical variables relate. In this case, the relationship seems to follow a linear trend. Given the linear nature observed in the scatter plot, a linear regression model seems appropriate for predicting pf_score using pf_expression_control. For accurate prediction, it's crucial to confirm the linearity of this relationship. This can be done by assessing the R^2 value, examining the residual patterns, and applying statistical tests, all of which help verify the linear model's effectiveness.

If the relationship looks linear, we can quantify the strength of the relationship with 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
```

Here, we set the `use` argument to "complete.obs" since there are some observations of NA.

Sum of squared residuals

In this section, you will use an interactive function to investigate what we mean by "sum of squared residuals". You will need to run this function in your console, not in your markdown document. Running the function also requires that the `hfi` dataset is loaded in your environment.

Think back to the way that we described the distribution of a single variable. Recall that we discussed characteristics such as center, spread, and shape. It's also useful to be able to describe the relationship of two numerical variables, such as `pf_expression_control` and `pf_score` above.

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 connection between pf_score and pf_expression_control seems to be mostly straight. This means that when pf_expression_control goes up, pf_score also tends to increase. The data doesn't make a perfect curve, but it mostly follows a straight line pattern. Also, The relationship between pf_expression_control and pf_score is positive. This means when pf_expression_control goes up, pf_score usually goes up too. And when pf_expression_control goes down, pf_score often goes down. The relationship is not super strong, but you can still see a clear trend. There's some variation in pf_score for a given pf_expression_control value. In the scatter plot, there aren't any weird or outlier points, but the variation in pf_score suggests that other things might also be affecting it.

Just as you've used the mean and standard deviation to summarize a single variable, you can summarize the relationship between these two variables by finding the line that best follows their association. Use the following interactive function to select the line that you think does the best job of going through the cloud of points.

```
# 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)
```

After running this command, you'll be prompted to click two points on the plot to define a line. Once you've done that, the line you specified will be shown in black and the residuals in blue. Note that there are 30 residuals, one for each of the 30 observations. Recall that the residuals are the difference between the observed values and the values predicted by the line:

$$e_i = y_i - \hat{y}_i$$

The most common way to do linear regression is to select the line that minimizes the sum of squared residuals. To visualize the squared residuals, you can rerun the plot command and add the argument `showSquares = TRUE`.

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

Note that the output from the `plot_ss` function provides you with the slope and intercept of your line as well as the sum of squares.

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?

After running plot_ss several times, the smallest sum of squares I got is 995.997

The linear model

It is rather cumbersome to try to get the correct least squares line, i.e. the line that minimizes the sum of squared residuals, through trial and error. Instead, you can use the `lm` function in R to fit the linear model (a.k.a. regression line).

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

The first argument in the function `lm` is a formula that takes the form `y ~ x`. Here it can be read that we want to make a linear model of `pf_score` as a function of `pf_expression_control`. The second argument specifies that R should look in the `hfi` data frame to find the two variables.

The output of `lm` is an object that contains all of the information we need about the linear model that was just fit. We can access this information using the summary function.

```
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
```

Let's consider this output piece by piece. First, the formula used to describe the model is shown at the top. After the formula you find the five-number summary of the residuals. The "Coefficients" table shown next is key; its first column displays the linear model's y-intercept and the coefficient of `pf_expression_control`. With this table, we can write down the least squares regression line for the linear model:

$$\hat{y} = 4.61707 + 0.49143 \times pf\_expression\_control$$

One last piece of information we will discuss from the summary output is the Multiple R-squared, or more simply, R^2. The R^2 value represents the proportion of variability in the response variable that is explained by the explanatory variable. For this model, 63.42% of the variability in runs is explained by at-bats.

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?

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

```r
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 of the regression line is:

