

Lab 8: Introduction to linear regression

João De Oliveira

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?

```
nrow(hfi)
```

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

```
ncol(hfi)
```

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

```
dim(hfi)
```

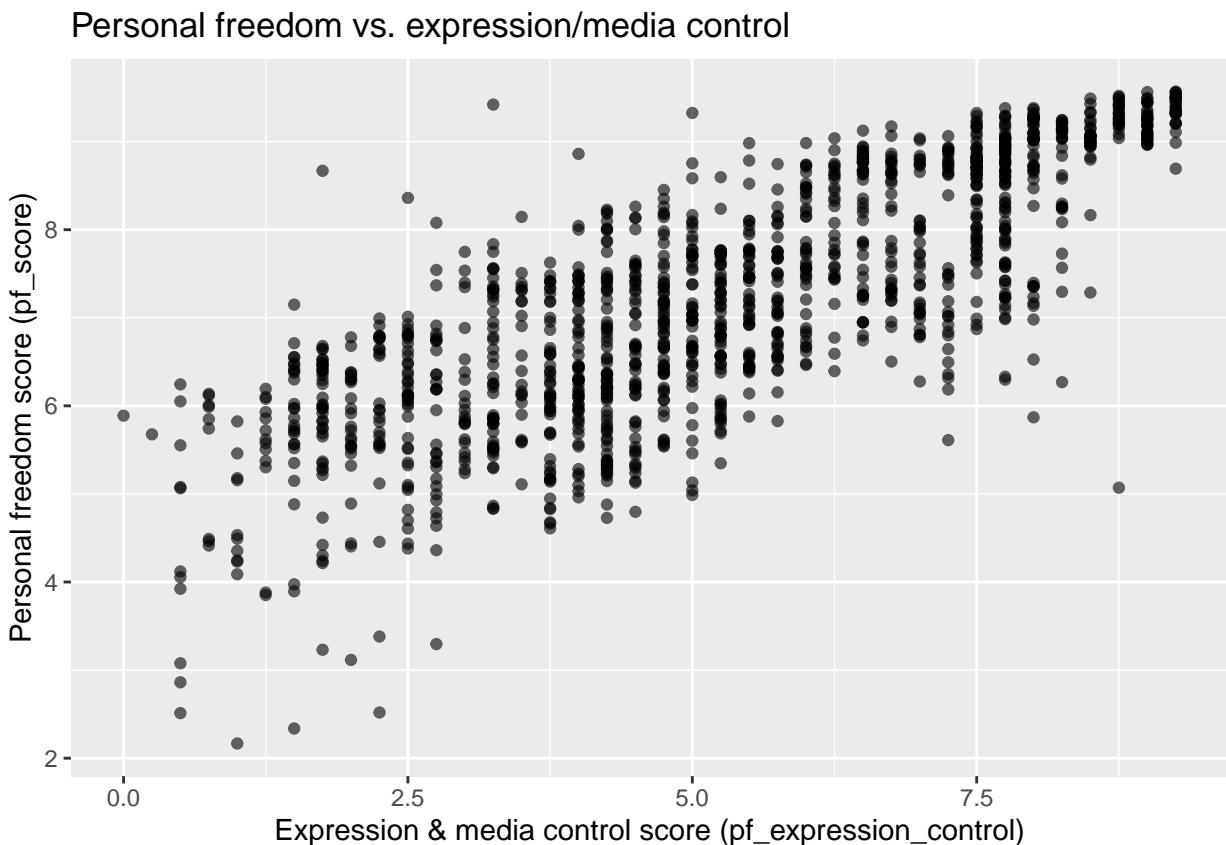
```
## [1] 1458 123
```

Answer 1 The dataset has 1458 rows and 123 columns.

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?

Answer 2 A scatterplot would be the appropriate plot to display the relationship between two numerical variables.

```
# Scatterplot of pf_score vs pf_expression_control
ggplot(hfi, aes(x = pf_expression_control, y = pf_score)) +
  geom_point(alpha = 0.6) +
  labs(
    x = "Expression & media control score (pf_expression_control)",
    y = "Personal freedom score (pf_score)",
    title = "Personal freedom vs. expression/media control"
  )
```



The relationship appears to be linear and positive. As `pf_expression_control` increases (meaning less political pressure on media), `pf_score` tends to increase as well. There is a clear upward trend with points generally following a linear pattern, though there is some scatter. Yes, I would be comfortable using a linear model to predict the personal freedom score based on `pf_expression_control` since the relationship appears reasonably linear.

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.

Answer 3

- **Form:** The relationship appears to be linear. Despite not following a perfect straight line, the distribution of the points is roughly linear.
- **Direction:** The relationship is positive. As `pf_expression_control` increases (less political pressure on media), `pf_score` also increases.
- **Strength:** The relationship is strong, with a correlation coefficient of 0.796. The points are relatively close to what would be a best-fit line, though there is some scatter, as mentioned above.
- **Unusual observations:** There are a few countries that appear to be outliers- some with relatively low `pf_expression_control` but moderate `pf_score`, and possibly some with high scores in both variables that extend beyond the main cluster. However, no extreme outliers that would drastically affect the linear relationship are immediately apparent.

