Part C [15 marks] - Graphing and programming

When designing an experiment, there are standard methods for calculating the minimum sample size that is required in order to detect a specified effect size, at a specified significance level and power, given various information about the group(s) of interest. However, in some cases, standard tools do not exist, and we must write our own simulator for a specific problem in order to assess the sample size that is required.

In this part, you are going to write a program/function to simulate a model with various predictors and plot some of the results.

The model of interest is a linear-regression model with two predictors, haemoglobin (g/dL) and serum cholesterol (mg/dL) that we know affect our outcome of interest – body mass index (BMI; kg/m^2). We are interested to know how many individuals we are required to sample, in order to detect a change of 0.5 in BMI for individuals who received a particular treatment (e.g., an exercise program) compared to those who did not (i.e., “treatment” is the third predictor in the model). Equation 3, below, presents the model of interest:

$$\text{BMI} = \beta_0 + \beta_1\text{*Hgb} + \beta_2\text{*Chol} + \beta_3\text{*Treatment} + \varepsilon \qquad (3)$$

Where:

- $\beta_0 = 18.06$, $\beta_1 = 0.297$, $\beta_2 = 0.014$;
- Haemoglobin values (Hgb; g/dL) are normally distributed with a mean of 14.4 g/dL and standard deviation of 1.5 g/dL (i.e. N(14.4,1.5) distributed);
- Serum Cholesterol values (Chol; mg/dL) are N(215,50) distributed; and
- Treatment is an indicator variable, where 1 represents those that received the treatment, and 0 represents those that did not; and
- The error term, $\varepsilon \sim N(0,1)$.

1) **[10 marks, R/Stata]** For this question, we will first simulate a dataset (Part C Q1a), and then create a function/program which extends Q1ai below (Part C Q1b), and then we will calculate power for varying sample sizes (Part C Q1c).

    a. Write code that:

    i) randomly samples 100 individual’s haemoglobin and serum cholesterol from the above distributions[§],

ii) randomly allocates exactly half of the subjects to the Treatment group, and

iii) calculates their BMI according to the above equation, assuming that β_3=0.5.

Store the simulated haemoglobin, serum cholesterol, treatment, and BMI values in a data frame/dataset.

§*Note*: To generate random normal values, you will need to use `rnorm()` in R, or `rnormal` in Stata. To randomly allocate individuals to each treatment group, you will need to use `sample()` in R, or the `sample` command in Stata.

b. Using your code from Part C Q1a, write a function/program which takes as input the value for the sample size, *n*, and returns an object/data frame containing *n* simulated individuals with their haemoglobin, serum cholesterol, treatment, and BMI values.

c. Using your code from parts a and b, modify the code and include a *for* loop that:

   i. Simulates the model 1000 times; and

   ii. fits a linear regression model* of "BMI ~ Hgb + Chol + Treatment" for each simulated data set, and counts the number of times that the p-value associated with Treatment is <0.05; and

   iii. cycles through sample sizes between 40 and 200 in step sizes of 20 and performs (a) and (b) and stores the proportion of simulated experiments that had a p-value for treatment <0.05. This is an estimate of the Power (i.e., 1 – Type II error rate).

   **Notes*:

   To fit the linear regression model in R, you can use:

   ```
   lm(BMI ~ Hgb + Chol + Trt, data=df)
   ```

   To fit the linear regression model in Stata, you can use:

   ```
   regress BMI Hgb Chol Trt
   ```

Make sure that you set the random number seed so that the results are reproducible. Note that it may take a few minutes to run, so test that your code is working on a smaller number of simulations (e.g., 100-200) and sample sizes (e.g., 20, 60, 100), before running the entire loop.

2) **[5 Marks, R/Stata]** Create a figure showing the sample size on the x-axis, and the estimated Power on the y-axis. Include the following items in your figure:
   a) A horizontal line at 80% Power.
   b) Use your favourite colour (ensure that this is not the default colour) to highlight the point that corresponds to the smallest sample size such that “we have at least 80% power to detect a difference of 0.5 in treatment between the two groups at the 5% significance level”. You will need to automate this, and not manually select this value (*Hint*: you want to find the minimum sample size that has an estimated Power ≥80%).
   c) Change the y-axis to have labels from 20% to 100% power in steps of 10%. Change the x-axis to go from a sample size of 20 to 300 in steps of 40. Make sure the axes have appropriate labels.