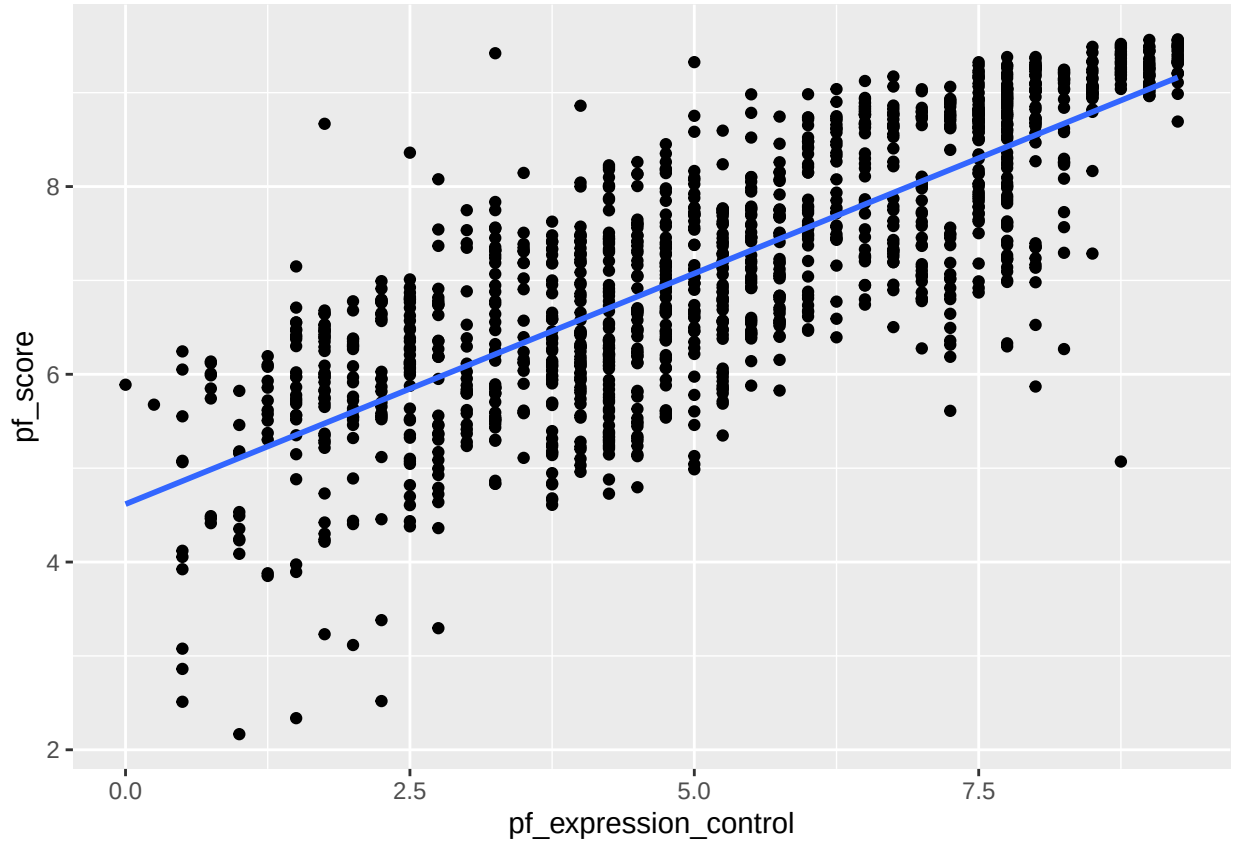
$$\hat{y} = 5.153687 + 0.349862 \times pf\_expression\_control$$

In this equation: 5.153687 is the intercept. This means if pf_expression_control is 0, the starting point for hf_score would be approximately 5.15.* 0.349862 is the slope. This tells us how much hf_score is expected to change for each one unit increase in pf_expression_control. In the context of the relationship between human freedom (hf_score) and the amount of political pressure on media content (pf_expression_control), the slope of 0.349862 indicates that for each one unit increase in political pressure on media content, the human freedom score increases by approximately 0.35.This suggests that there is a positive relationship between the two variables. In other words, as the amount of political pressure on media content increases, the human freedom score tends to increase as well, but the increase in human freedom score is less than the increase in political pressure.

Prediction and prediction errors

Let's create a scatterplot with the least squares line for `m1` laid on top.

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



Here, we are literally adding a layer on top of our plot. `geom_smooth` creates the line by fitting a linear model. It can also show us the standard error `se` associated with our line, but we'll suppress that for now.

This line can be used to predict y at any value of x. When predictions are made for values of x that are beyond the range of the observed data, it is referred to as *extrapolation* and is not usually recommended. However, predictions made within the range of the data are more reliable. They're also used to compute the residuals.

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?

```
predicting_pf_score <- 4.61707 + 0.49143*6.7
cat('pf_score predicted = ', predicting_pf_score)
```

```
## pf_score predicted =  7.909651
```

Since in hfi dataframe we do not have pf_expression_control = 6.7, but values arround it, we want to interpolate to get the actual pf_score corresponding to it.

```
actual_values_around_6.7_pf_exp <- hfi %>% filter(pf_expression_control >= 6.5, pf_expression_control <
```

Interpolation which will give the actual pf_score corresponding to pf_expression_control = 6.7:

```r
x_values <- actual_values_around_6.7_pf_exp$pf_expression_control
y_values <- actual_values_around_6.7_pf_exp$pf_score

# Interpolating to find pf_score for pf_expression_control = 6.7
interpolated_value <- approx(x = x_values, y = y_values, xout = 6.7)

# The interpolated pf_score corresponding to 6.7 pf_expression_control
interpolated_pf_score <- interpolated_value$y

actual_pf_score = interpolated_pf_score
cat('Actual value of pf_score for pf_expression_control = 6.7 is ', actual_pf_score)
```

```
## Actual value of pf_score for pf_expression_control = 6.7 is  8.038269
```

```r
cat("Residual = ", actual_pf_score - predicting_pf_score)
```

```
## Residual =  0.1286177
```