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?

```
hfi_clean <- hfi %>% filter(complete.cases(pf_expression_control, pf_score))

m1_temp <- lm(pf_score ~ pf_expression_control, data = hfi_clean)

sum(residuals(m1_temp)^2)

## [1] 952.1532
```

Answer 4 I ran `plot_ss` several times. The smallest sum of squared residuals I achieved was about 952.1532. Compared to my classmates, my best sum of squares was similar to most of them and actually the same when compared with a few students.

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
## -10.0000 -1.0000  0.0000  1.0000 10.0000
```

```

## -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:

$$\hat{y} = 4.61707 + 0.49143 \times \text{pf_expression_control}$$

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 `pf_free` is explained by `pf_expression_control`.

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?

```
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

```

Answer 5 The regression equation is:

$$\hat{hf_score} = 5.15369 + 0.34986 \times pf_expression_control$$

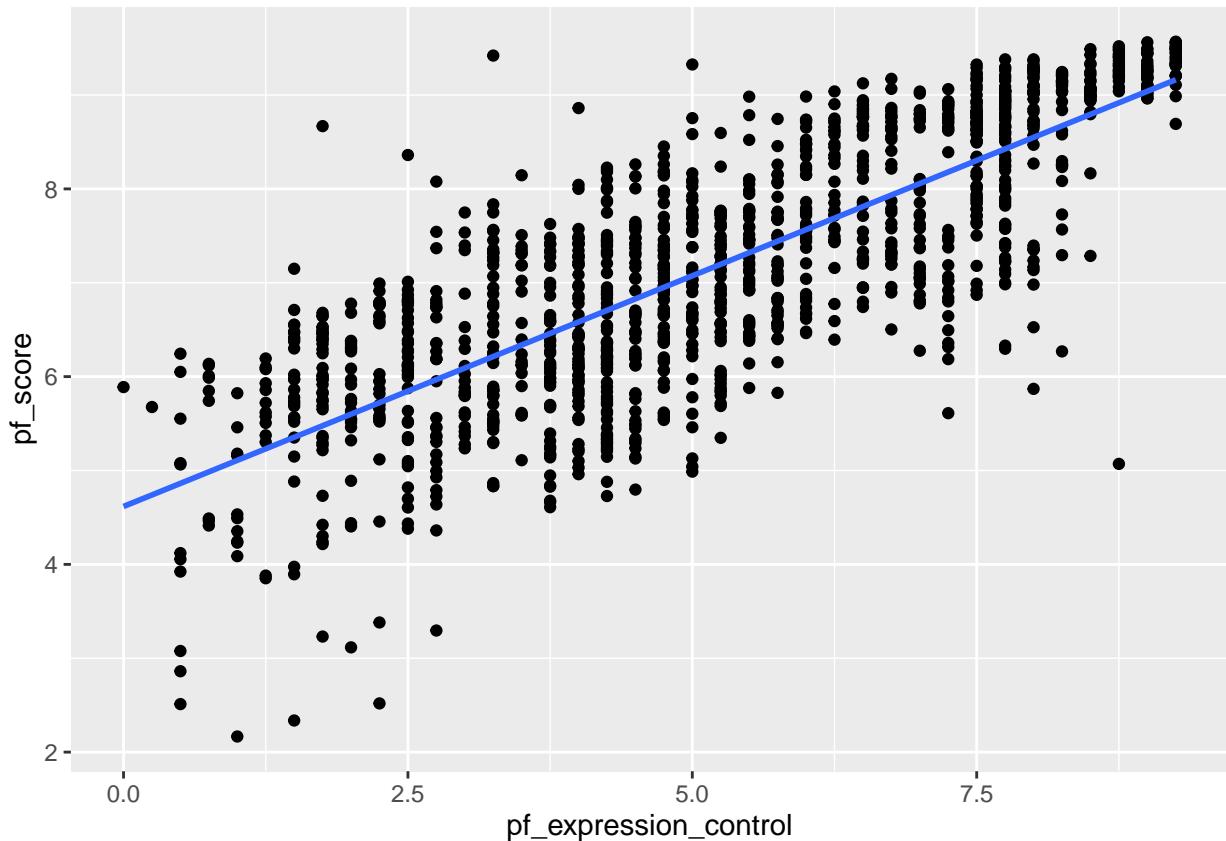
Interpretation of the slope: The slope of 0.34986 tells us that for each one-point increase in `pf_expression_control` (i.e., for each unit decrease in political pressure on media content, since 0 = most pressure), we expect the human freedom score to increase by approximately 0.35 points, on average. This indicates a positive relationship: countries with less political pressure on media content tend to have higher overall human freedom scores.

The R^2 value is 57.75%, meaning that `pf_expression_control` explains approximately 57.75% of the variability in `hf_score`.

Prediction and prediction errors

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

```
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?

