STAT202 Assignment 2: Introduction to multiple linear regression

Due on 11th August, 12 noon.

In this lab you will work with with a subset of the **fish.csv**, which contains information on the nutritional value of Antarctic fish species. Each fish was captured, frozen (following ethical research guidelines), put in a blender, turned in to a smoothie and analyzed for **energy** content (the response variable), and **fat**, **protein** and **ash** content (the predictors). The file also contains values for **mass**, **length** and **species**, which we will not use in this lab.

Remember to

a. create a new folder for this assignment. After that,
b. create a new RStudio project in that folder, as you did the previous week. This way, R will know where all files are.

You are going to read all the data located at `http://stats.apiolaza.net/data/fish.csv` into a data frame called `fish`, then take a sample of 18 observations (as done in the previous assignment, based on your student ID). Remember to use set.seed() at the beginning of your R script.

```
# Read student data set
library(tidyverse)

# Use YOUR student id number instead of 999999 to get
# your sample of the observations

set.seed(999999)

fish <- read_csv(URL between quotes here)

my_fish <- apply sampling here
```

Now my_fish contains the data you will use in your assignment.

1. Create a series of scatterplots between **energy** (as response, on y), and fat content (**fat**), protein content (**protein**) and ash content (**ash**). Write a 50-word comment on the relationships you observe between the variables, including direction (positive, negative) and strength (weak, medium, strong). Tip: when plotting you may want to try transparency of the points `geom_point(alpha = number)` with number between 0.1 and 1.

2. Fit 5 different linear regression models (call them m1, m2, m3, m4 and m5) using **energy** as the response variable and using protein, fat, ash, protein + fat, and protein + fat + ash as predictors respectively. Notice the changes of goodness of fit when moving from single-predictor models (m1 to m3), to two predictors (m4) and three predictors (m5). Have

a look at Rsquared, Adjusted R-squared and residual standard errors for the model. Write a 50-word comment on the improvement of fit when moving from m1 through m5.

3. Considering the regression with three predictors (m5) have a look at any potential outliers. Use plot(m5) to visualize the distribution of residuals. Write 50 words explaining what you learned about the residuals of your model.

4. Create a new version of the m5 model (call it m6) but using a centered version of the predictors. Create `cent_protein`, `cent_fat` and `cent_ash` in your `my_fish` dataset and use them as predictor for energy.

5. Produce the summary for m6 and the residual plots for m6 and write 30 words explaining any difference compared to m5. Explain the meaning of the intercept and slopes.

6. Create a Word file with your graphs, code and answers and submit it to Learn.

Note: another way to create the list of plots for question 1 is to use the function `ggpairs` from the package `GGally` (look it up in the web). When listing the variables to plot put **energy** at the end, so it is always on the y axis.