Lab 10 – Linear functions

ENVX1002 Handbook

Semester 1, 2025

Learning Outcomes

- Fit simple linear models and obtain associated model summaries in R
- Overlay fitted models onto scatterplots in R
- Undertake hypothesis testing to determine if slope $\neq 0$
- Check assumptions are met prior to assessing model output
- Assess model summary in terms of fit and P-values

Before you begin

Create your Quarto document and save it as `Lab-10.Rmd` or similar. The following data files are required:

1) ENVX1002_wk10_practical_data_Regression.xlsx

Last week you fitted models in R, now it is time to understand what the output means.

Before you begin, ensure you have a project set up in your desired folder. Then open up a fresh R markdown and save the file within this folder.

Don't forget to save as you go!

Exercise 1: Walkthrough - Fertiliser

Like last week, we will start off our R modelling journey by fitting a model to the fertiliser data.

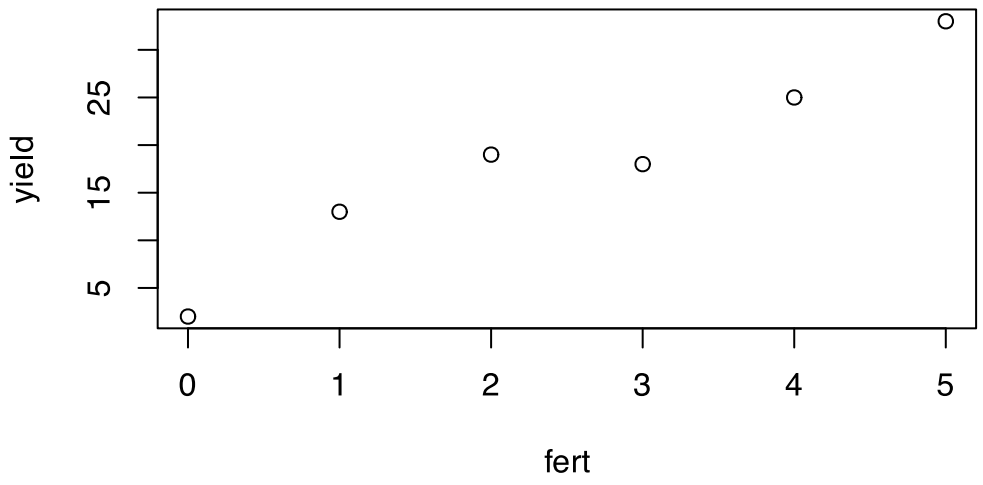
a) Read the following code into R:

```
# add the data to R Studio
fert <- c(0, 1, 2, 3, 4, 5)
yield <- c(2, 13, 19, 18, 25, 33)
```

1.1 Scatterplot and correlation

To visually identify any trends or relationships, last week we created a scatterplot of the data. This is helps us visually understand our points so we know what we might expect from the model, and possibly identify if the relationship is looking non-linear.

```r
# Create a scatterplot
plot(fert, yield)
```



Remembering back to last week, we then calculated the correlation coefficient to numerically determine whether there was a relationship between fertiliser and yield.

Using the code below, we found there was quite a strong relationship between fertiliser and yield (0.964):

```r
# Correlation coefficient
cor(fert, yield)
```

```
[1] 0.9636686
```

1.2 State hypotheses

Remembering back to the lecture and tutorial, the general equations for our hypotheses are:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

In the context of our data, the hypotheses would be:

H_0: Slope = 0; fertiliser is not a significant predictor of yield.

H_1: Slope $\neq$ 0; fertiliser is a significant predictor of yield.

If P > 0.05, we fail to reject the null hypothesis that the true slope (β_1) is equal to 0. If this is the case, it means our model does not predict better than the mean of our observations, and so there is no advantage to using our model over the mean of y ($\bar{y}$).

If we find there is a high probability of the slope not being equal to 0 (P < 0.05), we can reject the null hypothesis and conclude our model is better at predicting than the mean of our observations.

Now we understand what we are testing for, we can fit the model.