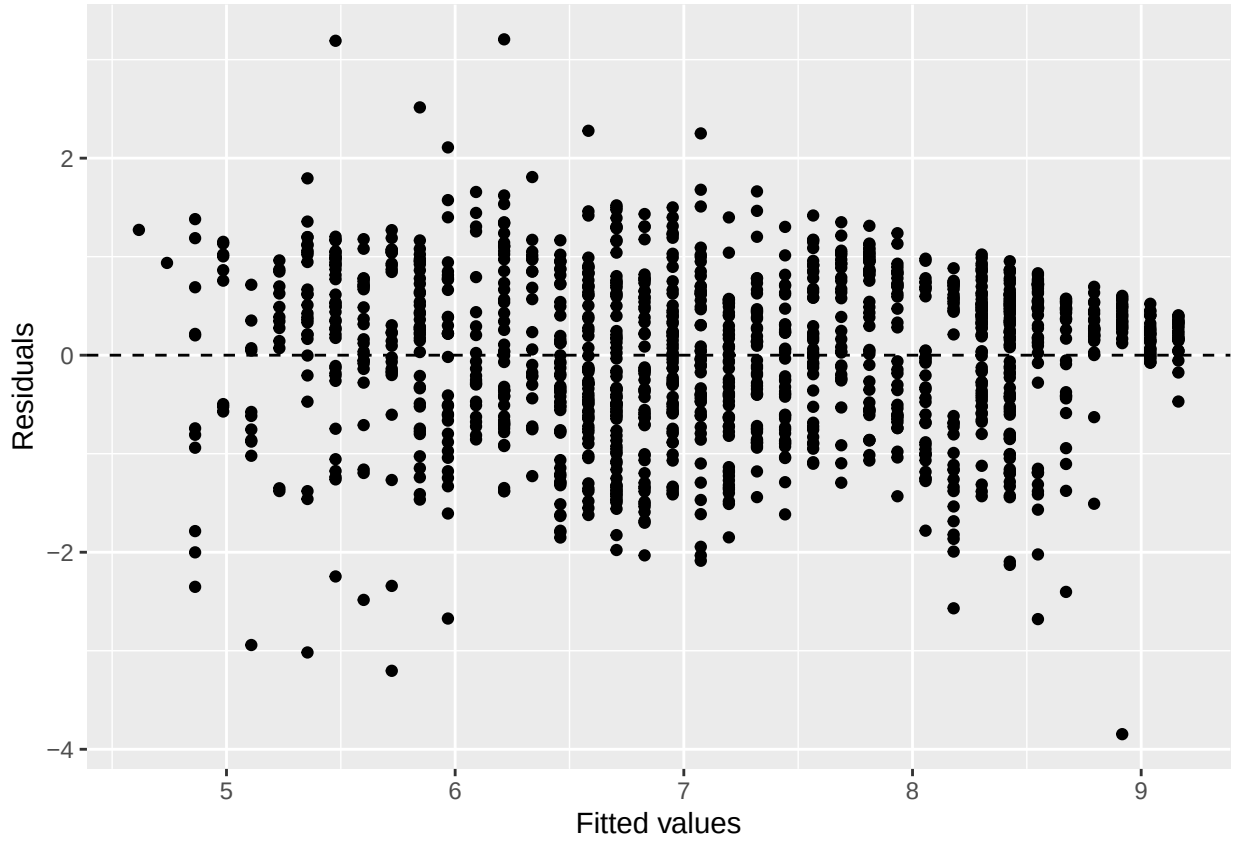
The residual for this prediction is approximately 0.1286177. Since the residual is positive, it means that the model's prediction was an underestimate. In other words, the actual pf_score was about 0.1286 units higher than what the model predicted.

Model diagnostics

To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

Linearity: You already checked if the relationship between `pf_score` and 'pf_expression_control' is linear using a scatterplot. We should also verify this condition with a plot of the residuals vs. fitted (predicted) values.

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



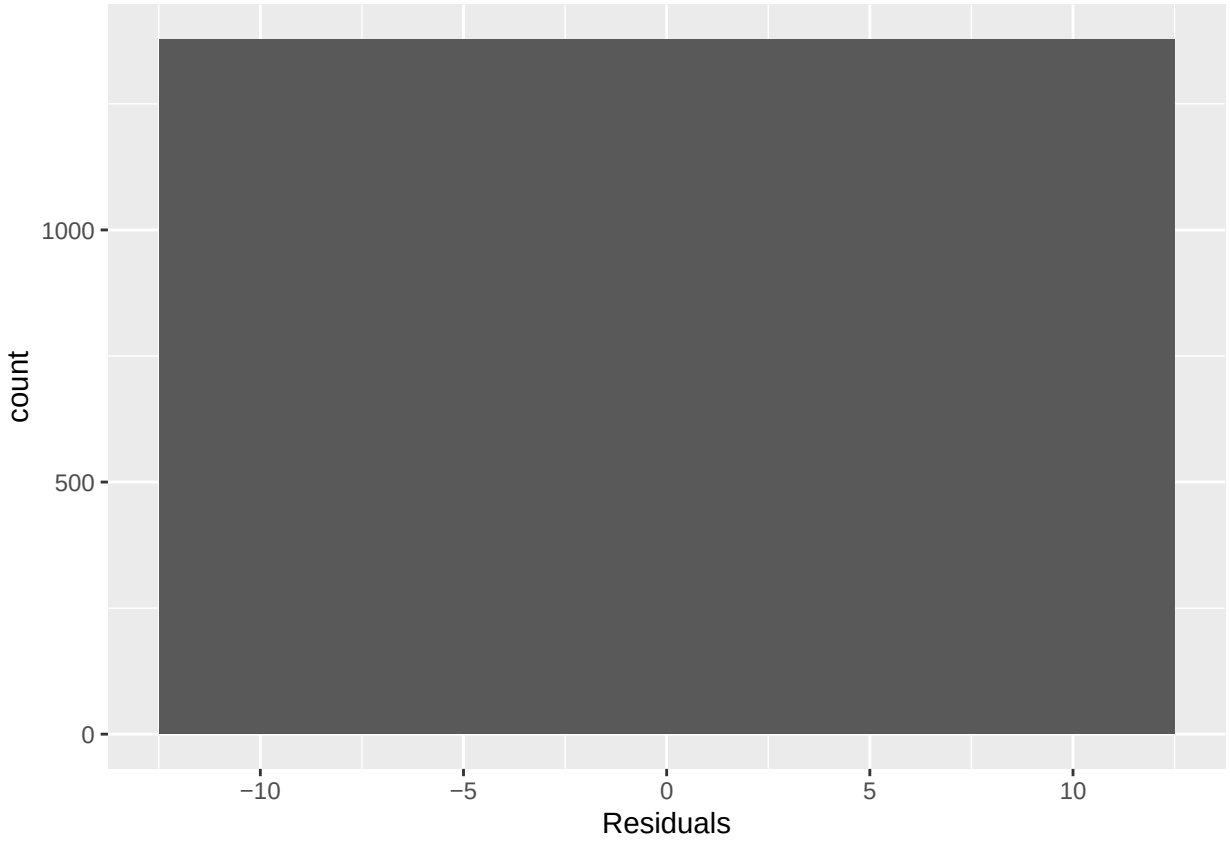
Notice here that `m1` can also serve as a data set because stored within it are the fitted values ($\hat{y}$) and the residuals. Also note that we're getting fancy with the code here. After creating the scatterplot on the first layer (first line of code), we overlay a horizontal dashed line at $y = 0$ (to help us check whether residuals are distributed around 0), and we also reanme the axis labels to be more informative.