```
# countries with pf_expression_control close to 6.7
hfi_67 <- hfi %>%
  filter(!is.na(pf_expression_control), !is.na(pf_score)) %>%
  mutate(diff = abs(pf_expression_control - 6.7)) %>%
  arrange(diff) %>%
  head(8)

hfi_67

## # A tibble: 8 x 124
##   year ISO_code countries region pf_rol_procedural pf_rol_civil pf_rol_crime
##   <dbl> <chr>    <chr>     <chr>          <dbl>        <dbl>        <dbl>
## 1 2016 BLZ      Belize     Latin~       4.75        4.74        3.32
## 2 2016 CHL      Chile      Latin~       7.71        6.29        5.55
## 3 2016 FRA      France     Weste~       6.82        7.02        6.47
## 4 2016 GHA      Ghana      Sub-S~      5.85        6.24        5.15
## 5 2016 NAM      Namibia    Sub-S~      NA          NA          NA
## 6 2016 PNG      Pap. New~  Ocean~       NA          NA          NA
## 7 2016 SUR      Suriname   Latin~       4.61        5.02        5.17
## 8 2015 CHL      Chile      Latin~       7.85        6.42        5.76
## # i 117 more variables: pf_rol <dbl>, pf_ss_homicide <dbl>,
## #   pf_ss_disappearances_disap <dbl>, pf_ss_disappearances_violent <dbl>,
## #   pf_ss_disappearances_organized <dbl>,
## #   pf_ss_disappearances_fatalities <dbl>, pf_ss_disappearances_injuries <dbl>,
## #   pf_ss_disappearances <dbl>, pf_ss_women_fgm <dbl>,
## #   pf_ss_women_missing <dbl>, pf_ss_women_inheritance_widows <dbl>,
## #   pf_ss_women_inheritance_daughters <dbl>, pf_ss_women_inheritance <dbl>, ...

# predicted value for pf_expression_control = 6.7
predicted_score <- coef(m1)[1] + coef(m1)[2] * 6.7
predicted_score

## (Intercept)
## 7.909663

countries_67 <- hfi %>%
  filter(!is.na(pf_expression_control), !is.na(pf_score),
         abs(pf_expression_control - 6.7) < 0.5) %>%
  select(year, countries, pf_expression_control, pf_score) %>%
  mutate(predicted = coef(m1)[1] + coef(m1)[2] * pf_expression_control,
         residual = pf_score - predicted)

countries_67
```

```

## # A tibble: 144 x 6
##   year countries    pf_expression_control pf_score predicted residual
##   <dbl> <chr>                <dbl>      <dbl>      <dbl>      <dbl>
## 1 2016 Belize             6.75      7.43      7.93     -0.503
## 2 2016 Burkina Faso       7          7.46      8.06     -0.602
## 3 2016 Chile              6.75      8.22      7.93      0.282
## 4 2016 France             6.75      8.77      7.93      0.833
## 5 2016 Ghana              6.75      7.87      7.93     -0.0621
## 6 2016 Israel             6.5        7.54      7.81     -0.266
## 7 2016 Korea, South       6.5        8.77      7.81      0.955
## 8 2016 Malawi             6.25      7.45      7.69     -0.243
## 9 2016 Mongolia           7          8.00      8.06     -0.0588
## 10 2016 Namibia            6.75      7.39      7.93     -0.539
## # i 134 more rows

```

Answer 6 Using the regression equation $\hat{y} = 4.61707 + 0.49143 \times \text{pf_expression_control}$, for a country with `pf_expression_control = 6.7`, the predicted `pf_score` would be:

$$\hat{y} = 4.61707 + 0.49143 \times 6.7 = 7.91$$

To determine if this is an overestimate or underestimate, we would need to compare it to the actual `pf_score` for a country with `pf_expression_control = 6.7`. From the data above, we can see countries with values close to 6.7. The residual for each country shows whether the model overestimates (negative residual) or underestimates (positive residual) the actual score.

Model diagnostics

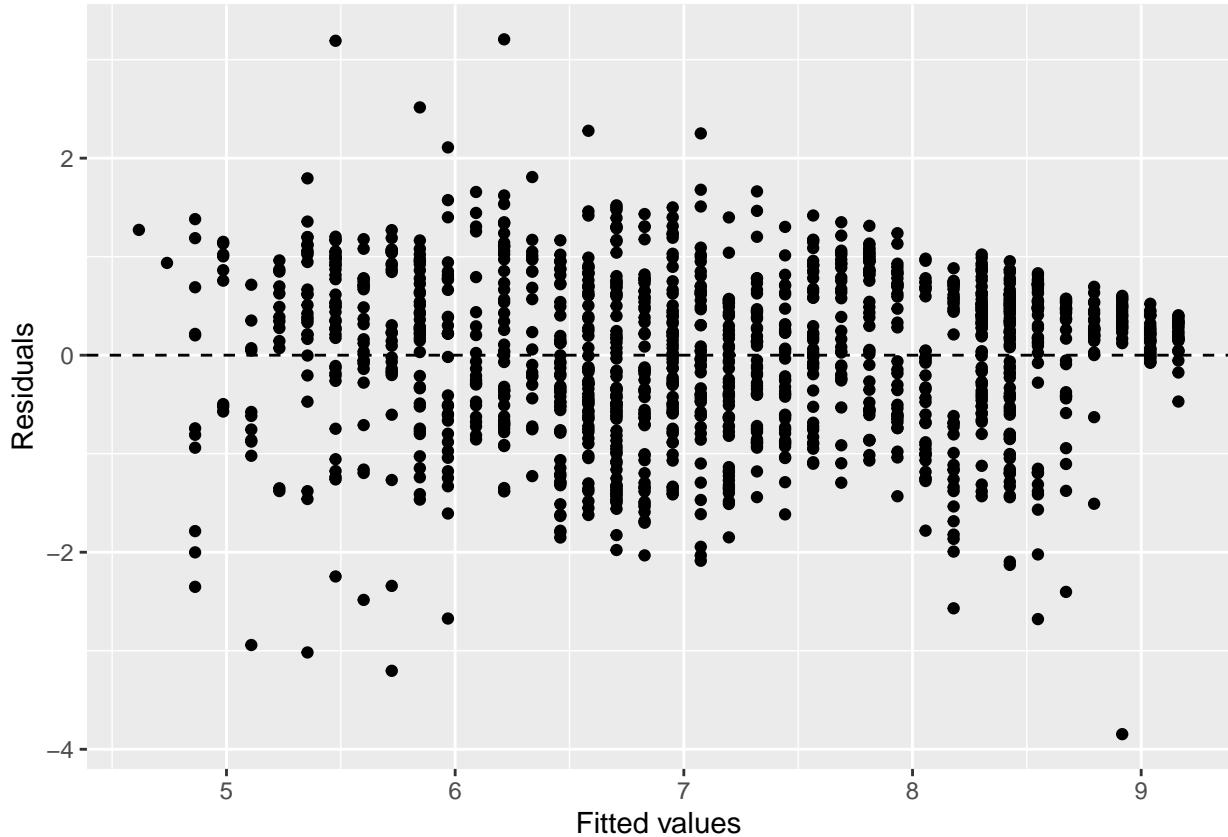
To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

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.