1.3 Fit the model

After checking the correlations and scatterplot, we need to fit the model using the `lm()` function. Remember to tell R the name of the object you want to store it as (in this case, `model.lm <-`), then state the name of the function. The arguments within the function (i.e. between the brackets) will be `yield ~ fert`, with `yield` being the response variable and `fert` being the predictor.

```
# Run your model
## yield = response variable (x)
## fert = predictor variable (y)
model.lm <- lm(yield ~ fert)
```

1.4 Check assumptions

This time, before obtaining our model summary, we need to check our assumptions.

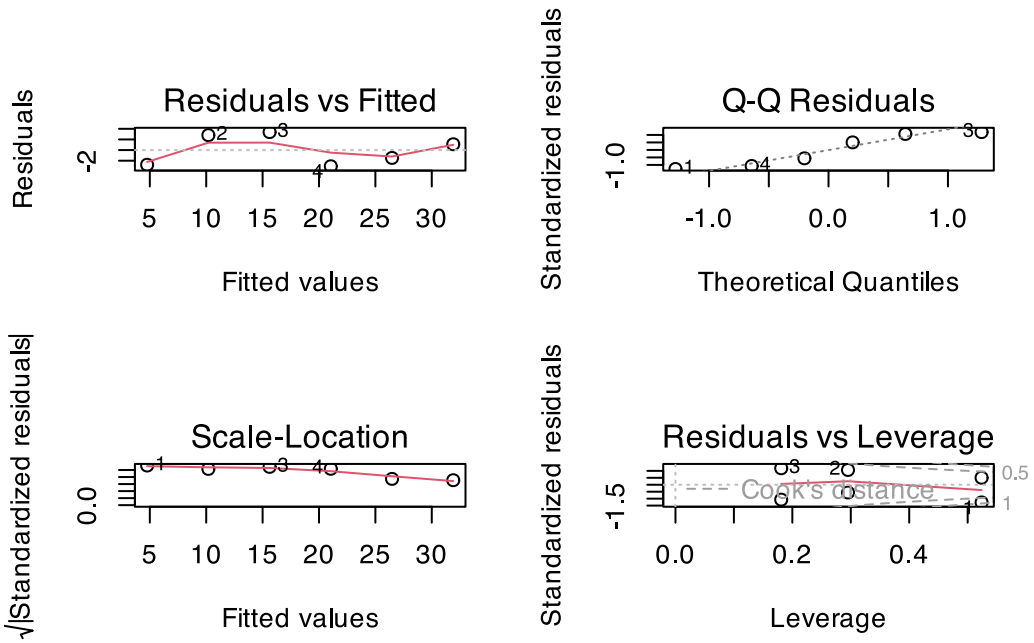
Smaller sample size (n = 6) makes it harder to check whether the assumptions have been met, but we will still run through the check.

Looking at each plot, we can see that the residual plots don't look the best;

- Residuals vs fitted: Will tell us if the relationship is linear. We are looking for an even scatter around the mean, and red line should be reasonably straight. In this case the red line is not too straight, but the scatter seems even.
- Normal Q-Q: If the residuals are normally distributed, most of the points should lie along the dotted line. Our points follow the line, but do not lie on it.
- Scale-Location: This is for testing whether the variance is equal in the residuals at each value of x. If the variance is equal, then we would expect to see an even scatter and no fanning. In this case, there is no fanning.
- Residuals vs Leverage: This will help us identify whether there are any single points influencing the slope or intercept of the model. We can see in the output plot there is a point sitting in the bottom-right corner, outside the dotted line, indicating that it may be having an influence on the model.

These plots are only useful as an example of how to obtain and interpret output. If we wanted to obtain a more reliable check of our assumptions (and a more reliable model), we would need a larger sample size (n > 10).

```r
# Check your assumptions!!
par(mfrow = c(2, 2)) # sets plots to show as 2x2 grid
plot(model.lm)
```



In this case, we will assume the assumptions have been met and continue to assess the model output.