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 points are spread out randomly around the zero line, which means the model is probably doing a good job of explaining the relationship between the variables. There's no obvious pattern like a curve or specific shape, and the points are spread out evenly, no matter the value on the x-axis. This even spread is what we expect in a good model. There are a couple of unusual points, but they're not enough to be worrisome, showing that the data mostly follows a straight-line relationship.

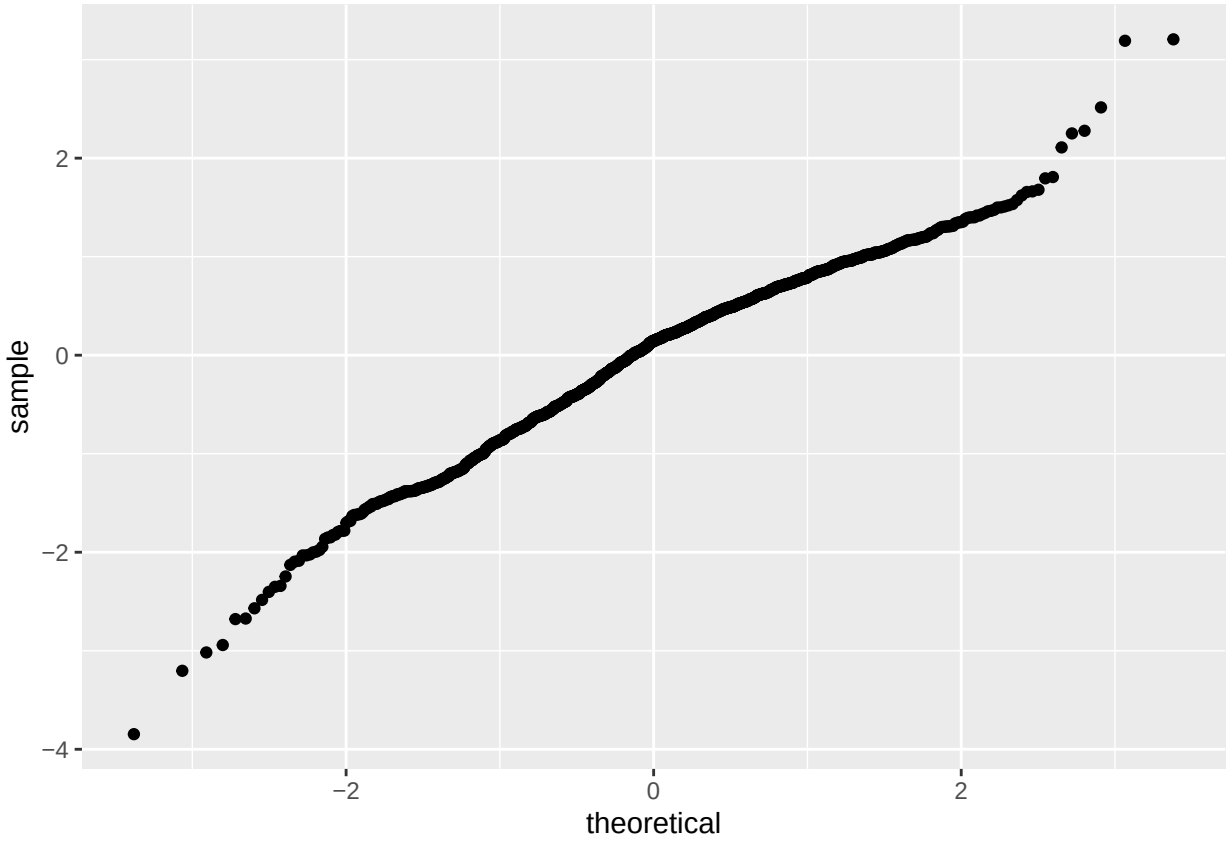
Nearly normal residuals: To check this condition, we can look at a histogram

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



or a normal probability plot of the residuals.

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



Note that the syntax for making a normal probability plot is a bit different than what you're used to seeing: we set `sample` equal to the residuals instead of `x`, and we set a statistical method `qq`, which stands for "quantile-quantile", another name commonly used for normal probability plots.

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

The histogram of residuals looks roughly like a normal distribution. It's not a perfect bell curve, but the residuals are mostly balanced around a central value of zero, without any separate peaks or clear signs of outliers. In the normal probability plot, the residuals mostly line up in a straight line, suggesting they're close to a normal distribution. There's no clear bending up or down to show skewness. A few points stray from the line at the higher end, but it's not by much.

