Lab 2: Starting multiple linear regression

Luis A. Apiolaza
School of Forestry, University of Canterbury

Assignment due on 7^{th} August 2024

1 Lab activities

In this lab you will work with a data set on biometric features of FORE224-STAT202 students. Students recorded their `weight_kg` (response variable in kg), and the following predictors: `height_cm` (in cm),`sex`, `handspan_cm` (distance in cm between the tips of their thumb and little finger), `finger2_cm` (length of index in cm) and `finger4_cm` (length of annular in cm). We will ignore `sex` in this lab.

Remember to create a folder for this assignment and to create a new project in that folder. This way R will know where to read files from and where to save output files.

In this assignment you will read a MS Excel file, for which you will need to use the `read_excel()` function from the `readxl` package. If you do not have the package installed go to `Tools - Install packages` in RStudio to install it.

Your script will need to start with the following code:

```
library(tidyverse)
library(readxl)
library(performance)
```

1. Read the `biometrics.xlsx` file and assign it to the name `biomet`.

2. Create a scatterplot where `weight_kg` is the response and `height_cm` is the predictor. Note: all the scatterplots in this assignment must use theme `theme_bw()`.

3. Remove the two outliers using code, also remove any observations that have missing values, and take a sample of size 150 observations. Assign the sample to a data frame called `my_biom`. You will use `my_biom` for the following questions. Remember to use `set.seed()` with your student number before taking the sample.

4. Create a series of scatterplots between `weight_kg`, `height_cm`, `handspan_cm` and `finger2_cm` and write a 50-word comment on the relationships you observe between the variables, including direction (positive, negative) and strength. This requires either creating plots one at the time or just one scatterplot matrix. For the latter you'll need the `ggpairs()` function of the `GGally` package.

5. Fit five different linear regression models (call them `m1`, `m2` ... `m5`) using `weight_kg` as the response variable and using `height_cm`, `handspan_cm`, `finger2_cm`, `height_cm + handspan_cm` and `height_cm + handspan_cm + finger2_cm` as predictors. Notice the changes of goodness of fit when moving from models with 1, 2 and 3 predictors. Have a look at R-squared, Adjusted R-squared and residual standard errors for the models. Write a 50-word comment on the improvement of fit when moving from `m1` through `m5`.

6. Compare models m4 and m5 using the `anova(m4, m5)` function. Which model fits best?

7. Considering the best model, have a look at any potential outliers. Use the `check_model()` function to visualize the distribution of residuals. Write 50 words explaining what you learned about the residuals of your model.

8. Now we will use the `tricarpa.csv` data set. Read it in R and name the data frame `tricarpa`.

9. Reset the random seed to your student ID number and take a sample of 900 observations, called this sample `my_tri`.

10. Produce a scatterplot of `MOE` vs `acoustic_velocity`; that is, `MOE` in the y-axis and `acoustic_velocity` in the x-axis. This time we want both axes to go all the way to 0. You can add `+ lims(y = c(0, 18))` to the plot. Use something similar to change the scale of x, taking into account the ranges of the variables.

11. Fit a linear regression model of `MOE` vs `acoustic_velocity` and explain the meaning of the regression coefficients (using the values you estimated) in the context of the problem.

12. Create a new variable in `my_tri`, called `c_acoustic`, which is the centered version of `acoustic_velocity`. Produce a scatterplot of `MOE` vs `c_acoustic`, with the y axis going all the way to 0.

13. Fit a linear regression model of `MOE` vs `c_acoustic` and explain the meaning of the regression coefficients (using the values you estimated) in the context of the problem. Refer to the plots produced in items 10- and 12- to help in your answer.

14. Create a Word file with your graphs, code and answers and submit it in Learn. For each question present question, your code/figures and your comments.