```

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

```



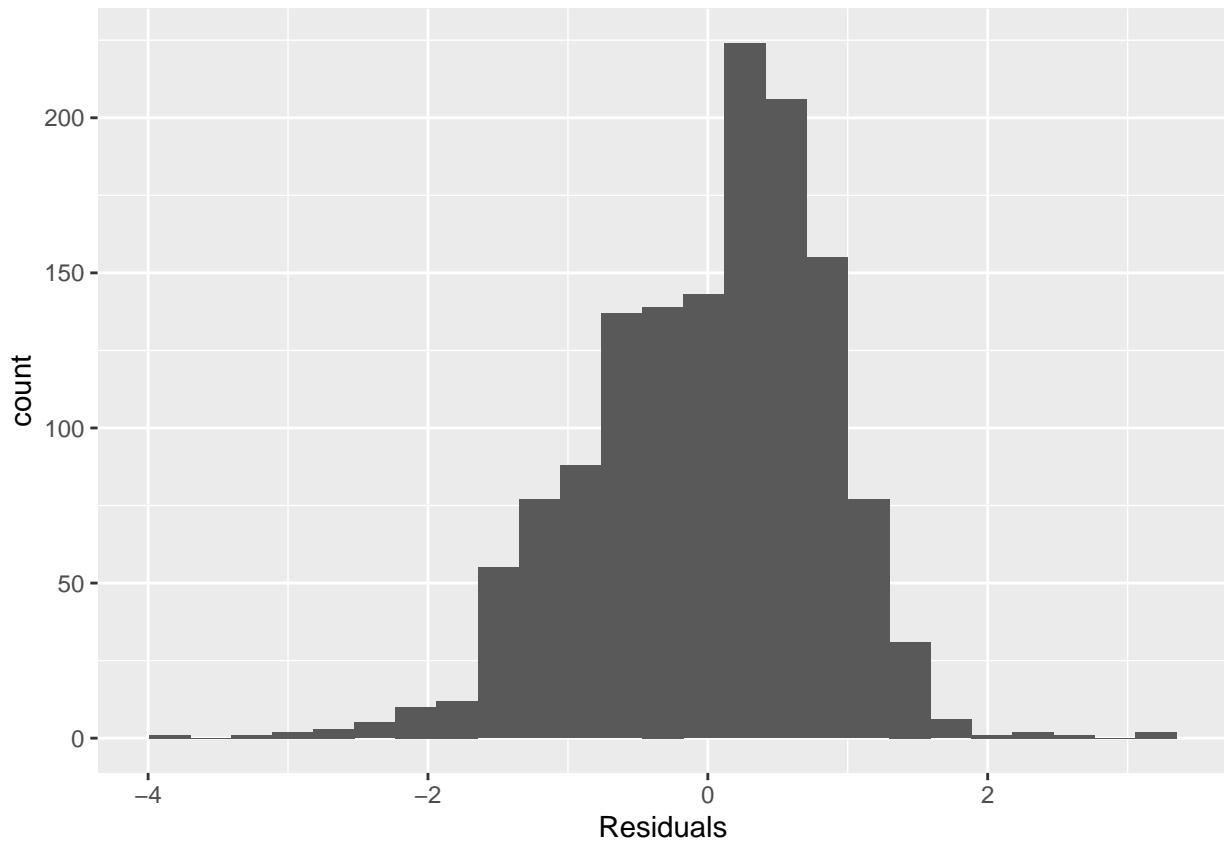
Notice here that `m1` can also serve as a data set because stored within it are the fitted values (\hat{y}) and the residuals. Also note that we're getting fancy with the code here. After creating the scatterplot on the first layer (first line of code), we overlay a horizontal dashed line at $y = 0$ (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?