Constant variability:

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

The residuals are spread out evenly without any noticeable pattern, which suggests that the assumption of constant variability is probably true. The width of the spread of residuals doesn't change in a regular way as we look across the fitted values, which also supports the idea of constant variability. Basically, the plot shows a consistent, random scatter of residuals around the zero line across all the fitted values, indicating that the variability is constant.

More Practice

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?

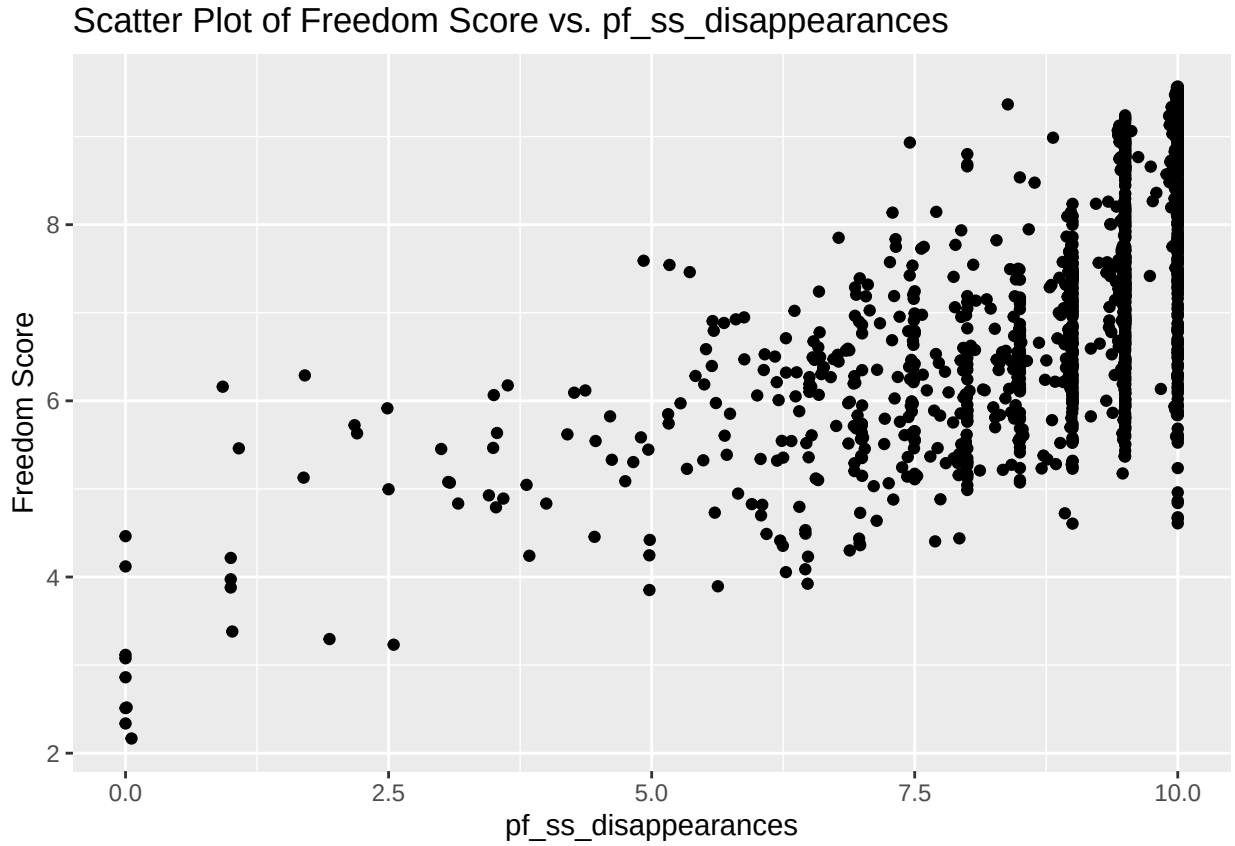
freedom variable: pf_score versus pf_ss_disappearances

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

```
## # A tibble: 1 x 1
##   `cor(pf_ss_disappearances, pf_score, use = "complete.obs")`
##                                                         <dbl>
## 1                                                       0.638
```

Plot

```
ggplot(hfi, aes(x=pf_ss_disappearances, y=pf_score)) +
  geom_point() +
  labs(x = "pf_ss_disappearances", y = "Freedom Score") +
  ggtitle("Scatter Plot of Freedom Score vs. pf_ss_disappearances")
```



The link between pf_score and pf_ss_disappearances looks like it's only slightly linear. This means when pf_ss_disappearances goes up, pf_score usually increases just a little bit. The

points don't make a clear curve. There's a general trend where higher pf_ss_disappearances is associated with a higher pf_score, and lower pf_ss_disappearances comes with a lower pf_score. But the connection isn't very strong; while you can see a trend, it's not very consistent. There's a lot of variation in pf_score even for the same pf_ss_disappearances, suggesting that there are other things affecting pf_score too. The scatter plot doesn't show any weird or extreme points, which means there aren't any major outliers throwing off the results.