1.5 Model output

Use the `summary()` function to obtain output for your model:

```r
# Obtain model summary
summary(model.lm)
```

```
Call:
lm(formula = yield ~ fert)

Residuals:
     1      2      3      4      5      6
-2.762  2.810  3.381 -3.048 -1.476  1.095

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7619     2.2778   2.091  0.10476
fert          5.4286     0.7523   7.216  0.00196 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.147 on 4 degrees of freedom
Multiple R-squared:  0.9287,    Adjusted R-squared:  0.9108
F-statistic: 52.07 on 1 and 4 DF,  p-value: 0.001956
```

In the model output obtained from `summary(model.lm)` the model parameters will be listed under 'Estimate' for the intercept and 'fert'. Last week we concluded the equation to be:

$$Yield = 4.7619 + 5.4286 * fert$$

Furthermore, from our model estimate, we can say that as fertiliser increases by 1, yield will increase by 5.4286.

1.6 Is the model useful?

When looking at the model summary output, we obtain the p-value from the coefficients table. We are interested in the P-value for `fert` and not the intercept.

The significance of the intercept P-value depends on our scientific question. We only really look at our intercept P-value when we want to extrapolate our line to the intercept, and know if the intercept is equal to zero (H_0) or not (H_1). This depends on your dataset and whether it makes sense to do so.

Also notice how the p-values for the F-test at the bottom of the summary output, and the t-test p-values we are using are the same. The F-test gives us an idea whether our overall model is significant and in this case, as we are only using a single predictor, the P-values will be the same.

Therefore we can say the following:

Observing the model output, we can see that the P-value for fert is significant (P = 0.00196) and we can say that as P < 0.05, we reject the null hypothesis. We can conclude our slope is not 0 and our model is a better way to predict yield than the mean of our observations.

1.7 How good is the model?

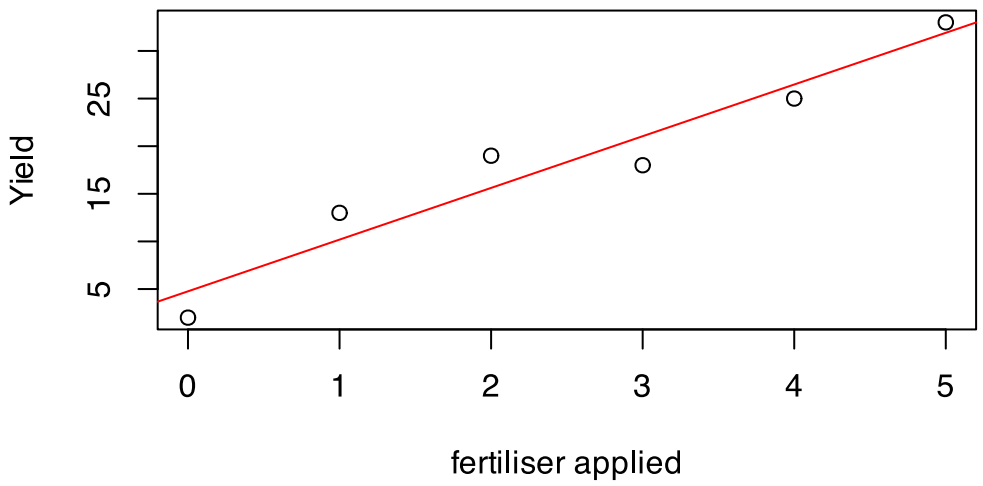
To assess how well the model fits the data, we need to look at the Residual standard error (3.147) and the r-squared value (0.9287).

- We can say that our residual standard error is relatively low in terms of our response variable.
- Our r-squared indicates that fertiliser accounts for 92.9% of variation in yield. That's pretty good!

Note how Multiple R-squared and Adjusted R-squared are similar. For simple linear models we can opt for the multiple r-squared, but when using multiple predictors we need to use adjusted r-squared.

Finally, to visually present our results, we can provide a scatterplot with the model overlaid.

```r
# Add the linear model to your scatterplot
plot(fert, yield, xlab = "fertiliser applied", ylab = "Yield")
abline(model.lm, col = "red")
```



1.8 Our conclusions

Now we can put our interpretations together to form the conclusion:

Observing the model output, we can see that the P-value for fert is significant (P = 0.00196) and we can say that as P < 0.05, we reject the null hypothesis. We can conclude our slope is not 0 and our model is a better way to predict yield than the mean of our observations.

