MATH 4044 – Statistics for Data Science

Assessable Practical Exercise 2 (SP5 2022)

Due 28 August 2022 by 11:59pm

Instructions:

- This exercise is worth 2.5% of your final mark and it is due no later than **11:59pm on Sunday 28 August** in Week 5.
- The exercise will be marked out of 20.
- You will need to submit your **individual** work via Learnonline as a **single file**, in either a Microsoft Word (doc or docx) or pdf file format. Your submission should consist of the SAS output you have generated (40%), plus the requested interpretation (50%). Please include only the most relevant SAS output (10%).
- You are welcome to discuss the exercise, give advice and share tips with other students, but there should be no sharing of files or output.

Assessment task:

Breakfast cereals are big business. According to *Choice* magazine, in 2011 '…we spent $1.17 billion on ready-to-eat cereal, and munched our way through 169,470 tonnes of it – that's about 10 large (750g) boxes for every man, woman and child…' But are breakfast cereals healthy? One variable of particular interest is the amount of sugar, which plays an important role in the tastiness of the product but can make for a less than healthy breakfast. The data file for this exercise contains nutritional information and ratings for 77 breakfast cereals.

(a) Apply the log transformation to variable `rating` to create a new variable `Lrating`. Check Normality of `rating` and `Lrating` and briefly discuss the effect of the log transformation on the distribution.

(b) Obtain a Pearson correlation matrix relating variables `Lrating`, `sugars`, `fiber` and `sodium`, and comment briefly on these correlations.

(c) Obtain a scatterplot matrix relating `Lrating`, `sugars`, `fiber` and `sodium`, and briefly discuss the resulting relationships. Based on your scatterplots and results from part (b), which variable would you recommend as the single explanatory variable in a simple linear regression model for `Lrating`.

(d) Fit a simple linear regression model relating `Lrating` to the variable you have identified in part (c), with `Lrating` as the dependent variable. Interpret the model equation. Obtain and discuss fit diagnostics. Are there any observations that would require closer inspection? Explain briefly.

Data file for this exercise:

The data is stored in a SAS data file called `cereals.sas7bdat` located in `mydata` library on the SAS OnDemand server. Variables in that file are as follows:

Variable	Description
name	Name of cereal
mfr	Manufacturer of cereal where A = American Home Food Products; G = General Mills; K = Kelloggs; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina
type	C = cold, H = hot
calories	Calories per serve
protein	Grams of protein
fat	Grams of fat
sodium	Milligrams of sodium
fiber	Grams of dietary fibre
carbo	Grams of complex carbohydrates
sugars	Grams of sugar
potass	Milligrams of potasium
vitamins	Vitamins and minerals, 0, 25, or 100, indicating the typical percentage of FDA recommended
shelf	Display shelf (1 = bottom, 2 = middle, or 3 = top, counting

	from the floor)
weight	Weight in ounces of one serving
cups	Number of cups in one serving
rating	Rating of the cereals calculated from Consumer Reports, out of 100. The higher the score, the healthier the cereals