11. 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?

```
m3 <- lm(pf_score ~ pf_ss_disappearances, data = hfi)
summary(m3)
```

```
##
## Call:
## lm(formula = pf_score ~ pf_ss_disappearances, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2207 -0.7855 -0.0039  0.9149  3.1307
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.54116    0.15442   16.46   <2e-16 ***
## pf_ss_disappearances   0.52886    0.01722   30.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.4066, Adjusted R-squared:  0.4062
## F-statistic: 942.9 on 1 and 1376 DF,  p-value: < 2.2e-16
```

The R^2 score, which tells us how closely related two things are, is higher for pf_score and pf_expression_control (0.6342) than for pf_score and pf_ss_disappearances (0.4066). This means pf_score and pf_expression_control have a stronger connection. In simpler terms, how well we can guess pf_score based on pf_expression_control is better compared to using pf_ss_disappearances, because the R^2 score is higher

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

Let s examine these relationship: Freedom of Movement (pf_movement) and Economic Freedom (ef_score); Women's Security (pf_ss_women) and Human Freedom(hf_score); Religious Freedom (pf_religion) and Rule of Law (pf_rol); Freedom to Establish and Operate Businesses (ef_regulation_business) and Economic Growth (ef_score); and Freedom of Expression and Personal Freedom.

Freedom of Movement (pf_movement) and Economic Freedom (ef_score):

```
model4 <- lm(pf_movement ~ ef_score, data = hfi)
summary_model4 <- summary(model4)
```

Women's Security (pf_ss_women) and Human Freedom(hf_score):

```
model5 <- lm(pf_ss_women ~ hf_score, data = hfi)
summary_model5 <- summary(model5)
```

Religious Freedom (pf_religion) and Rule of Law (pf_rol):

```
model6 <- lm(pf_religion ~ pf_rol, data = hfi)
summary_model6 <- summary(model6)
```

Freedom to Establish and Operate Businesses (ef_regulation_business) and Economic Growth (ef_score):

```
model7 <- lm(ef_regulation_business ~ ef_score, data = hfi)
summary_model7 <- summary(model7)
```

Freedom of Expression and Personal Freedom:

```
model8 <- lm(pf_score ~ pf_expression, data = hfi)
summary_model8 <- summary(model8)
```

This is the dataframe of all the five models above R^2 (residual squared):

```
Models <- data_frame(
  model_number = c('model4', 'model5', 'model6', 'model7', 'model8'),
  R_Squared = c(summary_model4$r.squared, summary_model5$r.squared, summary_model6$r.squared, summary_m
))
Models
```

```
## # A tibble: 5 x 2
##   model_number R_Squared
##   <chr>            <dbl>
## 1 model4          0.204
## 2 model5          0.424
## 3 model6          0.0356
## 4 model7          0.583
## 5 model8          0.635
```

Analyzing the residual squared table (or dataframe):

- Model 4 (Freedom of Movement and Economic Freedom): With an R^2 of 0.2044, this model suggests a low to moderate relationship. This might be surprising if one expected that the freedom of movement would have a strong correlation with economic freedom.

- Model 5 (Women's Security and Human Freedom): The R^2 here is 0.4238, which indicates a moderate relationship. This could be surprising if one assumed that women's security would be highly indicative of a country's overall human freedom.

- Model 6 (Religious Freedom and Rule of Law): With an R^2 of 0.0356, this model shows a very weak relationship, which might be surprising if the expectation was that religious freedom is closely tied to the rule of law.

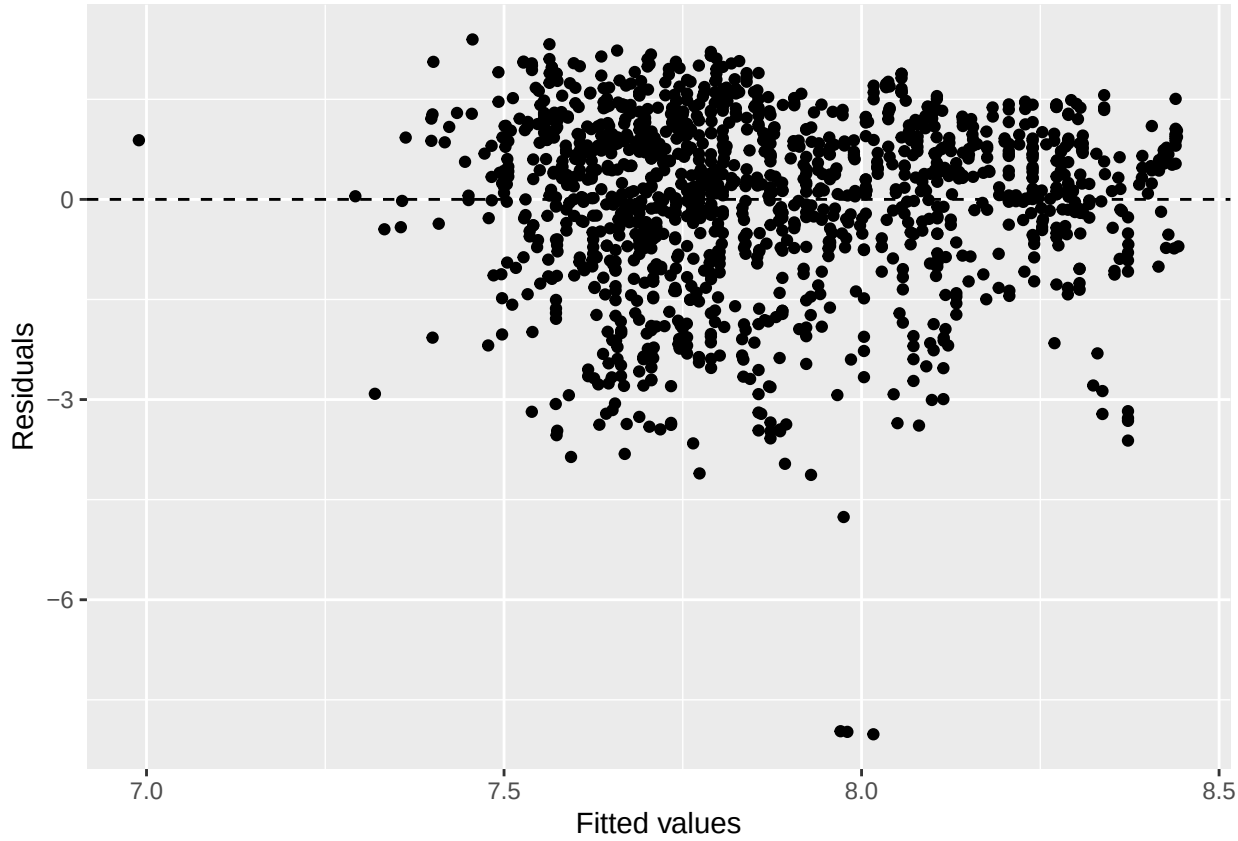
- Model 7 (Freedom to Establish and Operate Businesses and Economic Growth): This model has an R^2 of 0.5828, suggesting a moderately strong relationship. This might be less surprising since it's often assumed that business freedom would correlate with economic growth.

