Fundamental Mathematics and Statistics for Health Data Science

Assessment 2:

*** Released on 10 Dec 2021, due on 16 Jan 2022 ***

1. Read ‘iq.txt’ into R. The data were collected from 38 subjects, with measurements of their IQ score (`PIQ`), brain size (`brain`), height (`height`) and weight (`weight`). Our interest is whether the IQ score can be predicted by the other variables.

   a. Express data by a pairwise scatter plot of the four variables. Comment on the relationship between `PIQ` and the other three variables. Test the correlation between `PIQ` and `Brain`. (10)

   b. Choose an appropriate regression model where the response is `PIQ`, and the explanatory variables include the remaining three variables. Select your final model by stepwise AIC model selection, starting with the null model. Interpret the results of the final model. What percentage of variation does the selected model explain? (20)

   c. Plot the diagnostics of the selected model and check the assumptions of the model. (10)

2. The data file ‘d.cancer.csv’ comprises 56 samples and 855 columns -1st column: cancer status (0: cancer-free; 1: cancer) and gene expression data (854 genes).

   a. Carry out principal component analysis on the gene expression data. How many components do you think give an adequate representation of variation of the gene expression data? Why? (12)

   b. Please choose a proper regression model to test if any of the first 5 principal components are predictive of cancer. Comment on the usefulness of PCA for this dataset. (13)

3. A new test for a virus becomes available. The test is claimed to have 95% sensitivity, and 89% specificity. At a given time, the virus is thought to have 3% prevalence in the population.

   a. Assuming the above figures are correct:

      i. Show that the probability of a test conducted on a random individual yields a positive result with probability 13.5%. (8)

      ii. If such a positive result is observed in an individual, what is the probability that individual truly has the virus? (3)

   b. Some public health professionals do not believe that the 3% prevalence figure is correct. To test this, n = 1000 individuals are randomly selected to be tested. Of these, 165 were tested positive.

      i. Does this support the *hypothesis* of the public health professionals, i.e. that the prevalence is incorrect? (18) (HINTS: note the word in italic, and that you will need to use information from part a)).

      ii. The prevalence is later found, through further testing, to indeed be 3%. Knowing this, how can the findings in part (i) be explained? (6)

Please specify the hypotheses clearly for each of the tests, and include your R commands with the answers.