

Lab 8 Pricilla Nakyzze Introduction to linear regression

2025-03-24

```
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'  
## (as 'lib' is unspecified)
```

```
install.packages('openintro')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'  
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --  
## v dplyr      1.1.4      v readr      2.1.5  
## v forcats    1.0.0      v stringr   1.5.1  
## v ggplot2    3.5.1      v tibble    3.2.1  
## v lubridate  1.9.4      v tidyr     1.3.1  
## v purrr      1.0.4
```

```
## -- Conflicts ----- tidyverse_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag()     masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

```
## Loading required package: airports  
## Loading required package: cherryblossom  
## Loading required package: usdata
```

```
library(statar)
```

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

```
glimpse(hfi)
```

```
names(hfi)
```

Exercise 1

What are the dimensions of the dataset?

```
dim(hfi)
```

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

There 1458 observations/rows and 123 dimensions/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?

I would be comfortable using a linear model if the relationship between the independent and dependent variable is linear. Which is a straight line relationship between the variables.

```
ggplot(data = hfi, aes(x= pf_expression_control, y = pf_score)) +  
  geom_point(aes(color = pf_score), alpha = .7)+  
  geom_smooth(method = 'lm',  
             se = FALSE,  
             color = "red")+  
  labs(title = "Freedom score VS Expression Control",  
       x = "Expression control score (0-10)",  
       y = "Personal Freedom Score") +  
  theme_minimal()
```

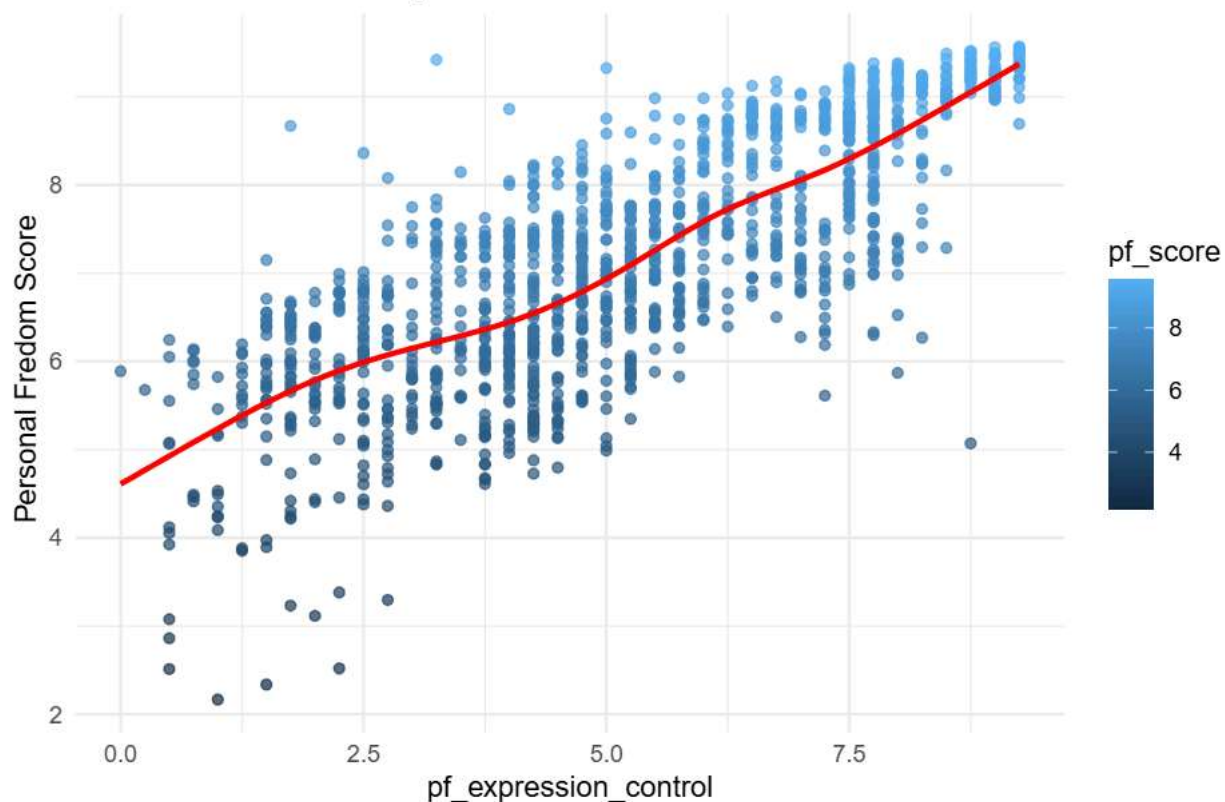
```
## Warning in geom_smooth(method = "lm", se = FALSE, color = "red"): Ignoring  
## unknown parameters: `method`
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 80 rows containing non-finite outside the scale range  
## (`stat_smooth()`).
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range  
## (`geom_point()`).
```

Freedom score VS Expression Control