Answer 7 Looking at the residuals vs fitted values plot, the residuals seem to be randomly scattered around the horizontal line at $y = 0$, with no clear patterns such as curves, funnels, or systematic trends. This means that the linearity condition is satisfied. If there were a pattern, it would suggest that a linear model may not be appropriate and that the relationship might be non-linear. The random scatter suggests that a linear model is appropriate for this relationship.

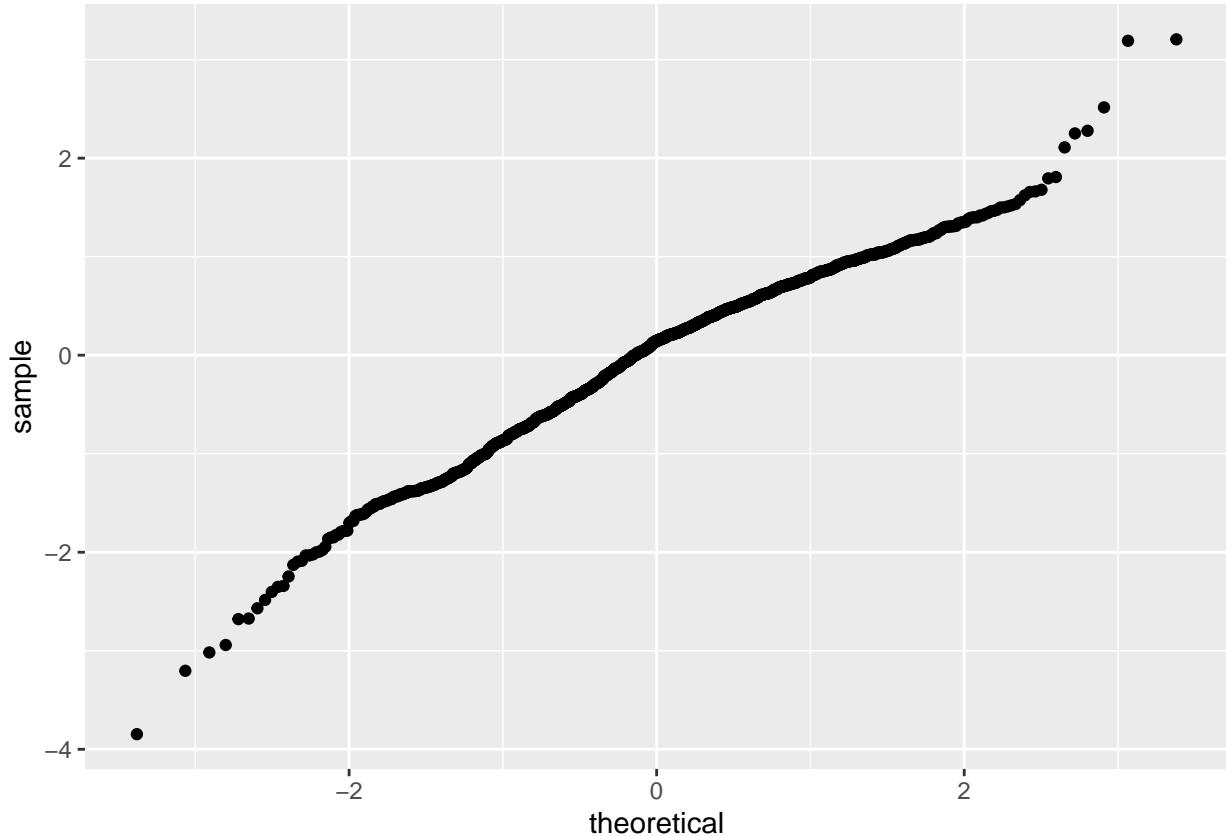
Nearly normal residuals: To check this condition, we can look at a histogram