- Model 8 (Freedom of Expression and Personal Freedom): With the highest R^2 of 0.6351, this model indicates a strong relationship, which might be expected, as freedom of expression is a significant component of personal freedom.

Surprising relationship: **the most surprising relationship is the one between Religious Freedom and Rule of Law (Model 6), given its very low R^2 value. It might be surprising because one could hypothesize that countries with a higher rule of law would naturally allow for greater religious freedom. However, the data does not strongly support this hypothesis, indicating other factors might play a more critical role in the rule of law or that religious freedom is influenced by other, more complex societal factors not captured by this model. Now let s do the model 6 diagnostics below.**

Linearity:

```
ggplot(data = model6, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```

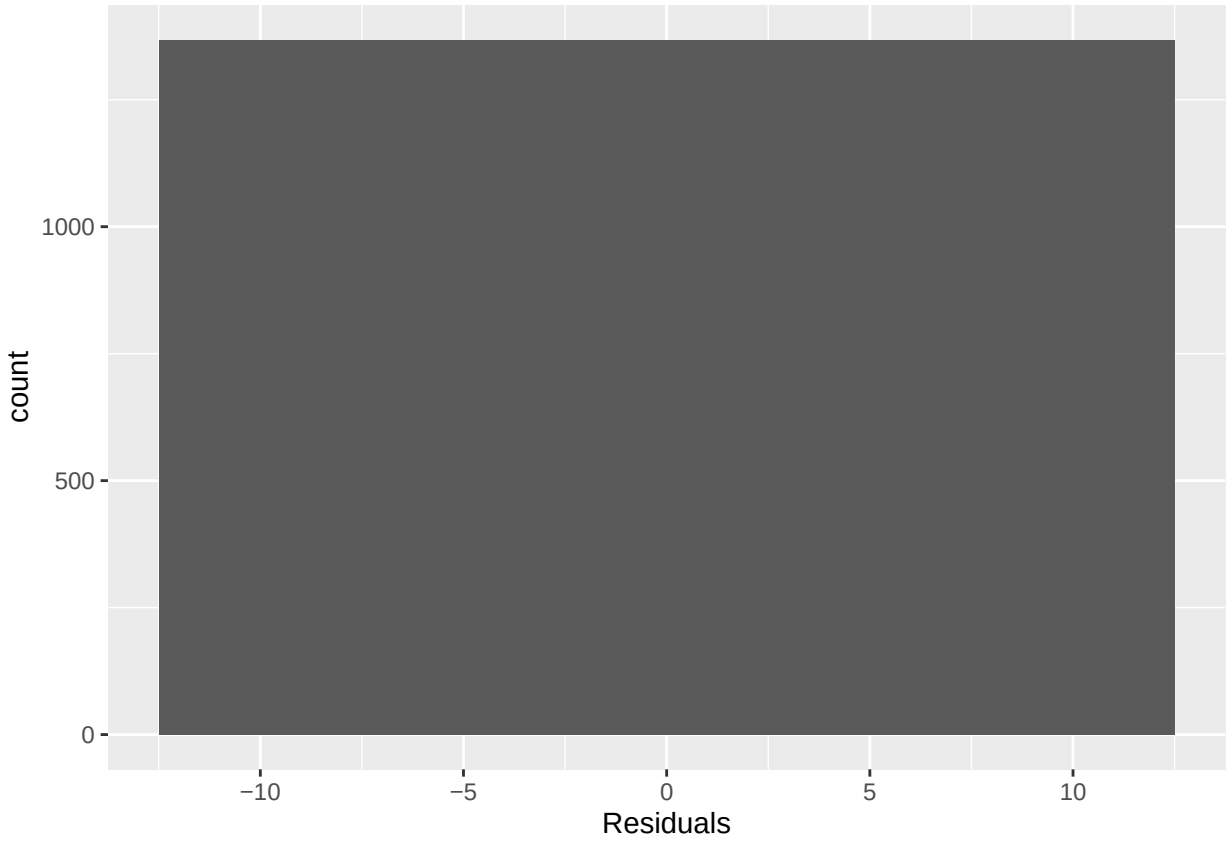


Interpretation: Looking at this plot, the residuals seem to be randomly scattered without a

clear pattern, which would initially suggest that the relationship is linear and the assumption of linearity is met. However, there appears to be a cluster of points and possibly some outliers, particularly for lower fitted values, which could indicate potential issues with outliers or influential points.

histogram of the residual:

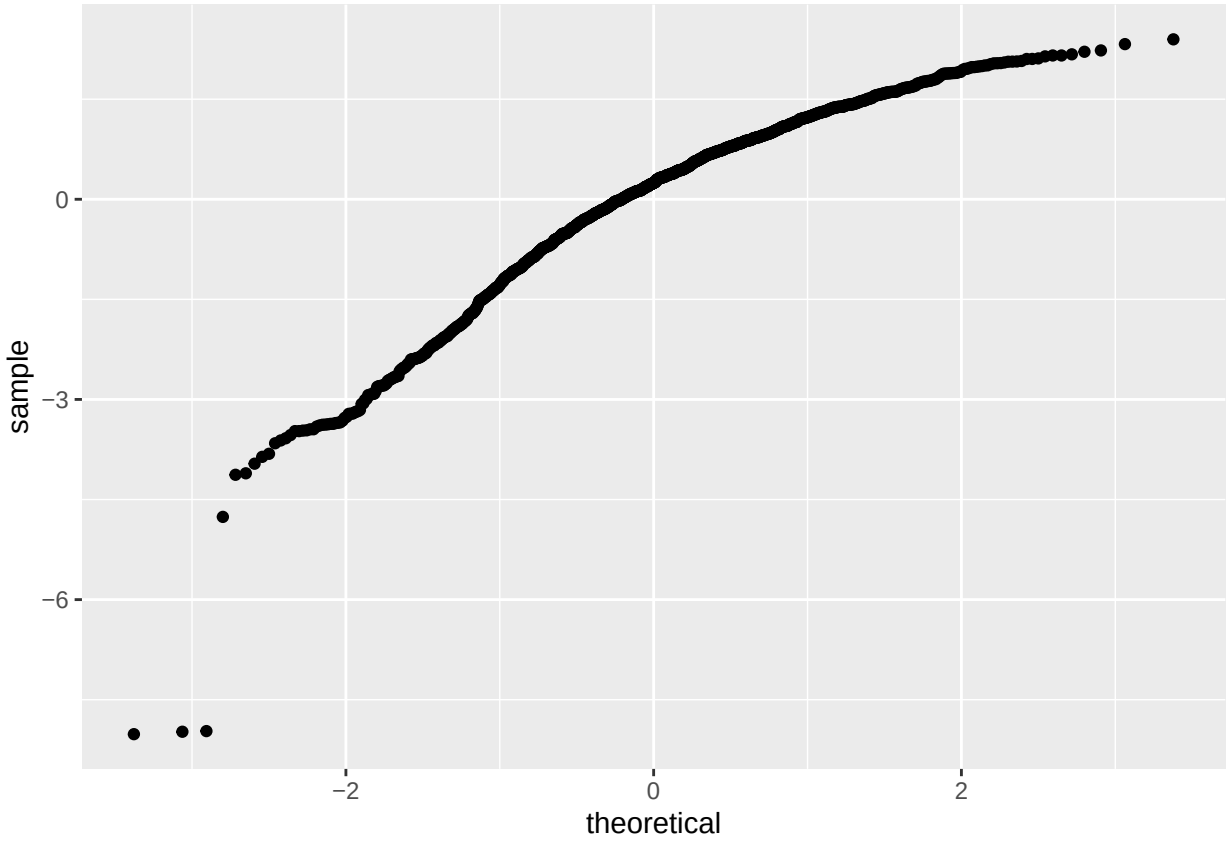
```
ggplot(data = model6, aes(x = .resid)) +
  geom_histogram(binwidth = 25) +
  xlab("Residuals")
```



Interpretation: we would expect to see a histogram that resembles a bell curve (normal distribution). However, the histogram here appears to be a single, large bar, which suggests that either the residuals are not normally distributed or there could be an issue with how the histogram was generated