```
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.

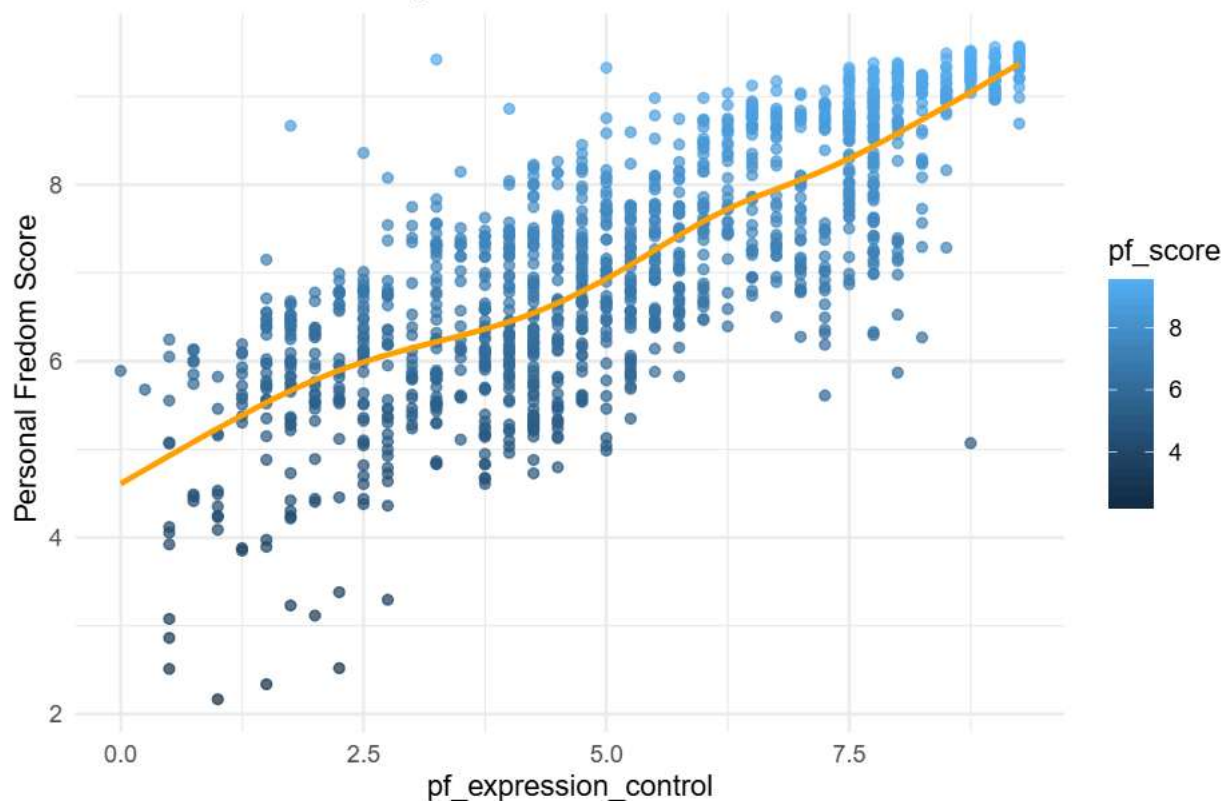
The relationship and correlation between the two variables is positive. As one increase the other increases as well. There are some outliers on the bottom left of the plot.