```
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(bins = 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?

Answer 8 - Histogram: The histogram of residuals appears to be roughly symmetric and bell-shaped, centered around 0, which suggests the residuals are approximately normally distributed.

- **Q-Q Plot:** The points in the Q-Q plot generally follow the diagonal line, with some minor deviations at the tails. This indicates that the residuals are approximately normally distributed. The points staying close to the line suggests that the nearly normal residuals condition is reasonably well met.

Overall, I think the nearly normal residuals condition appears to be met for this model.

Constant variability:

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

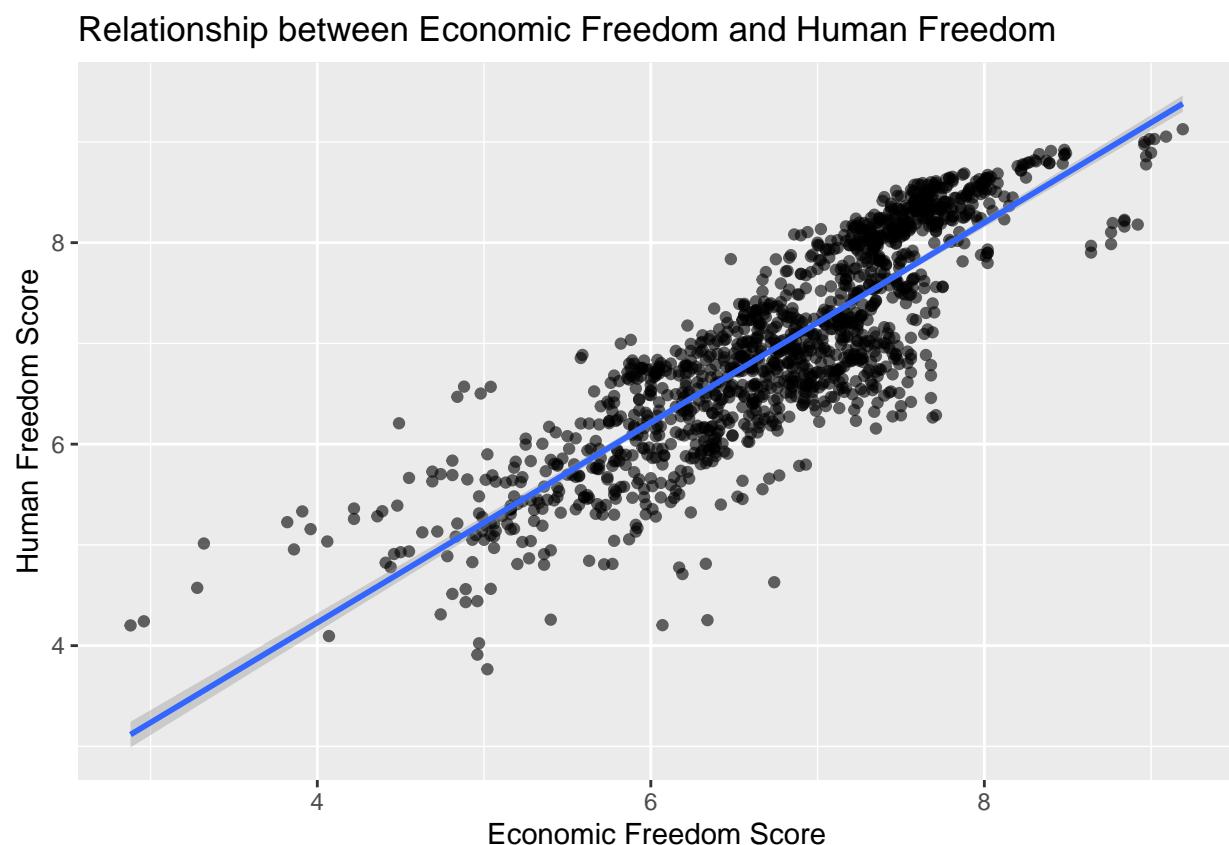
Answer 9 Looking back at the residuals vs fitted values plot, the constant variability condition appears to be met. The spread of the residuals around the horizontal line at $y = 0$ appears to be relatively constant across all fitted values. There is no obvious funnel shape. The residuals maintain roughly the same vertical spread throughout the range of fitted values, which satisfies the constant variability assumption for linear regression. * * *

More Practice

- 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?

Answer: Exploring the relationship between economic freedom score and human freedom score, we would expect these to be positively correlated.

```
# scatterplot
ggplot(data = hfi, aes(x = ef_score, y = hf_score)) +
  geom_point(alpha = 0.6) +
  stat_smooth(method = "lm", se = TRUE) +
  labs(x = "Economic Freedom Score",
       y = "Human Freedom Score",
       title = "Relationship between Economic Freedom and Human Freedom")
```



```
# linear model
m3 <- lm(hf_score ~ ef_score, data = hfi)
summary(m3)
```

```
##
## Call:
## lm(formula = hf_score ~ ef_score, data = hfi)
##
```

```

## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.31864 -0.36668  0.05449  0.41767  1.49198
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.25906   0.11112   2.331   0.0199 *
## ef_score    0.99245   0.01624  61.117 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5324 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.7308, Adjusted R-squared:  0.7306
## F-statistic: 3735 on 1 and 1376 DF,  p-value: < 2.2e-16

# correlation
cor(hfi$ef_score, hfi$hf_score, use = "complete.obs")

```

```
## [1] 0.8548651
```

Yes, there appears to be a strong positive linear relationship between economic freedom and human freedom. The correlation coefficient is approximately 0.855, and the points follow a clear linear pattern. The R^2 value is 73.08%, meaning economic freedom explains about 73.08% of the variability in human freedom scores.

- 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?

```

# R-squared for m1: pf_expression_control predicting pf_score
r2_m1 <- summary(m1)$r.squared

# R-squared for m2: pf_expression_control predicting hf_score
r2_m2 <- summary(m2)$r.squared

# R-squared for m3: ef_score predicting hf_score
r2_m3 <- summary(m3)$r.squared
cat("R-squared for pf_expression_control -> pf_score:", round(r2_m1 * 100, 2), "%\n")

```

```
## R-squared for pf_expression_control -> pf_score: 63.42 %

cat("R-squared for pf_expression_control -> hf_score:", round(r2_m2 * 100, 2), "%\n")
```

```
## R-squared for pf_expression_control -> hf_score: 57.75 %

cat("R-squared for ef_score -> hf_score:", round(r2_m3 * 100, 2), "%\n")
```

```
## R-squared for ef_score -> hf_score: 73.08 %
```

Comparison:

