INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

27th July 2022

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

Time allowed: 2 Hours (14.30 - 16.30 Hours)

Total Marks: 100

Q. 1) Given below is the PDF h(x) derived from the CDF H(x) of a GEV distribution

CDF:

In the case where $\gamma \neq 0$:

$$H(x) = \exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)$$

PDF:

$$h(x) = \frac{dH(x)}{dx}$$

$$= \frac{dH(x)}{dv} \times \frac{dv}{du} \times \frac{du}{dx}$$

$$= -\exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right) \times -\frac{1}{\gamma}\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)} \times \frac{\gamma}{\beta}$$

$$= \frac{1}{\beta}\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\left(1+\frac{1}{\gamma}\right)} \exp\left(-\left(1+\frac{\gamma(x-\alpha)}{\beta}\right)^{-\frac{1}{\gamma}}\right)$$

Based on the historical data, the maxima value for three years are 8, 8 and 9 units

i) Write a function in R to compute the log-likelihood of the above distribution. (3)

ii) Use the function in (i), to find the Maximum Likelihood Estimates (MLE) of the three parameters α, β and γ. You can use α = 6, β = 4 and γ = 5 as the initial estimates of the parameters. (4)

iii) Using the MLE parameters obtained in (ii) above, estimate the probability that maximum claim 'p' in any given year is greater than X where X = {10, 15, 20,….100}. (6)

iv) Plot a curve of the probabilities against the maximum claim in a year and do appropriate labelling. (4)

v) Explain the reason associated with the shape of the curve generated in (iv). (3)

[20]

Q. 2) You have been asked to analyse the number of deaths per month in the country X from lung disease over the period from Jan 2015 to Dec 2020. This information is contained in a time series called "Lung_Deaths.csv".

i) Convert the data into a time series format representing the above period after loading it into R. (2)

ii) Plot this time series giving appropriate labels for each axis. (3)

iii) Plot, on two separate graphs, the sample autocorrelation function (sample ACF) and sample partial autocorrelation function (sample PACF) of the series with appropriate labelling of the axes. (6)

iv) Justify the presence of seasonality in the data based on (ii) and (iii) above. (3)

v) Apply seasonal differencing to the time series to remove the element of seasonality. (4)

vi) Plot the ACF and PACF of this time series generated in (v) above up to lag k = 5 years, giving appropriate labels for each axis. (4)

vii) Comment on your plots in part (vi), making reference to the stationarity. (3)

[25]

Q. 3) The number of customers arriving to a grocery store can be modelled by a Poisson process with rate of 30 customers per hour.

i) Find the probability that there are 2 customers between 10:00 AM and 10:15 AM. (2)

ii) Find the probability that there are 7 customers between 11:00 AM and 11:20 AM and 15 customers between 11:20 AM and 12 noon. (4)

iii) Prepare a probability distribution table for different number of customers (0, 1, 2…20) at any given 10 minute time interval. (5)

[11]

Q .4) Refer to "data_1.csv".

i) Write down formulae for the estimated integrated hazard and its estimated variance, using the Nelson-Aalen model. (3)

ii) Load the table into R. Add two new columns to the table and populate them with estimated integrated hazard, and the estimated variance for each t_j of the given data. (3)

iii) Produce a scatterplot showing the values of the estimated integrated hazard across time t_j, for j = 0, 1,2, 3,…,20, together with the corresponding 90% confidence interval values. You need to do a proper labelling of the axes. (6)

[12]

Q. 5) **i)** Write a general function to compute the truncated moments for a lognormal distribution using k, σ, μ, L (the lower bound) and U (the upper bound) (4)

Given

$\sigma = 0.7$
$\mu = 1.2$
$L = 10$
$U = \infty$

ii) Use the function generated in (i) to compute the first and second order truncated moments. (3)

iii) Using the function in (i) or otherwise, compute the first two moments for a non-truncated normal distribution with the same μ and σ. (2)

iv) Compare the results generated in (ii) and (iii) with appropriate reasoning. (3)

[12]

Q. 6) Refer to the data set “Cricket.csv”. The data contains the following columns

A. Player – Name of the Player
B. Team – The team he represented in IPL 2022
C. Init_Group – His initial group (Batsman/ Bowler/ Alrounder)
D. Runs – Number of runs scored in IPL 2022
E. Ave_Bat – Batting Average
F. Wickets – Number of wickets taken as a bowler
G. Economy – Economy as a bowler

Load the data into R and answer the following questions. Name the data frame as “Cricket”.

i) What is the average of “Runs” and “Wickets” for three groups of players based on the initial grouping. (3)

ii) Create a copy of “Cricket” and rename it as “Cricket1”. In “Cricket1” keep only the four numerical columns and remove the rest of them. (2)

Perform a feature scaling on Cricket1 by executing the following code
Cricket1 = as.data.frame(scale (Cricket1))

After scaling, set a seed value of 100 using the following code set.seed(100).

iii) Execute k-means clustering algorithm on “Cricket1” and assign it to a variable “clust_Cricket”. Print the cluster means of the generated clusters. (4)

iv) Add the cluster memberships of each of the players to the original data frame (Cricket) by creating a new column called “Clust_Membership”. (2)

v) Rename the cluster memberships as “Batsman”, “Bowler” and “Alrounder” by following the rules given below:

- The cluster with the highest average runs is renamed as “Batsman”
- The cluster with the highest average number of wickets is renamed as “Bowler”
- The remaining cluster is renamed as “Alrounder” (3)

vi) Compute the error rate in clustering by comparing the initial grouping with the clusters created. (3)

vii) Comment on the possible reasons for misclassification in the clustering. (3)

[20]

****