INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

25th May 2023

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

Time allowed: 1 Hour 45 Minutes (14.45 – 16.30 Hours)

Total Marks: 100

INSTRUCTIONS TO THE CANDIDATES

1. ***Mark allocations are shown in brackets.***

2. ***Attempt all questions beginning your answer to each question on a new page.***

3. ***Attempt all sub-parts of the question in one document only, unless otherwise instructed to do so.***

4. ***All the detailed guidelines are available on exam screen.***

5. ***Do save your work in solution template on a regular basis.***

6. ***If Any, Data set file(s) accompanying the question paper is available for download on the exam screen.***

7. ***You need to import the same into R studio as soon as you begin the exam.***

8. ***Ensure to copy and paste R codes and output at regular intervals onto the solution template.***

9. ***Please check if you have received complete Question Paper and no page is missing. If so, kindly get new set of Question Paper from the Invigilator.***

AT THE END OF THE EXAMINATION

Please return this question paper to the supervisor separately. You are not allowed to carry the question paper in any form with you. You are requested to save and submit the work before leaving the examination premises.

Q. 1) The data set “Monthly_Corn.csv” contains monthly average closing prices and the volume of corn traded on the Chicago Mercantile Exchange from January 2017 to March 2023. Load the data into R.

i) Create a time series of the closing price by using an appropriate R function. (1)

ii) Plot this series, labelling each axis appropriately. (2)

iii) Based on visual inspection of the series, comment on whether the series is stationary or not. (1)

iv) Plot the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) of the closing prices, labelling each axis appropriately. (4)

v) Comment on the stationarity of the closing price by observing the plots in part (iv). (2)

vi) Create another time series “monthly_returns” by using the formula $\log(P_t/P_{t-1})$, where P_t and P_{t-1} correspond to the closing prices of month t and t-1 respectively. (2)

vii) Plot the ACF and PACF of the monthly returns series, labelling each axis appropriately. (3)

viii) Comment on the stationarity of the monthly returns by observing the plots in part (vii). (2)

ix) Based on the ACF and PACF of the returns series obtained in part (vii), identify the most appropriate ARMA model to be fitted to the series. (1)

x) Fit the following four ARMA models to the returns series:

a) ARMA(0,0),
b) ARMA (0,1),
c) ARMA (1,0) and
d) ARMA (1,1)

and identify the model with the least Akaike Information Criteria (AIC) value. (4)

xi) Compare the result in part (ix) and part (x) and give a suitable explanation for any deviation observed. (2)

[24]

Q. 2) The dataset “CSK.csv” contains details of 38 players who represented the team Chennai Super Kings (CSK) in the Indian Premier League (IPL) in more than 15 games as of 2022.

The details of the variables are as follows:

Bat_Avg: Batting Average of the Player

Bat_SR: Batting Strike Rate of the Player

Bound_Sixes: Percentage of the total runs scored in the form of boundaries and sixes

Bowl_Avg: Bowling Average

Bowl_Econ: Bowling Economy

Bowl_SR: Bowling Strike Rate

Initial_Class: Initial classification (Initially the players are classified as either “Batsman” or “Bowler” based on the historical performance

Load the data into R and name the data frame as “CSK”.

Use the following code

```
rownames(CSK)<-CSK$Player
CSK$Player<-NULL
```

Perform feature scaling of all the numerical columns

```
CSK[,1:6]<-scale(CSK[,1:6])
```

i) Compute the mean value of all the scaled columns for Batsman and Bowler. (3)

ii) Comment on the characteristics of different features based on the mean values for Batsman and the Bowler. (2)

iii) Consider the mean values of Batsman and Bowler as the cluster centroids for the next iteration.

