Lab 4: Data Types and more Regression

When carrying out regression analysis on a dataset in R, the explanatory variable(s) used to predict the response variable can be of a qualitative nature in opposed to the conventional quantitative data types.

Thankfully, R has embedded the tools required to analyse qualitative and quantitative data types.

Variables in R are categorized as continuous, ordinal and nominal. Continuous variables contain quantitative data, wherein the variable can assume any value within a range of values.

Nominal variables are qualitative and the order in which the data is defined is not relevant. A numerical value is used to denote or represent a description of the variable.

With regards to ordinal variables, they are also qualitative however an order is implied with the definition of the data. The arrangement of the numerical values being used to represent the variable is of importance.

In the event that datasets requiring analysis contain ordinal and nominal variables, in R these variables will be treated as factors (numerical variables with qualitative connotations).

Section A

The first section of our lab will exhibit handling datasets containing qualitative data. The table below contains a patient dataset that will be entered in R as a data frame.

Patient ID	Date of Admission	Age	Diabetes	Status
0001	17/06/18	25	Type 1	Poor
0002	11/07/18	53	Type 2	Improved
0003	21/08/18	37	Type 1	Excellent
0004	17/09/18	59	Type 1	Poor

- Assign each variable names, except “Date of Admission”, a vector containing the data points (Character data points are entered as string- in “”)
- Now assign a name to store the data frame being created
- Identify variables that will be treated as factors, and use the function “factor()” to convert the data

Section B- Using Lab 4 Dataset

- Make a plot of data , using interest rates as the explanatory variable and median house price as the response variable
- Create a model to predict housing prices using interest rates. Are the results satisfactory?
- If possible, create another model which may be a better fit.
- Compare both models with regards to the significance of their respective estimated coefficients, the amount of variability that can be explained and the quality of the residuals.
- Using the more suitable model, produce the following:
    - Confidence intervals for the coefficients
    - Prediction intervals for the following interest rates: 7.0,6.5,5.8