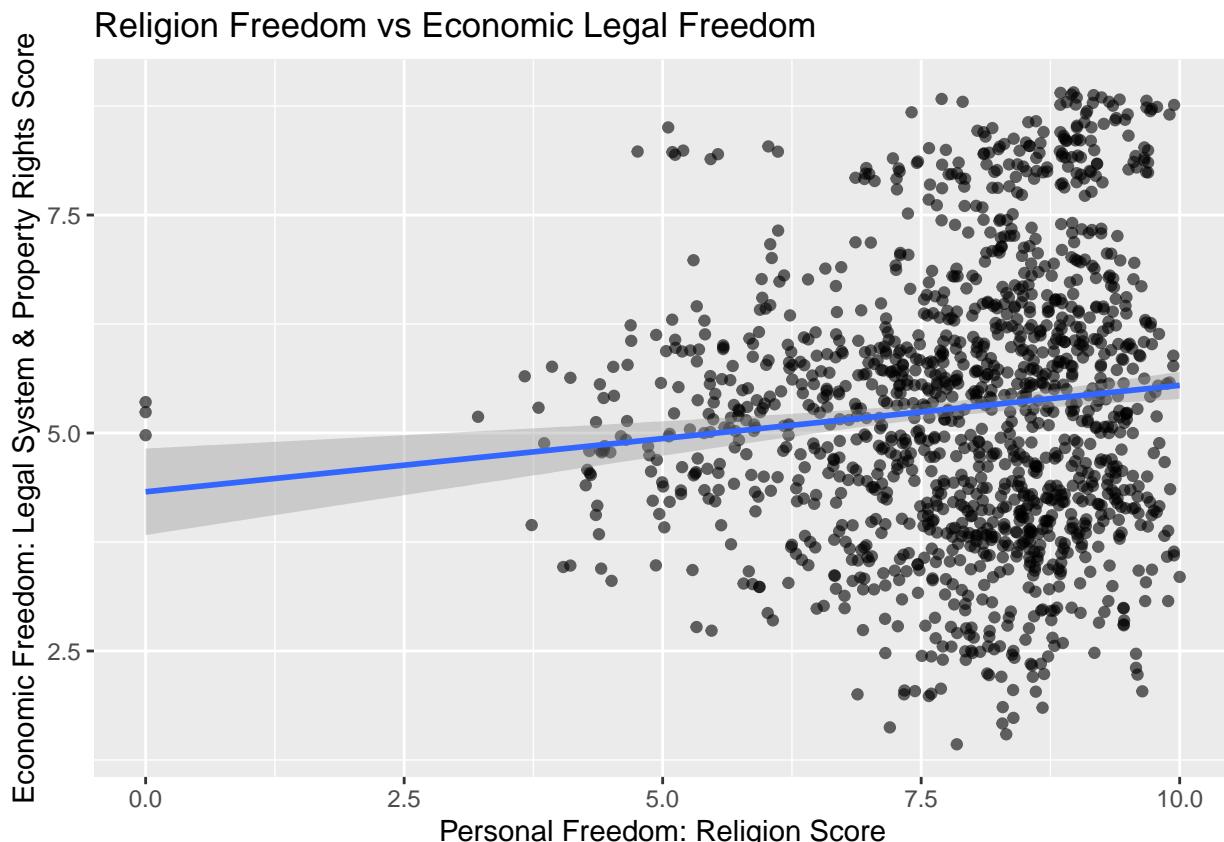
The relationship between `ef_score` and `hf_score` (Model 3) has a higher R^2 value (73.08%) compared to `pf_expression_control` predicting `hf_score` (Model 2, 57.75%). This makes sense because: 1. **Economic freedom is a component of human freedom:** The human freedom index likely includes economic freedom as one of its components, so there's a more direct relationship. 2. **Broader measure:** Economic freedom is a broader concept that encompasses many aspects of economic life, making it a stronger predictor of overall human freedom. 3. **Component vs. aggregate:** `pf_expression_control` is a more specific measure (focused on media/political expression), while `ef_score` is a composite measure that aligns more closely with the overall `hf_score` structure.

However, Model 1 ($\text{pf_expression_control} \rightarrow \text{pf_score}$) has a strong R^2 of 63.42% because `pf_expression_control` is directly related to personal freedom.

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

Answer: Exploring the relationship between personal freedom related to religion and economic freedom related to legal system and property rights:

```
# scatterplot
ggplot(data = hfi, aes(x = pf_religion, y = ef_legal)) +
  geom_point(alpha = 0.6) +
  stat_smooth(method = "lm", se = TRUE) +
  labs(x = "Personal Freedom: Religion Score",
       y = "Economic Freedom: Legal System & Property Rights Score",
       title = "Religion Freedom vs Economic Legal Freedom")
```



```

# linear model

m4 <- lm(ef_legal ~ pf_religion, data = hfi)
summary(m4)

## 
## Call:
## lm(formula = ef_legal ~ pf_religion, data = hfi)
##
## Residuals:
##       Min     1Q Median     3Q    Max
## -3.8538 -1.2050  0.0014  0.9572  3.5657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.32569   0.25296 17.100 < 2e-16 ***
## pf_religion 0.12209   0.03166  3.856 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.579 on 1366 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.01077,   Adjusted R-squared:  0.01004
## F-statistic: 14.87 on 1 and 1366 DF,  p-value: 0.0001206

cor(hfi$pf_religion, hfi$ef_legal, use = "complete.obs")

## [1] 0.1037701

```

There is a moderate positive correlation (0.104), which might be somewhat surprising. It suggests that countries with greater religious freedom also tend to have stronger legal systems and property rights protections. This could indicate that societies that value personal freedoms (like religious freedom) also tend to establish institutions that protect economic freedoms.