a) Compute the Euclidean distance between the batsman cluster centroid and each observation and store the distances in a new variable “D1” of the data frame CSK. (4)

b) Compute the Euclidean distance between the bowler cluster centroid and each observation and store the distances in a new variable “D2” of the data frame CSK. (4)

iv) Assign a new cluster category (Batsman or Bowler) to each player (create a new column called “Iteration1” to the CSK data frame) based on which of the two distances are lower. (3)

v) Compute the proportion of cluster transfers that occurred after the first iteration. (2)

vi) Repeat the steps (i), (iii) and (iv) for another iteration using the new cluster memberships of each of the players and assign the new cluster category to each player after the second iteration (create a column called “Iteration2”). (10)

vii) The players who were a part of two different categories in Iteration 1 and Iteration 2 are called as “Allrounders”. Identify the players who can be categorized as “Allrounders” in the CSK team. (2)

viii) Your friend suggested that you can use “k-means” function to do the clustering of batsman and bowler groups.

You need to execute the k-means clustering to create two clusters based on the six attributes of each player. Before executing the k-means clustering function, it is mandatory to set the seed value to 100 using set.seed(100). Print the cluster means of the two clusters formed. (3)

ix) Add the k-means cluster memberships to the CSK data frame. Rename cluster 1 as “Batsman” and cluster 2 as “Bowler”. (2)

x) Identify the names of the players who are differently classified between Iteration 2 and the k-means cluster. (2)

xi) Comment on the reason for difference in classifications between Iteration 2 and k-means cluster. (2)

[39]

Q. 3) The village "Patasnagar" is very famous for manufacturing crackers for Diwali. Most of the people in that village work with one of two manufacturing firms "Patake" and "Tapake" for their livelihood. It has been observed that the people in this village are highly exposed to respiratory disorders due to unfriendly working conditions. A research firm was able to convince the Patake's management to manufacture "green crackers" on an experimental basis from 1-Jan-2022 and funded for the expenses involved for setting up the equipment. Tapake did not listen to the request of the research firm and continued with the same type of crackers in 2022. The research firm started monitoring the villagers for 365 days from 1-Jan-22 to 31-Dec-22. The details of the observations are tabulated in "Crackers.csv".

There are five variables in the dataset

VillagerID: The ID of each villager
Green: The value is 1, if the villager works with Patake and 0 otherwise.
Male: The value is 1, if the villager is a male worker and 0 for a female worker
Status: The value is 1, if the villager is exposed to respiratory disorder in 2022 and stopped working for the year. The value is 0, if the villager did not get exposed to the disorder and survived the entire 2022 or has left the village in between and got censored.

Use the "survival" library to work with this question.

Load the data into R and name the data frame as "crackers".

i) Consider the entire data. Plot the Kaplan-Meier (KM) survival function estimate for all villagers, including 95% confidence interval of the estimate by appropriately naming the axes. (6)

ii) Using the output of part (i), compute the probability that a villager survived from getting the respiratory disorder at the end of the investigation period. (2)

iii) Consider the four groups of villagers:

a) Males producing green crackers,
b) Males producing non-green crackers,
c) Females producing green crackers,
d) Females producing non-green crackers.

Plot the KM survival estimates without any confidence intervals for these four groups in a single plot. Use different coloured lines for each plot and give a proper legend to understand the same. (8)

iv) Comment on the plots developed in part (iii). (2)

v) Instead of using the KM method, use the Cox proportional hazard model (CoxPH) with two covariates (Green and Male) without their interaction term, and respiratory disorder as the event. Any ties can be handled using Breslow method. Paste the results. (4)

vi) Comment on the effects of "Green" and "Male" (Gender) on the respiratory disorders based on the results obtained in part (v). (4)

vii) Add the Green-Gender (Male) interaction effect to the Cox proportional hazards model of part (v) and paste the results. (4)

viii) Compute the reduction/increase of disorder hazard rate for males and females by using the result of part (vii). (3)

ix) Comment on the interaction effect of Green and Male on the respiratory disorders based on the results obtained in part (vii). (2)

x) Compare and comment on the results obtained using KM and CoxPH models. (2)

[37]

****