We can therefore conclude that fertiliser is a significant predictor of crop yield as the slope is not equal to zero (P < 0.05), and it accounts for 92.9% of the variation in yield.

Exercise 2: Toxicity in peanuts

Data: *Peanuts* spreadsheet

The data comprise of, for 34 batches, the average level of the fungal contaminant aflatoxin in a sample of 120 pounds of peanuts and the percentage of non-contaminated peanuts in the whole batch.

The data were collected with the aim of being able to predict the percentage of non-contaminated peanuts (`Peanuts$percent`) from the aflatoxin level (`Peanuts$toxin`) in a sample. We will now investigate whether this is the case.

First thing's first! Let's read in the data using `read_xlsx` command:

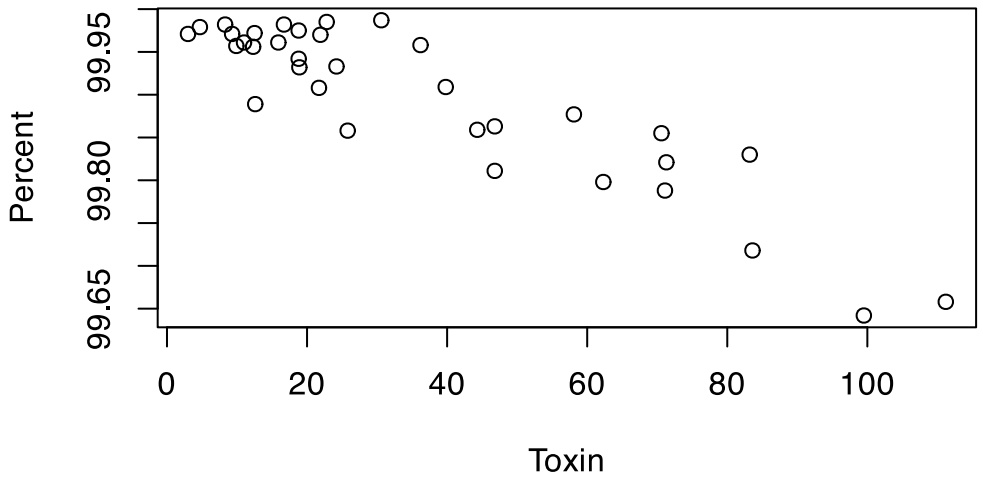
```
library(readxl)
Peanuts <- read_xlsx("data/ENVX1002_wk10_practical_data_Regression.xlsx", sheet
= "Peanuts")
head(Peanuts)
```

```
# A tibble: 6 × 2
  Percent Toxin
    <dbl> <dbl>
1    100.   3
2    100.   4.7
3    100.   8.3
4    100.   9.3
5    100.   9.9
6    100.  11
```

2.1 Scatter plot

Make a scatter plot of the data.

```
plot(Percent ~ Toxin, data = Peanuts)
```




```
#Alternate syntax:
#plot(Peanuts$Percent, Peanuts$Toxin)
```

a) Describe the relationship between the two variables.

b) Would you say that the percentage of non-contaminated peanuts in a batch could be predicted accurately from the level of aflatoxin in a sample via a linear relationship?

2.2 State Hypotheses

a) What are the hypotheses we are testing? State them as the formulae and in the context of the study.

