STAT330/430 - Statistical Learning

Assignment 2
Due date - April 11th

Note: You must submit two files for this assignment:

- a pdf document of your solutions, and
- a complete and concisely annotated *R Script* file of all **R** analysis that was undertaken to produce your results.

In addition, to receive marks for Question 1 you are required to engage in the Topic 2 moodle discussion forum.

Please refer to the Assignment assessment criteria document for additional guidance.

Question 1 [5 marks]

In each assignment you will be able to earn up to 5 marks based on your engagement on the moodle Discussion forums from the topics associated with the given assignment. See the Assignment assessment criteria document for more details.

Question 2 [10 marks]

The net reproductive rate (R_0) of an animal is defined as the total number of offspring that an individual can produce during its lifetime and can be calculated as follows:

$$R_0 = \sum_{x=0}^{\infty}(l_x m_x)$$

where l_x is the proportion of females surviving to each age and m_x is the average number of offspring produced at each age. Hence, if you had data on two different time-periods or ages you would calculate (R_0) as:

$$R_0 = (l_1 m_1) + (l_2 m_2)$$

The *aphid* data-set contains records of the number of eggs produced on three different days by each of 20 aphids.

Variable	Description
ID	Individual aphid in the experiment
Day1	Number of eggs produced on Day 1
Day2	Number of eggs produced on Day 2
Day3	Number of eggs produced on Day 3

Note: *NA* indicates that the animal has died.

Using the *aphid* data-set:

(a) Create a function to calculate R_0 for the population.

(b) Estimate a boot-strapped standard error and 95% confidence interval of R_0 using 100,000 replicates.

(c) Provide a brief summary of your result as well as fully annotated code in your *R Script* file.

Question 3 [30 marks]

The *Body* dataset contains 21 body dimension measurements as well as age, weight, height, and gender on 507 individuals. The 247 men and 260 women were primarily individuals in their twenties and thirties that all reported to exercise on a regular basis.

Variable	Description
BA_diam	Biacromial diameter
PB	Biliac diameter, or "pelvic breadth"
BI_diam	Bitrochanteric diameter
Chest_dep	Chest depth between spine and sternum at nipple level
Chest_diam	Chest diameter at nipple level, mid-expiration
Elbow_diam	Elbow diameter, sum of two elbows
Wrist_diam	Wrist diameter, sum of two wrists
Knee_diam	Knee diameter, sum of two knees
Ankle_diam	Ankle diameter, sum of two ankles
Shoulder_g	Shoulder girth over deltoid muscles
Chest_g	Chest girth
Waist_g	Waist girth, narrowest part of torso below the rib cage
Navel_g	Navel (or "Abdominal") girth at umbilicus and iliac crest
Hip_g	Hip girth at level of bitrochanteric diameter
Thigh_g	Thigh girth below gluteal fold
Bicep_g	Bicep girth, flexed
Forearm_g	Forearm girth, extended, palm up
Knee_g	Knee girth over patella, slightly flexed position
Calf_g	Calf maximum girth
Ankle_g	Ankle minimum girth
Wrist_g	Wrist minimum girth
Age	Age (years)
Weight	Weight (kg)
Height	Height (cm)
Gender	Gender (1 - male, 0 - female)

(a) Produce some appropriate exploratory graphics of the *Body* dataset. Given the number of variables ensure that all graphics are easily readable and understandable. Provide a general summary of the trends seen in your graphics. Note: These graphics do not have to be exhaustive, just useful.

(b) Split the *Body* data-set into 50:50 testing:training data-sets. Use the validation set approach to implement forward *or* backward selection to select an optimal subset of predictors of *Weight*. Provide a summary of your results which includes (but is not limited to) appropriate tables, metrics and commentary on the relative accuracy of your results.

(c) Using the testing and training data-sets created in (b) use ridge *or* lasso regression to constrain or regularise the predictors of *Weight*. You should employ cross-validation to tune the value of λ during the training phase of your model. Provide a summary of your results which includes (but is not limited to) appropriate tables, metrics and commentary on the relative accuracy of your results.