and normal probability plot of the residuals:

```
ggplot(data = model6, aes(sample = .resid)) +
  stat_qq()
```



Interpretation: In The Q-Q plot we would expect a perfect match for the normal distribution which would show points forming a straight diagonal line. But, In this Q-Q plot, most of the points follow the line but deviate in the tails, especially in the lower tail, which suggests that the residuals may have a distribution with heavier tails than a normal distribution. This could indicate potential outliers or that the data is skewed.

Conclusion:

After reviewing the model diagnostics for the relationship between Religious Freedom and Rule of Law, which yielded an R^2 of 0.0356, the conclusion is as follows: The weak R^2 value initially indicated a surprising finding, as one might expect a stronger connection between Religious Freedom and Rule of Law. The residuals plot did not show any clear patterns that would suggest non-linearity, indicating that the model's assumptions of linearity and equal variance are reasonable. However, the normality diagnostics revealed some concerns. The histogram suggested potential issues with the data distribution or the plotting scale, and the Q-Q plot showed deviations from normality, particularly in the tails, hinting at possible outliers or heavy-tailed distribution of residuals. Together, these diagnostics suggest that while the linear model may capture the essence of the relationship, there are underlying complexities—possibly extreme values or other variables at play—not captured by the model, leading to a lower than expected R^2 value. This surprising result underscores the multifaceted nature of how religious freedoms interact with legal systems and possibly other societal factors not included in the model6.