Exercise 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?

```
hfi2 <- hfi %>%
  select(pf_expression_control, pf_score) %>%
  na.omit()
ggplot(data = hfi, aes(x = pf_expression_control, y = pf_score, showSquares = TRUE, showSquares = TRUE))
  geom_point(aes(color = pf_score), alpha = .7) +
  geom_smooth(method = 'lm',
              se = FALSE,
              color = "orange") +
  labs(title = "Freedom score VS Expression Control",
       x = "Expression control score (0-10)",
       y = "Personal Freedom Score") +
  theme_minimal()
```

```
## Warning: Duplicated aesthetics after name standardisation: showSquares
## Warning in geom_smooth(method = "lm", se = FALSE, color = "orange"): Ignoring
## unknown parameters: `method`
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## Warning: Removed 80 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

Freedom score VS Expression Control



call

```
lm(formula = y-x, data = pts)
```

coefficients:

(intercept) x

4.6171 0.4914

sum of Squares: 952.153

The smallest sum of squares is 952.153

```
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?

$y = 5.153687 + .349862 * \text{pf_expression_control}$

The positive correlation slope implies human freedom declines as political pressure on media content increases.

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

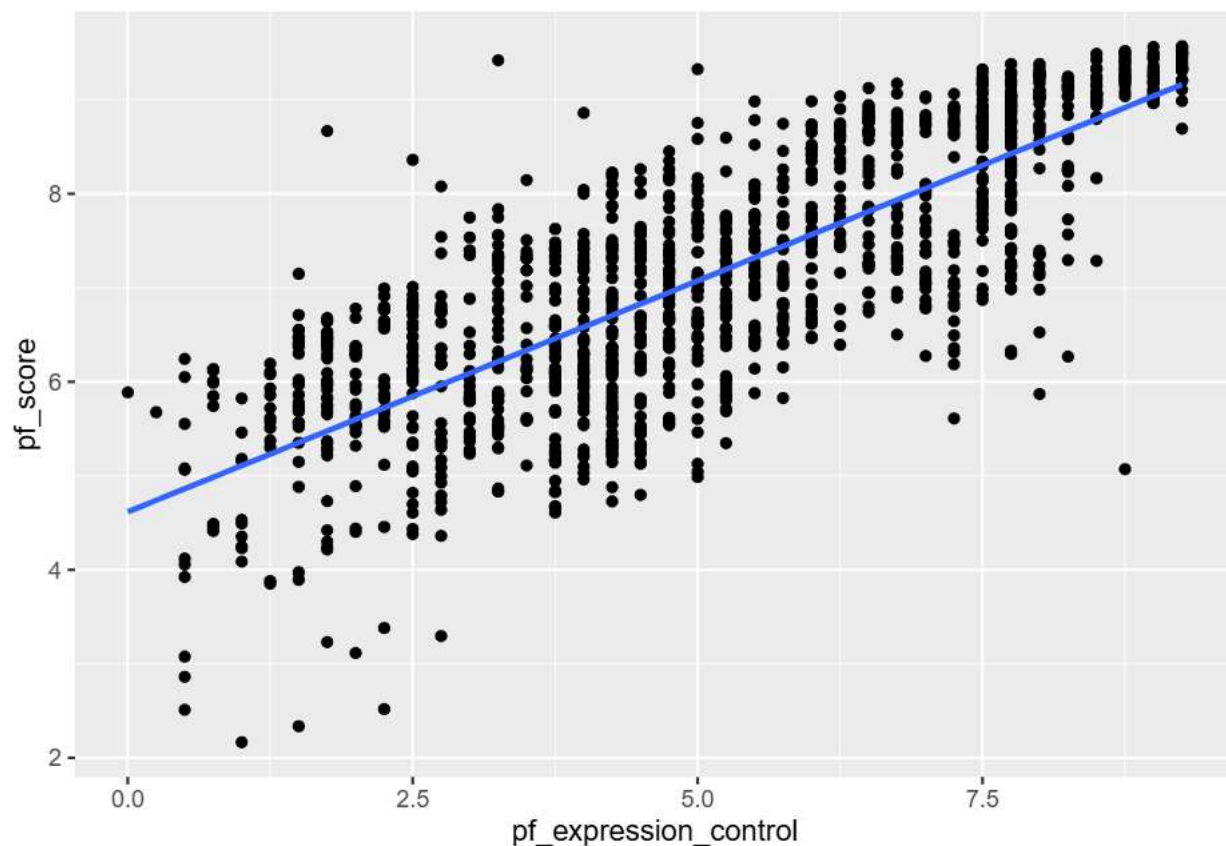
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 80 rows containing non-finite outside the scale range
```

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

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
```

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



Exercise 6

If someone saw the least squares regression line and not the actual data, how would they predict a country's personal freedom score 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?