Now let's display the model diagnostics:

```

# residuals vs Fitted
p1 <- ggplot(data = m4, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals") +
  ggtitle("Residuals vs Fitted Values")

# histogram of residuals
p2 <- ggplot(data = m4, aes(x = .resid)) +
  geom_histogram(bins = 25, fill = "steelblue", alpha = 0.7, color = "white") +
  xlab("Residuals") +
  ggtitle("Distribution of Residuals")

# Q-Q plot
p3 <- ggplot(data = m4, aes(sample = .resid)) +

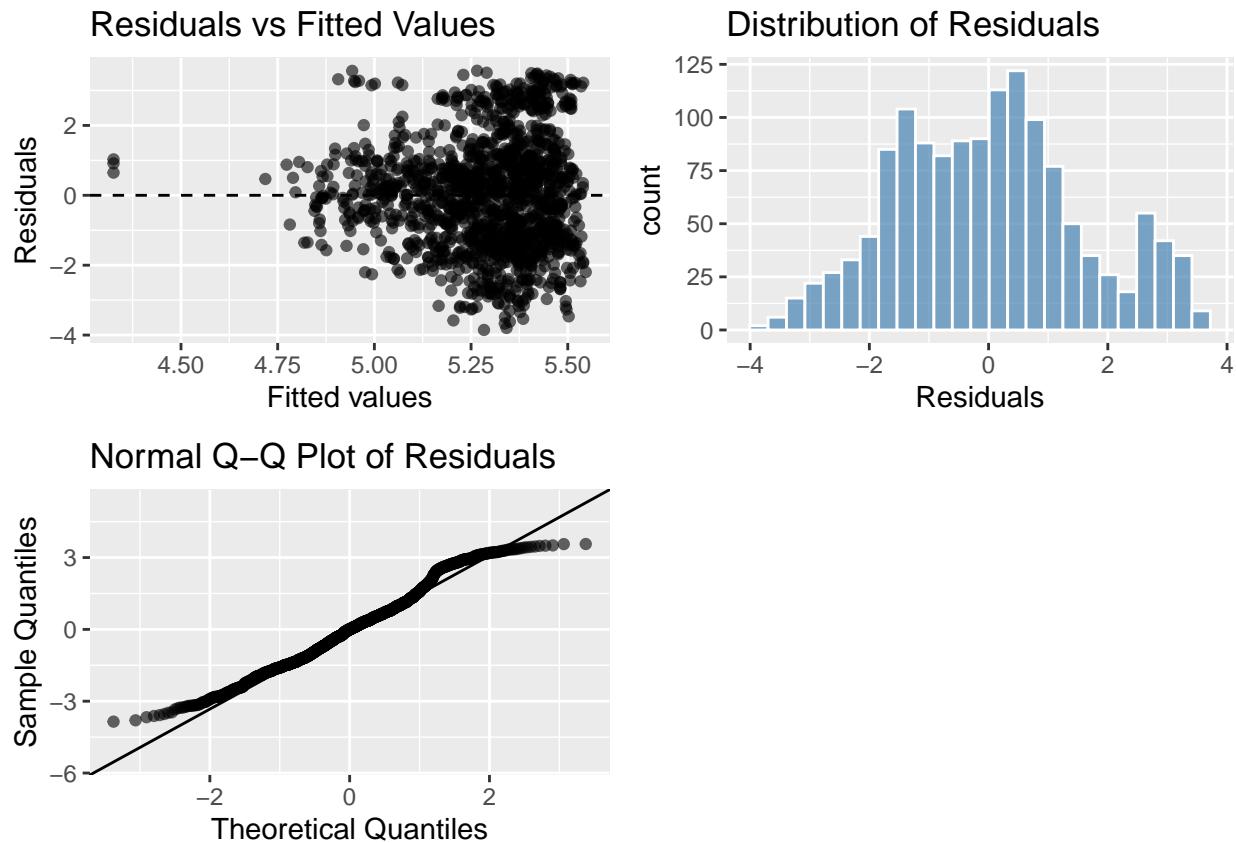
```

```

stat_qq(alpha = 0.6) +
stat_qq_line() +
xlab("Theoretical Quantiles") +
ylab("Sample Quantiles") +
ggtitle("Normal Q-Q Plot of Residuals")

# Display plots
library(gridExtra)
grid.arrange(p1, p2, p3, ncol = 2)

```



Model Diagnostics Assessment:

- Linearity (Residuals vs Fitted):** The residuals appear randomly scattered around 0 with no clear patterns, suggesting the linearity condition is met.
- Nearly Normal Residuals:** The histogram shows a roughly bell-shaped distribution centered around 0, and the Q-Q plot shows points generally following the diagonal line, indicating the residuals are approximately normally distributed.
- Constant Variability:** The spread of residuals appears relatively constant across fitted values, suggesting constant variability.

Why this relationship is interesting/surprising: The positive relationship between religious freedom and economic legal freedom suggests that societies that protect personal freedoms also tend to have stronger institutions protecting economic rights. This could reflect broader cultural and institutional factors that support both personal and economic freedoms. The R^2 of 1.08% indicates that religious freedom explains

a meaningful portion of the variability in economic legal freedom, though there are clearly other factors at play as well.
