PES University, Bangalore
UE15CS322 - Data Analytics

Session: Aug – Dec 2017
Week 4 – Assignment 4

Date of Submission: **Max Marks:** 20

NOTE: *In your assignment books, write the question, R commands required to get the desired output and the obtained output. For questions that require you to plot graphs, you will have to also print the obtained output (graphs/charts) and attach it in your assignment book. The solutions to the assignment must be* ***hand-written****.*

TOPIC: Sampling and Introduction to Filling Missing Values

Cancer is a difficult journey, both for the one who is suffering and their caregivers. Breast cancer is now the most common cancer in most cities in India, and second most common in rural areas. Breast cancer is a non-existent entity for a majority of the population until a near and dear one suffers from it. Healthcare is low on priority and even in major cities, screening is an 'alien' word for most people. So naturally, this results in most people presenting only when symptomatic, and on an average, most 'symptomatic' cancers are stage 2B and beyond (significant numbers in stages 3 and 4). So the breast cancer patients do not tend to survive for a longer time, as their western counterparts. In the West, a majority of breast cancers are diagnosed in stages 1 and 2, resulting in good survival. India needs to reach this achievement, and it is only with aggressive promotion of screening, awareness and proper treatment that India will achieve this.

Excerpts from website, 'Breast Cancer India'

The website manager, *Dr. Sumeet Shah* has also analysed 'Trends of Breast Cancer in India' and made it available on his website here. It is worth reading, not just for awareness, but also for the course to understand various ways in which data can be used, interpreted and visualized.

The dataset, Breast Cancer Wisconsin (Diagnostic) on UCI Machine Learning Repository contains the values of the features computed from a digitalized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. The description of the attributes present in the dataset can be viewed here. This dataset has been slightly modified to suit the requirements of this assignment. This modified dataset **cancer_tailored.csv (1)**[1] contains an additional column, 'predicted'. This is the prediction of a supervised learning algorithm called decision trees (which you will learn in Unit 3) based on the other attributes in the dataset, indicating if the cancer in that test case is **M**alignant or **B**enign. The column 'diagnosis' indicates the actual diagnosis of cancer in the test case.

Question – 1 (8 points)

The dataset **cancer_2015.csv (2)**[1] contains details of the number of patients admitted under breast cancer cases in 2015 every month. 12 values are missing in this dataset. Fill in these missing values using:

(a) Mean of all available samples for that attribute

[1] The assignment requires you to download files, (1) and (2) only

(b) Linear interpolation (based on the nearest available value to the top and the bottom of the chunk of missing values)
(c) Quadratic interpolation selecting the most proximate two values from the past and one from the future
(d) Linear regression to model the data available and then use the best fit line to arrive at predictions for the missing value
(e) The *package Mice* provides multiple imputations for the data. Use Mice to impute data (use 50 iterations) and plot three separate graphs – one for outpatients, the second for new registrations and the third for laboratory investigations to compare the nine versions of data you have produced (four from standard methods of filling in data and five from Mice's imputations).

Of all the four types of methods, you have resorted to in (a)-(d), which produces the closest values to any of the imputations output by Amelia?

It is sufficient for you to write the code and the 12 missing values (4 values for the three attributes) for each part. There is no need to print the graphs.

Question – 2 (10 points)

In Unit 1, you learnt other ways of handling missing data such as, dropping of rows and columns. Dataset (1) contains six missing values. Handle it by dropping off the rows containing missing data. Now, based on this dataset, answer the following questions:

(a) The dataset is still 30-dimensional which makes it difficult to analyse. Can any attributes be dropped to reduce the dimensionality of the dataset? If so, list the ones that can be dropped and the surviving attributes (with suitable reasons). What is the new dimensionality of the dataset? (Use visualizations for this)
(b) The entire series of tests and predictions were conducted to find out if the cancer in the patient was **M**alignant. Tabulate a confusion matrix for the obtained predictions. Label the corresponding values in the matrix as True Positive, True Negative, False Positive and False Negative.
(c) Calculate the following measures:
(**Note:** Most of the packages in R print the values of these measures along with the confusion matrix. Show the working done to obtain the value of these measures in the assignment book.)
    (i) Accuracy
    (ii) Misclassification Rate
    (iii) Recall
    (iv) Precision
    (v) Specificity
    (vi) F Score with $\beta = 2$ and 0.5 (also state what the corresponding values of β indicate)

Question – 3 (2 points)

Terms, 'correlation', 'covariance' and 'simple linear regression' are highly *correlated*, with subtle differences. Explain this subtle point.