# STAT10060 - Statistical Modelling Lab Report

Lab Report

Due Monday 3rd May 2021 at 12 noon (GMT)

## Instructions

In this assignment you will write a report constituting 50% of your final grade for the STAT10060 Statistical Modelling module. You will be provided with a unique data set upon which you will perform statistical analysis using the skills and methods you have learned in this module. You will use RStudio to do this. **This analysis should be presented in the form of a short report which makes clear the methods you implement, the assumptions underpinning these methods, and any conclusions you may draw from your analysis.** This document provides you with instructions for downloading your data set, some background on this data, guidelines for writing your report and the scientific questions to be answered.

## Data set

To access your data follow this link: https://suzywho92.shinyapps.io/STAT10060_data/.

This will bring you to a web application which allows you download your unique data set. Enter your student number in the box provided and select **Generate Data**. This should open your data set of 13 variables and 335 observations. Clicking **Download** will allow you to save your data set locally. This is a csv file and will be named `STAT10060_XXXXXXXX.csv` where `XXXXXXXX` denotes your student number.

This data set is best saved in a folder which will be your working directory for this assignment (e.g. a folder *LabReport* in My Documents). To open the data in RStudio, set your working directory to said folder (see video links on Brightspace for how to do this) and read your data in with the following command.

```r
data <- read.csv(file="STAT10060_XXXXXXXX.csv")
# Where XXXXXXXX denotes your student number.
```

When you have completed your report **upload both the report and your `R` script to Brightspace** before the deadline. The report should be submitted as a pdf (i.e. typed in Word or Latex).

## Guidelines

- Your report should be clearly titled and include your student number.
- It should have a clear structure including labelled sections.
- Your report should be **only 4 pages long**.
- The `R` script you submit should be clearly presented and well commented using #'s (see live sessions with R).
- This report will take some time. You need allow yourself enough time to carry out the analysis in R and then write up your findings. Marks will also be allocated for the quality of presentation within your report.

## Background

A pharmaceutical company in Ireland have proposed a new drug that could reduce a patient's risk of cardiovascular disease. Researchers at St Vincent's Hospital Dublin have been recruited to test this claim. They enrol some 335 patients with known high blood pressure and record their various measurements. Doctors assess each patient and label them as at Low, Medium or High risk of cardiovascular disease. The patients are randomly allocated to a group: treatment or placebo, so that the effects of the new drug can be significantly tested. After 6 months on this new drug, the patient's risk is reassessed and recorded.

The following table shows all the variables you have in your data set and a short description of each of them.

| **Variable** | **Description** |
|---|---|
| Patient ID | Patient identification number |
| Age | Age (in years) |
| Sex | Sex of the patient (Male/Female) |
| County | County of residence (ROI 26) |
| CardioRisk | Risk of cardiovascular disease (Low, Medium, High) |
| Height | Height (in cm) |
| Weight | Weight (in kg) |
| BloodGroup | Blood group (A, AB, B or O) |
| Stroke | Has the patient had a stroke (Y for yes, N for no) |
| RegularEx | Does the patient do regular exercise (Y for yes, N for no) |
| Group | What group the patient was randomly allocated to (Treatment or Placebo) |
| Cholesterol1 | Total cholesterol level (in mol/l) before the study |
| Cholesterol2 | Total cholesterol level (in mol/l) after the study |

## Questions of interest

### Exploratory Analysis

The researchers would like some summaries of the data. Answer the following questions in your report.

1. How many patients are in each of the Treatment and Placebo groups?
2. Provide a histogram and describe the distribution of age groups in your data.
3. Body Mass Index (BMI) is defined as $kg/m^2$ where $kg$ is the individual's weight in kilograms and $m$ is the individual's height in metres. Provide a histogram and describe the distribution of BMI in the whole sample. Calculate the sample mean BMI of both males and females.
4. Identify the counties with the highest and lowest mean weights.
5. What percentage of men belong to Blood Group B?
6. Calculate the mean difference in each subject's cholesterol level from the start to the end of the study. (Hint: create a new variable `choldiff`.)

### Hypothesis Testing

Now, the researchers want to compare differences between groups.

1. Is there a difference in the risk of cardiovascular disease between males and females?
2. Is there a difference in weight between the Treatment and Placebo groups?
3. Do the data suggest that the new drug reduces cholesterol level compared to the placebo?

### Model Fitting

The researchers want to explore the relationship between Cholesterol level and BMI.

1. Using the cholesterol level before the study (Cholesterol1), compute the correlation coefficient between BMI and cholesterol level.
2. Fit the regression line between BMI and cholesterol level, using cholesterol level as the response variable. Write down the equation of the estimated regresison line and your specify your parameter values.
3. What is the value of the coefficient of determination?
4. Perform a model utility test. Write down the hypothesis your are testing, the value of the test statistic and of the corresponding p-value.
   For a significance level of 0.05, what is the outcome of the test?

### Secondary Analysis

Lastly, some of the researchers have hypothesised that there is a potential relationship between height and Blood Group. They want to compare statistics across Blood Group factor levels.

1. Compute the average height for each level of the factor variable denoting the Blood Group.
2. Use ANOVA to test the hypothesis that the mean height is significantly different across the different Blood Groups.
3. Write down the value of the F statistic, the degrees of freedom and the corresponding p-value. For a significance level of 0.05, what is the outcome of the test?