University of Stirling *MATPMD1*
Department of Computing Science and Mathematics *Autumn 2021*
MATPMD1 : Statistics for Data Science

Lab Session 5

The focus of this week is gaining experience in running the lm() function in R and understanding its results.

The data for each of the questions can be found in the file MATPMD1LabSession5Data.xlsx. To read an excel spreadsheet in using R, install the ”readxl” library and use the following commands:

```
> library("readxl")
># read the data from the excel spreadsheet
># first the Blood Pressure Prediction Data
> Lab4dataBP <- read_excel("MATPMD1LabSession5Data.xlsx",
    sheet = "BP Prediction Data")
># now the Blood Pressure Training Data
> Lab5dataBPTraining <- read_excel("MATPMD1LabSession5Data.xlsx",
sheet = "BP Training Data")
># and now the Timber
> Lab5dataTimber <- read_excel("MATPMD1LabSession5Data.xlsx",
sheet = "Timber")
># see what the column names in the data are
> colnames(Lab5data)
```

1. Assume that the data we used in Example 10.2.8 was training data to build our statistical model which using multiple linear regression is:

$$\text{Systol} = 54.2960 - 0.6256\text{Years} + 1.3125\text{Weight} \tag{1}$$

To evaluate the performance of the prediction model, we will need to validate this on an external dataset, which has not been used to estimate the model parameters. For this reason, for model validation, before fitting the regression model to the data, the dataset wass randomly split into a training dataset and a test dataset:

(a) The training dataset is used to estimate the model parameters and select the model that provides the best fit to the training dataset.

(b) The test dataset is used for validation of the model estimated from the training dataset. We can then compare the predicted values against the observed values and calculate a coefficient of determination for the test dataset, which can be used as a measure of predictive performance.

The first sheet of MATPMD1LabSession4Data.xlsx (as described above) contains 15 records to test the prediction performance of the model selected earlier

based on the 24 records used to train the model. (Remember that our reduced model is explaining 47.83% of the variability in the response variable, i.e. R2=47.83% and when penalised by the numbers of parameters Adjusted R2=42.86%.)
Use the model (Equn. 1) to calculate the predicted values. Plot the predicted versus observed values of the response variable Systol. What do you think?

We can now measure the predictive performance of the model by comparing the predicted values against the observed values of the SBP and calculate the coefficient of determination:

$$R^2 = \frac{\mathrm{SS}_{REG}}{\mathrm{SS}_{TOT}} = 1 - \frac{\mathrm{SS}_{RES}}{\mathrm{SS}_{TOT}}$$

$$\begin{aligned} \text{where} \quad \mathrm{SS}_{RES} &= \sum \left((y_i - y_{pred}) - (\overline{y_i - y_{pred}})\right)^2 \\ \mathrm{SS}_{TOT} &= \sum \left(y_i - \bar{y}_i)\right)^2 \end{aligned}$$

What value do you get?

2. We are going to return to the Timber Example 10.2.7. This was the example where we produced a multiple linear regression model to predict the Volume of wood based on Diameter and Height of the tree. In the data file, MATPMD1LabSession4Data.xlsx (as described above), you will find a second sheet of data which contains the original three columns of data, plus two more factors to consider for the Timber model: Grade and Location. (Note these two factors are completely fictitious if they don't appear to make sense in the real world.)

   First performa a 2-Way ANOVA to determine if the Grade and Location of the trees are significant factors in explaining the total variation in Volume of wood obtained.

   Now add these factors to the linear model that we developed in Example 10.2.7 and perform a model comparison to determine if they add significant new information to the model.

3. Produce a multiple linear regression model to predict Diastolic Blood Pressure using the original training data in the Example 10.2.8 (24 records) as training data (this is in the BP Training Data sheet in MATPMD1LabSession4Data.xlsx). Then use the test data in MATPMD1LabSession4Data.xlsx (15 records, as described above) to assess whether the model for predicting Systolic Blood Pressure is better than this model predicting Diastolic Blood Pressure.