INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

8th December 2022

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) As an Actuary working for company Actuaria, you have been asked to build a loss model to forecast the size of potential losses over the next year for Employers' Liability class of business. You have been provided with data for 2,164 losses in the file *"CS2BQ1.csv"*.

Library required:

MASS

Set the seed in R to 1234 prior to carrying out the analysis
In order to build the model you have been asked to carry out the following analysis:

i) Load the loss data given in the csv file in R and create a data frame with name "loss". (1)

ii) Find the mean and standard deviation of the loss data. (2)

iii) Assuming that the loss data comes from a lognormal distribution, use the method of moments to calculate the 'mu' and 'sigma' parameters of a lognormal distribution. It is mandatory to compute the parameters using the "R" code only. (3)

iv) Fit a lognormal distribution using method of Maximum Likelihood Estimation. You should use an appropriate function from the MASS package for fitting the distribution. Create 2 variables "mu1" and "sigma1" to read the parameters of the distribution that has been fit. (3)

v) Compare the parameters estimated approaches in part (iii) and part (iv) above and comment on any differences. (3)

vi) Simulate 1,000 losses from a lognormal distribution using the mu1 and sigma1. Estimate the Simulated Mean and Standard deviation. (4)

vii) Create a 'qqplot' of the simulated losses against the actual losses and plot an 'abline' between 0 and 1 with appropriate labelling. (4)

viii) Comment on the fit of the loss distribution based on the qqplot. (3)

ix) Compute the percentiles from 0% to 100% using steps of 10% of the original loss data provided and the simulated losses. (3)

x) You have been asked to estimate the losses that would be expected to be retained within a policy excess of 20,000. Create two new columns "Retained" and "Transferred" to hold all the simulated losses that will be retained and not retained respectively. (3)

xi) Using the 'quantile' function estimate the percentiles from 0% to 100% using steps of 10% of the Retained and Transferred simulated losses. (3)

xii) Combine the estimated percentiles for the Original losses, Simulated Losses, Retained and Transferred losses into a data frame with the first column showing the percentiles (i.e. 0%, 10% and so on). (3)

xiii) Comment on the percentile distributions and differences across them. (2)

xiv) Based on the estimated percentiles, you have been asked to estimate a technical price using the following formula:

$$Technical\ Premium = Mean\ of\ Transferred\ Losses + 90th\ Percentile\ of\ Transferred\ Loss\ distribution * 10\% \quad (2)$$

xv) Compare the Technical Premium to the actual quote Premium of 70,000. (2)

xvi) Plot the cumulative distribution function for the Original, Simulated, Retained and Transferred Loss distributions in one plot. (4)

[45]

Q. 2) Load the dataset "ovarian.csv" and save this as a data frame called 'ovarian'. The columns in the dataset can be defined as:

futime:	survival or censoring time
fustat:	censoring status
age:	in years
resid.ds:	residual disease present (1=no,2=yes)
rx:	treatment group
ecog.ps:	ECOG performance status

Carry out the following analysis:

i) Compute the Cox regression using the column 'rx' as the covariate. (4)

ii) Summarise the results of the regression. (2)

iii) Comment on the hazard ratio and statistical significance of the 'rx' (3)

iv) Compute the Kaplan-Meier estimators for the two groups of 'rx' and save the results to a variable called 'KM'. (3)

v) Summarise the Kaplan-Meier estimator results. (2)

vi) Plot the 'KM' results and set the colour for group 1 of 'rx' to 'red' and group 2 to 'green'. Add a plot label and x and y axes labels to your plot. (5)

vii) Comment on the KM analysis using the summary and plot results. (3)

[22]

Q. 3) Set the simulation seed to 1234 prior to carrying out the analysis.

i) Create a vector of values from 1 to 1000 and store in a variable 'x'. Create another vector 'y' by using the rule: $y = \sin(x/40)+\varepsilon$, where ε is a normal random variable with a mean of 0 and a standard deviation of 0.1. (4)

ii) Plot the x and y as a line and add grey colour to the line. Add gridlines to the plot. (2)

iii) Create a vector to capture the 'y' values with a lag of the past 40 values (including the current value). Plot the lag as a red line added to the existing plot. (3)

iv) Create a vector to capture the 'y' values with a lag of the past 20 values, the current value and 20 future values. Plot the lag as a blue line added to the existing plot. (4)

[13]

Q. 4) Consider the Dataset "fraud.csv". There are two feature variables (1) State and (2) Sum_insured indicating the geography of the insured and the size of the sum insured for 1000 different life insurance policy claims. You are asked to perform the following computations:

i) Compute the number of fraudulent claims as a proportion of total number of claims. (2)

ii) Compute the proportion of total claims associated with a medium sum insured policy from state C3, based on the data provided. (3)

iii) Given the claim is a fraudulent claim compute the proportion of claims to be associated with a medium sum insured policy from state C3, based on the data provided. (4)

iv) Using the results in (i), (ii) and (iii) compute the probability of a fraudulent claim given that a new claim comes from state C3 and has a medium sum insured. (4)

v) Compute the proportion of fraudulent claims from medium sum insured policies from C3 as a proportion of total claims from medium sum insured policies from C3. (3)

vi) Compare the results obtained in (iv) and (v) with suitable explanation. (2)

vii) Provide two major limitations of computing the probability of a fraudulent claim using this method. (2)

[20]

****