To calculate the residual prediction. We find the difference between the actual value and predicted value. ($y - \hat{y}$) for any given point.

predicted value

```
Pv <- 4.61707 + 0.4913 * 6.7
```

find any actual value of `pf_score` with `pf_expression_control` that is = 6.7

```
hfi3 <- hfi %>%
  filter(pf_expression_control == 6.7) %>%
  select(pf_expression_control, pf_score, countries)
hfi3

## # A tibble: 437 x 3
##   pf_expression_control pf_score countries
##   <dbl> <dbl> <chr>
## 1 7.75 9.18 Australia
## 2 8 9.25 Austria
## 3 7.25 7.45 Bahamas
## 4 7.5 7.71 Barbados
## 5 9.25 8.99 Belgium
## 6 6.75 7.43 Belize
## 7 7.25 7.50 Benin
## 8 7 7.46 Burkina Faso
## 9 8.25 9.15 Canada
## 10 7.75 7.99 Cape Verde
## # i 427 more rows
```

find any actual value of `pf_score` with `pf_expression_control` that is = 6.7

```
Actual <- hfi %>%
  filter(pf_expression_control == 6.7) %>%
  select(pf_expression_control, pf_score)
Actual

## # A tibble: 0 x 2
## # i 2 variables: pf_expression_control <dbl>, pf_score <dbl>
```

prediction residual for belize

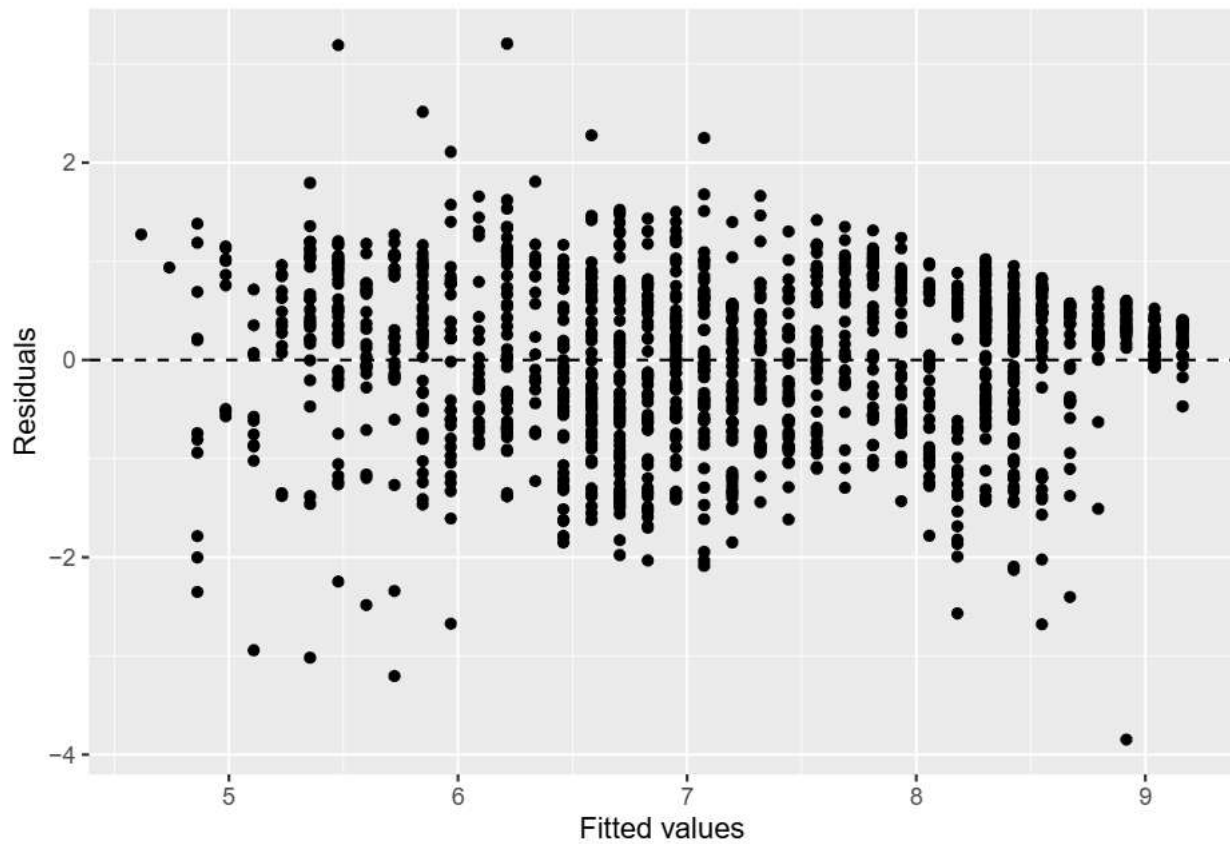
```
Prediction_ex <- 7.430864 - Pv
Prediction_ex
```

```
## [1] -0.477916
```

