ASSIGNMENT II

BIOSTATISTCAL INFORMATICS

Deadline: 12 noon, 19th April 2023

Submit via Canvas

Questions: j.blayney@qub.ac.uk

NOTES

1 Include a narrative (1/2 paragraphs) to explain and describe your answers. Use plots and tables where appropriate (these must be numbered, with legends and referred to in the text).

You are not expected to provide a clinical interpretation of the results.

2. R code: The entire analysis should be carried out in R. Your R code (plus annotations) will be assessed. Provide evidence of assumption testing, where appropriate. State significance levels.

3. Report: The narrative/description, tables and plots should be combined into one document.

The tables and plots must include titles where appropriate. The R code that you use MUST also be included as an appendix.

There are no maximum nor minimum word limits for this assignment, though a report of around six to twelve (dependent on line spacing, preferably set to 1.5 lines) pages of A4 (including tables, figures) would be sufficient.

The final report must be in .docx (or Open Source equivalent) or in .PDF format.

The presentation, formatting (including spelling/grammar) and structure of the report will also be assessed. **DO NOT USE SCREEN-SHOTS OF CODE.**

Question		**Weighting**	**Subtotal**
1	5	X1	5
2	5	X3	15
3	5	X4	20
4a	5	X3	15
4b	5	X3	15
R Code	5	X3	15
Report	5	X3	15

BACKGROUND

For this assignment, you will look at an adaptation of a breast cancer dataset (**assignOct3.txt**). There are 686 patients in this dataset, with five variables available (two of which are continuous, two binary and one factor), including overall survival and overall survival status.

The variables are:

Patient: Patient ID PT1 – 686

Age: Age at Diagnosis, Years

MenopauseStatus: Menopausal Status 1 = Pre, 2 = Post

HormoneTherapy: Hormone Therapy 1 = No, 2 = Yes

TumourSize: Tumor Size (length) mm

TumourGrade: Tumor Grade 1 – 3

Survival: Time to Death, Days

Event: Death Alive/Dead

1) Summarise the composition of the data
- *Use both a table and a descriptive format*

2) Considering all patients develop a FULL multivariate model from the clinico-pathological variables provided. Select the variables which best explain survival to establish a FINAL multivariate model.

3) A collaborator is interested if age behaves differently in the pre-menopausal and post-menopausal groups of patients. Compare and contrast the behaviour of age (using univariate and multivariate) in both groups.

4) Your collaborator defines "good" survival as those patients who have survived beyond five years and those with "poor" survival as dying before the first year.

a) Using expression levels of genes 1 to 5 in a second data file (**assignOct4.txt**), which genes, if any, have different expression levels between the "good survival" and "poor survival" patient groups?

b) Use a semi-parametric (univariate) method to consider the relationship of each gene to overall survival. Comment on the similarities/differences with the results in part a).