INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

24th March 2022

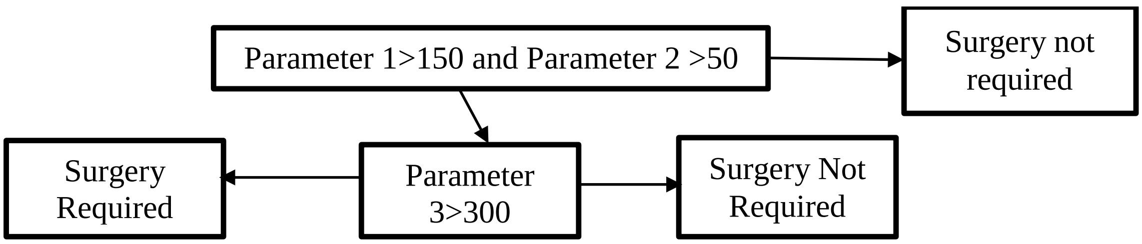
Subject CS2B – Risk Modelling and Survival Analysis (Paper B)

Time allowed: 2 Hours (14.30 - 16.30 Hours)

Total Marks: 100

Q. 1) A statistician uses Decision Tree to reach the final decision of whether there is a need to conduct a surgery for a patient or not. This is based on testing the presence of three parameters in human body. You need to test the effectiveness of the decision tree in predicting the need for surgery correctly.

Below is the decision tree as shown:



i) Write a function in R to decide whether a surgery is needed or not based on the value of three parameters. Denote Surgery required as 1 and Surgery not required as 0. (3)

ii) Check the value of function with parameter 1=180, parameter 2=75, parameter 3=350 (1)

iii) Load CS2B_Mar22_Dataset1.csv file into R and name it as *decisionData*. This data contains the three parameter values for 50 patients and the actual decision taken with respect to the surgery. Apply the function created in (i) to this data. Paste the output after applying the function in your answer sheet. (3)

iv) Compare the predicted and actual decisions by computing precision, recall and F1 score. (6)

v) Comment on the results obtained in (iv). (2)

[15]

Q. 2) An Actuary tried to predict the COVID-19 evolution by using Markov chain which involved modelling transitions from one state to another according to a transition probability matrix. These predictions are expected to help the authorities to set up adequate protocols for managing the post-confinement due to COVID-19.
There are five states in the model namely Infected, Quarantined, Hospitalised, Recovered and Death which are measured at the end of each week.

The ‘CS2B_Mar22_Dataset2.csv’ file contains the recorded weekly transitions between different states. Each row contains a pair of State_From and State_To where State_From denotes the starting state and State_To denotes the state after a week’s transition.
The following notation has been used:
I-Infected, Q-Quarantined, H-Hospitalised, R-Recovered, D-Death

Those states which have no transition to any another state can be assumed to be absorbing.

Load ‘CS2B_Mar22_Dataset2.csv’ file into R and name it as *coviddata.*

i) Using transition pair data, compute the transition probabilities between each state. (6)

ii) Find out the absorbing states in the model. (2)

iii) Create Markov chain model named “covid19” in R with the respective state space and transition probability matrix. Plot this model and paste the plot into your answer sheet. (Use library-markovchain) (4)

iv) Using the Markov chain, given that the individual is infected, compute the probability that:

a) He will be in quarantine after a week (1)

b) He will either be in hospital or in quarantine after two weeks (2)

c) He will recover after 3 weeks (2)

v) COVID-19 pandemic is expected to end in a quarter's time.
Estimate the number of people who will be (a) Quarantined (b) Hospitalized (c) Recovered and (d) Died out of the current 20,000 quarantined and 10,000 hospitalised individuals. (Answer to nearest whole number, Assume 13 weeks in a quarter) (5)

vi) Comment on your results obtained in (v). (3)

[25]

Q. 3) An ARMA(2,2) time series Z_t is defined by the following equation

$Z_t=2 + 0.8Z_{t-1} + 0.1Z_{t-2} + e_t + 0.4e_{t-1} + 0.1e_{t-2}$, where e_t are white noise terms which are IID random variables with mean 0 and $\sigma^2= 7$

i) Simulate 300 observations of the above time series using seed value equal to 100 and plot it giving appropriate labels for each axis. Paste the plot into your answer sheet. (4)

ii) Using the data simulated in (i), compute the mean and standard deviation of

a) The entire series

b) The first 150 observations of the series

c) The last 150 observations of the series (6)

iii) Plot the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) of the simulated data in (i), giving appropriate labels for each axis. Paste the plots into your answer sheet. (4)

iv) Comment on the stationarity of data from (ii) and (iii) above. (3)

v) Comment whether ARMA(2,2) can be inferred from the ACF and PACF plots. (3)

[20]

Q. 4) A random variable Z which represents claim amount in insurance is lognormally distributed with parameters μ=10 and σ^2= 4.

i) Compute

a) Likelihood that the claim will be of 5000. (to 2 significant figures) (2)

b) Probability that the claim pay-out will be greater than 5000 (to 2 significant figures) (2)

c) Maximum claim pay-out in the confidence interval [0.9,0.99] (3)

d) Median of Z (2)

e) Interquartile range of Z (2)

ii) Generate a random sample of 100 observed claim amounts using seed value equal to 50 from this distribution.

a) Calculate the mean, median, standard deviation and skewness coefficient of the sample. (4)

b) Plot a histogram of the sample vector showing the probability densities, setting the y-axis range from 0 to 10^-5 and x axis range from 0 to 10^5 for this graph and paste the plot into your answer sheet. (5)

[20]

Q. 5) A study aimed at estimating hazard rates and survival probabilities has produced the results of N_j=Number at risk and D_j=Decrements for different time intervals t_j as shown in CS2B_Mar22_Dataset3.csv. The researchers analysed the results using Nelson-Aalen model of comparison.

i) Write down the formulae for the estimated integrated hazard and the estimated variance of this estimator, using the Nelson-Aalen model. (2)

ii) Find the estimated integrated hazard and the estimated variance of this estimator for the given data for each time t_j. (6)

iii) Produce a scatter plot in R along with corresponding 95% confidence intervals for the Nelson-Aalen model. Paste the plot in the answer. (6)

iv) Compute the survival function according to the Kaplan Meier estimate. (3)

v) Demonstrate numerically the validity of the inequality relationship between the survival function using the Nelson Aalen model and the Kaplan Meier model using the data in this study. (3)

[20]

****