2.3 Fit a linear model

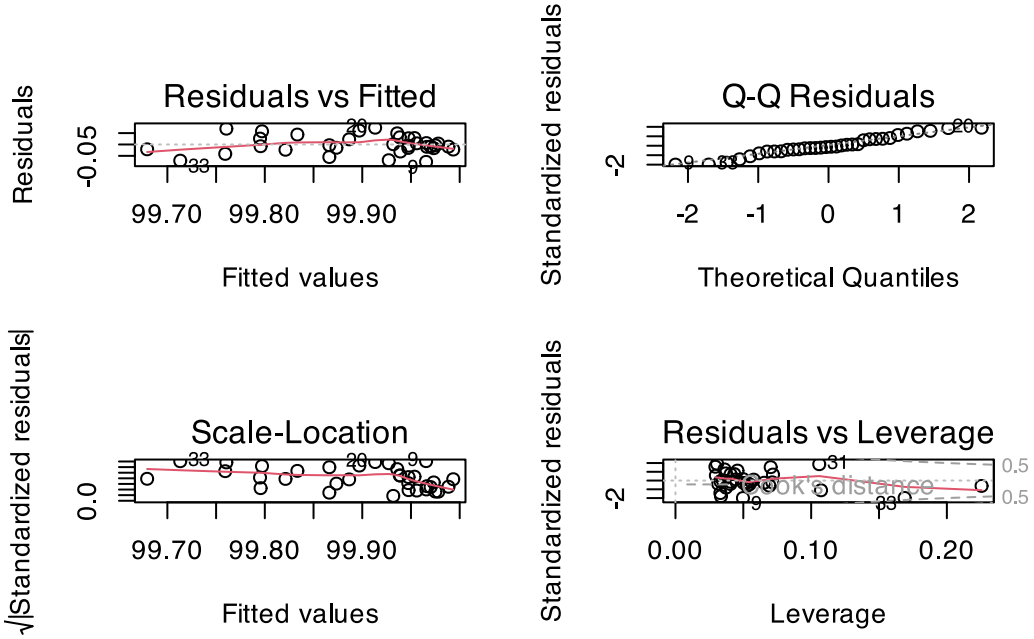
Use fit a linear model (`lm()`) to the Peanut data.

```
# fit a linear model using lm()
mod <- lm(Percent ~ Toxin, data = Peanuts)
```

2.4 Check assumptions

a) Inspect and comment on the residual plots- have the assumptions been met?

```
par(mfrow = c(2, 2))
plot(mod)
```



2.5 Observe model output

Once you are certain the assumptions are met, you can proceed to look at the regression output.

a) Comment on the overall fit of the regression, i.e. Is the model fit good? Is the model significant, and how much variation in percentage of non-contaminated peanuts does aflatoxin level account for?

```r
# Look at output with summary
summary(mod)
```

```
Call:
lm(formula = Percent ~ Toxin, data = Peanuts)

Residuals:
      Min        1Q    Median        3Q       Max
-0.076516 -0.020012 -0.004806  0.027094  0.073747

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.000e+02  1.089e-02 9184.91  < 2e-16 ***
Toxin       -2.903e-03  2.335e-04  -12.44 8.54e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03933 on 32 degrees of freedom
Multiple R-squared:  0.8285,    Adjusted R-squared:  0.8232
F-statistic: 154.6 on 1 and 32 DF,  p-value: 8.538e-14
```

b) Is toxin a significant predictor of percentage non-contaminated peanuts? If so, how can we tell?

c) Interpret the slope parameter in terms of quantifying the relationship between toxin and percent.

Exercise 3: Dippers

Data: *Dippers* spreadsheet

The file, Breeding density of dippers, gives data from a biological survey which examined the nature of the variables thought to influence the breeding of British dippers.

Dippers are thrush-sized birds living mainly in the upper reaches of rivers, which feed on benthic invertebrates by probing the river beds with their beaks.

Twenty-two sites were included in the survey. For the purpose of fitting a simple linear model, the dataset has been reduced to two variables:

- The number of breeding pairs of Dippers per 10 km of river
- The numbers of caddis fly larvae (Log(Number+1) transformed)

Now it is your turn to work through the steps as above. Does the number of caddis fly larvae influence the number of breeding pairs of Dippers?

a) Read in the data from today's Excel sheet, the corresponding sheet name is "Dippers"