In this case the residual would be slightly overestimated.

```
ggplot(data = m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
```

```
geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```

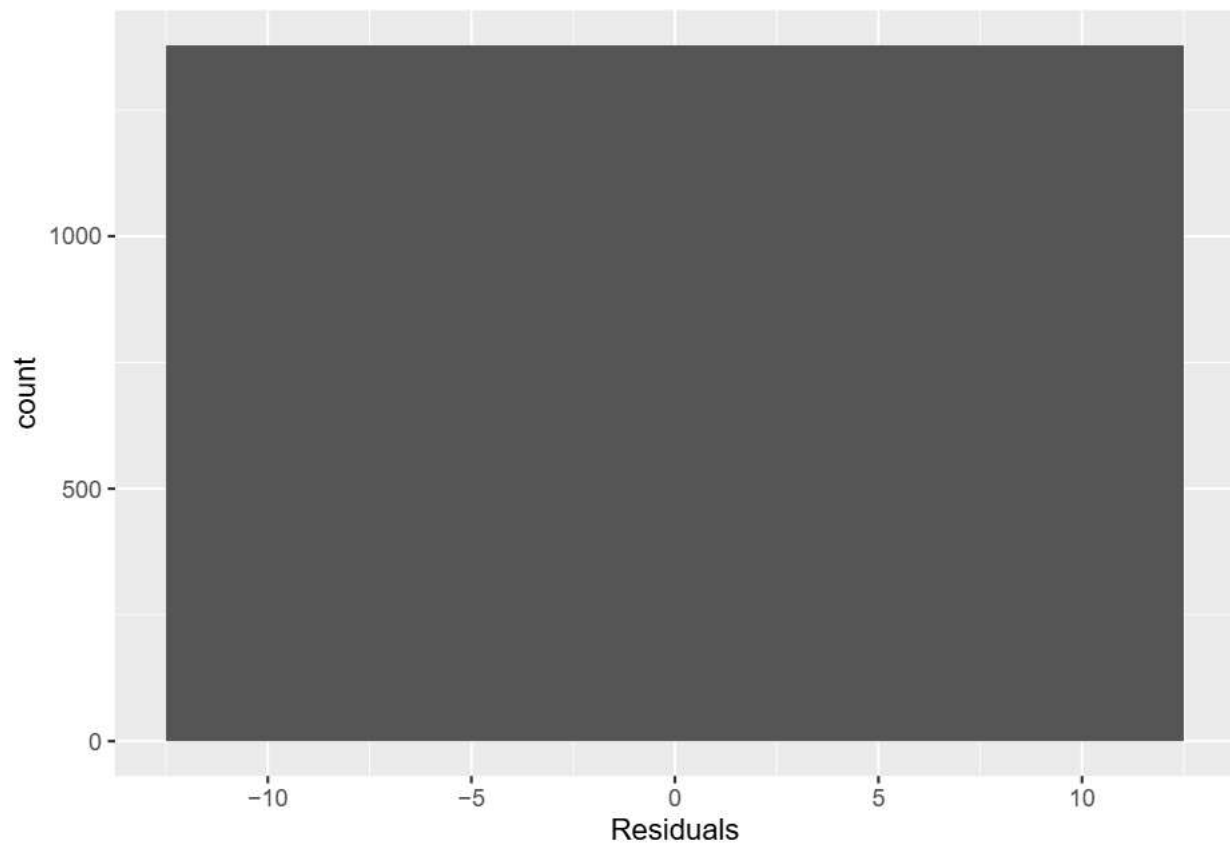


Exercise 7

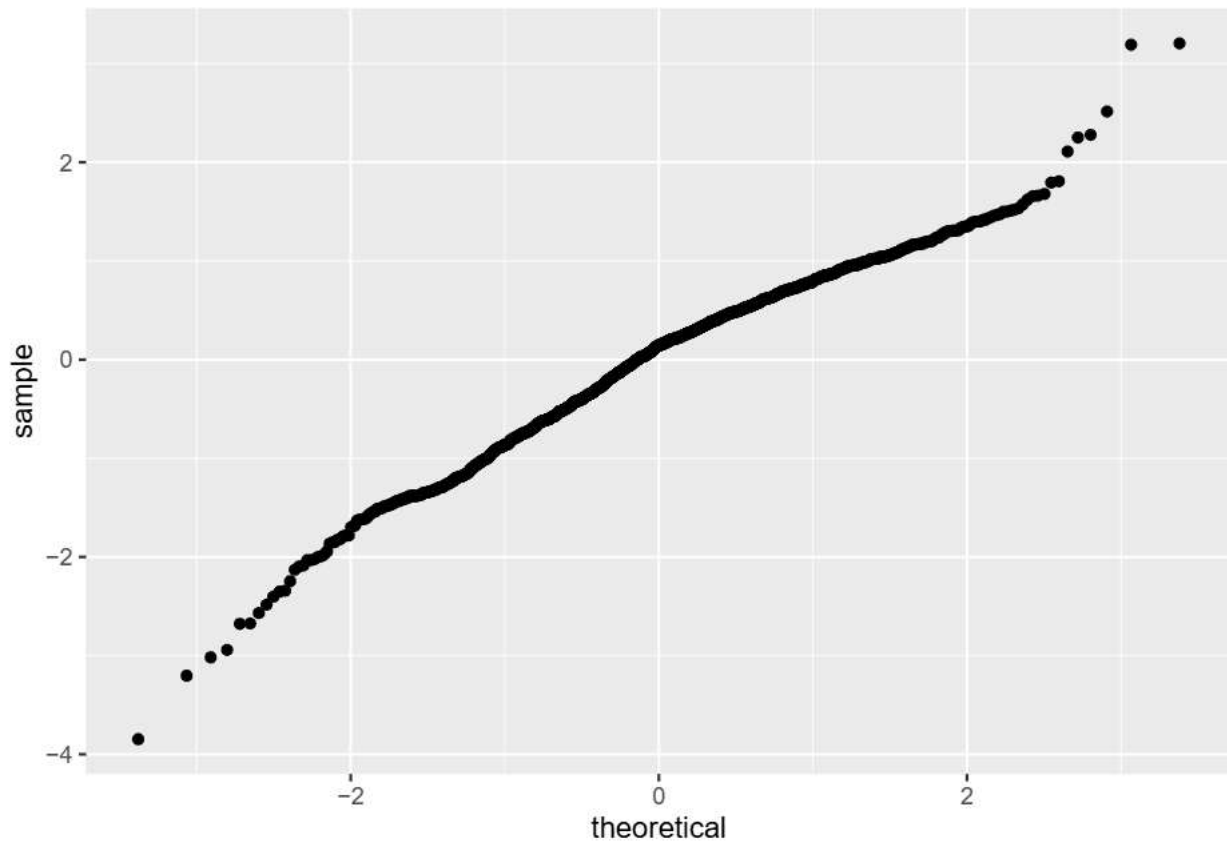
Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between the two variables?

The horizontal line signifies that the independent variable x has no predictive power for the dependent variable y . Meaning the slope of regression line is 0. So a change in x does not lead to a change in y .

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



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

Exercise 8

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

Yes. The residuals are considered normally distributed because the difference between the sample values and theoretical values are normally distributed.

Exercise 9

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

Yes constant variability is met because the scattering residuals are only around 0 and at the top with no trend in the residual plot.

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?