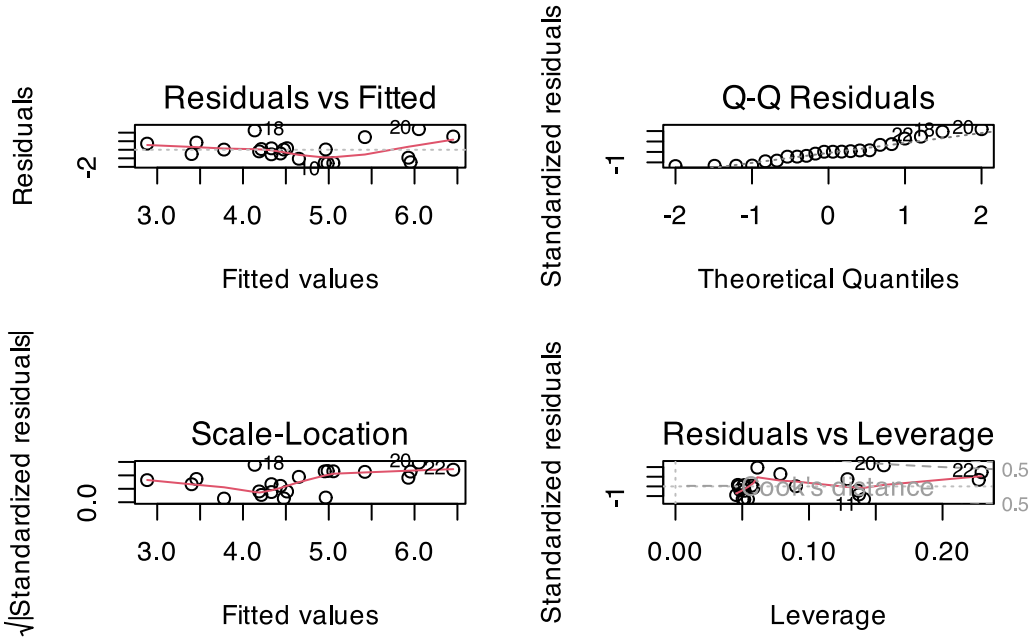
b) Obtain a scatterplot, are there signs of a relationship between breeding pair density and caddis fly larvae?

c) What are the hypotheses we are testing? State them as the formulae and in the context of the study.

d) Let's investigate further. Run the model, but before looking at our model output, are the assumptions ok?

```
# Run model
dipper.lm <- lm(Br_Dens ~ LogCadd, data = Dippers)

# Check assumptions
par(mfrow = c(2, 2))
plot(dipper.lm)
```



e) Once you are happy assumptions are good, you can use `summary()` to interpret the model output.

f) What is the equation for our model, incorporating our coefficients?

g) Based on the F-statistic output, is the model significant? How can we tell? Is it different to the significance of LogCadd?

h) Is LogCadd a significant predictor of Dipper breeding pair density? How can we tell?

i) How good is the fit of our model?

j) What conclusions can we make from this model output?

k) A final thought; Does our result make sense within the context? i.e. why might the Dipper breeding pair density be related to LogCadd?

Great work fitting simple linear models! Next week we step it up with *multiple* linear regression.

Bonus take home exercises

Use the template below to test the hypotheses for each exercise.

1. Scatterplot and correlation
2. State Hypothesis
3. Fit the model
4. Check assumptions
5. P-value and model fit
   a) Is the model significant?
   b) Is the predictor significant?
   c) How good is the model fit?
6. Conclusions

Exercise 1: Cars stopping distance

Use the `cars` dataset from last week to test if `speed` (mph) is a predictor of `dist` (stopping distance, fft).

```
head(cars)
```

```
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

Exercise 2: Penguins

Use the palmer penguins dataset to test if `flipper_length` is a significant predictor of `bill_length`.

```
#Load libraries
library(palmerpenguins)
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
#Clean data
penguins <- penguins%>%
  na.omit()#remove missing data

head(penguins)
```

```
# A tibble: 6 × 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           36.7          19.3               193        3450
5 Adelie  Torgersen           39.3          20.6               190        3650
6 Adelie  Torgersen           38.9          17.8               181        3625
# ℹ 2 more variables: sex <fct>, year <int>
```

Exercise 3: Old Faithful Geyser Data

Using the inbuilt `faithful` dataset, test whether waiting time (`waiting`) is a significant predictor of eruption time (`eruption`).

```
head(faithful)
```

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```