INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

29th May 2024

CS2B - Risk Modelling and Survival Analysis

Time allowed: 1 Hours 45 Minutes (14.45 -16.30 Hours)

Total Marks: 100

Instructions: Candidates are required to paste the necessary R code along with the output(s).

Q. 1) A research organisation has been asked to study possible adverse effect of a new drug administered to individuals, with particular reference to the possibility of blood clots in the brain within the first 30 days of receipt of the drug.

For this question, the survival package should be loaded into R using the following code:

install.packages("survival")
library(survival)

Load the .csv file named "*Q1data.csv*" into R and assign it to a data frame called '**datamain**'. This .csv file contains results of investigation for 2500 individuals under the heads of the following six variables:

- Life: patient identifier (integers 1, 2, ….2500)
- Drug: indicative variable (0 = received no drug, 1 = received drug)
- Age: indicative variable (0 = age less than equal to 50, 1 = age more than 50)
- Co-morbidity: indicative variable (1 = individual having another chronic disease at the time of receipt of the drug, 0 = no chronic disease)
- Status: indicative variable (0 = censoring due to the end of period, 1 = censoring due to death (for unknown reason), 2 = admission to hospital due to blood clots within 30 days of administering the drug, 3 = admission to hospital due to reasons other than blood clots within 30 days of administering the drug)
- Time: duration in days at which admission to hospital/censoring occurred (integers with range 0 to 30, 0 = day of administering the drug)

i) Comment on whether "censoring" in this investigation is likely to be informative. (3)

ii) Construct a new table 'datanew' from 'datamain' adding a new column ST with the following properties:

- ST=0 if 'Status' in the 'data' is 0 or 1 or 3
- ST=1 if 'Status' in the 'data' is 2

Display the last 20 rows of 'datanew' in your answer script. (6)

iii) Assume that the censoring is non-informative, plot Kaplan-Meier survival functions for analysing the effect of the drug on blood clots. Plot both survival functions on the same axes. Use different colours to identify each survival function and a range from 0.98 to 1 on the y-axis. Use appropriate labelling in the plot. (6)

iv) Comment on the plot of part (iii). (4)

v) It has been decided to analyse the data further using Cox's proportional hazards model and by adding covariates into the investigations.

The following decisions were taken:

- Significance of covariates would be tested with interactions.
- At least two covariates would be used.

- Two covariates to be compulsorily used are ‘drug indicator’ and ‘age’.

The decision has been taken to add ‘co-morbidity’ as one more covariate. Test the hypothesis, using the likelihood ratio statistics, that ‘co-morbidity’ has no effect on blood clots allowing for ‘drug indicator’ and ‘age’. State the null and alternative hypotheses clearly and use Breslow method for tie handling. (7)

[26]

Q. 2) A new non-banking financial company is considering which products it should first launch in the market. One of the products being considered is a personal loan and the company employs an analyst for this purpose. The analyst wants to construct a decision tree machine learning model to determine which loan applications to approve.

For answering this question use the following R package for calculating Recursive Partitioning and Regression Trees:

install.packages(“rpart”)
library(rpart)

The analyst has decided to construct the decision tree model to classify each potential customer as either expected to default (default=1) or not expected to default (default=0). For this the analyst used two features namely, a categorical feature w_1 that takes values 0, 1, and -1with equal probability and a continuous feature w_2 that takes values between -1 and 1 with a uniform distribution on [-1,1].

As there was no previous customer data to draw on, the analyst decides to construct a specimen data set to train the decision tree model with each of 20,000 hypothetical customers. For this the analyst used a field containing the value 1 if the customer defaults on the loan or 0 otherwise.

The analyst decides to model default of a hypothetical customer as:

$$\exp(w) * (1+\exp(w))^{(-1)} \qquad \text{(Eqn. 1)}$$

where,

$$w\text{=a +b*}w_1\text{+c*}w_2 \qquad \text{(Eqn. 2)}$$

a, b and c are parameters.

Before constructing the specimen dataset, the analyst first needs to generate three sets of 20,000 observations from the uniform distribution on [0,1].

i) Generate a 20,000×3 matrix named V of observations from the uniform distribution on [0,1]. Each column of V should contain 20,000 observations that have been generated together but separately from the observations in other columns. Set a random number generator seed of 1234. Display the first six rows of V in your answer script. (3)

Let V_{ij} represent the (i,j) entry of matrix V.

n = 20000

The analyst proceeds with creating the specimen data set as follows:

- An *n*-component vector w_1 whose *i*'th component is equal to:

$$\begin{cases} -1 \text{ if } V_{i1} \leq ⅓ \\ 0 \text{ if } ⅓ < V_{i1} \leq ⅔ \\ +1 \text{ if } V_{i1} > ⅔ \end{cases}$$

- An *n*-component vector w_2 whose *i*'th component is equal to:

$$2^*V_{i2} – 1$$

- An *n*-component vector *w* whose *i*'th component is the value of *w* given by the formula above [Eqn. (2)] for customer *i.*
- An *n*-component vector *defualtprob* whose *i*'th component is the probability of default given by the formula above [Eqn. (1)] for customer *i.*
- An *n*-component vector *default* whose *i*'th component is equal to 1 if V_{i3} is less than or equal to the probability from the vector *defaultprob* for the customer i and 0 otherwise.
- A dataframe in the format required to construct decision trees using the R package *rpart* , incorporating the vectors w_1, w_2 and *default*.

The R code to create the data frame in the format required, given the vectors w_1, w_2 and *default,* with the names of columns being “w1”, “w2” and “default” respectively, is as follows:

data.frame(“w_1”= w_1, “w_2”= w_2, “*default*”= default)

ii) Generate a data frame named *sample*, in line with the analyst's construction rules, corresponding to the parameter values *n*=20,000, a=0, b=c=0.5 and display in your answer script the first six rows. (10)

iii) Calculate the expected probability of customer default based on the *sample* data generated. Comment on how realistic the result is. Also suggest, giving reasons, a value of the parameter *a* that can give a realistic probability of default. (4)

iv) Use parameter *a = -0.6* and keep other parameter values unaltered; Calculate the expected probability of customer default based on the *sample* data thus generated. Comment on the result. (4)

v) Now use a new *defaultprob* function $0.5^*exp(w) * (1+0.5^*exp(w))^{(-1)}$, where *w*=a +b*$w_1$+c*$w_2$ and a, b and c are parameters as given in (i). Calculate the expected probability of customer default based on the *sample* data thus generated. (4)

The R code for fitting and plotting a decision tree object called *tree*, that predicts whether a customer will default or not, is as follows:

tree = rpart(default ~ ., data = sample, method = "class") (Eqn. 3)

The analyst collects actual customer data from another financial institution to assess the predictive power of the decision tree model. “*Q2data.csv*” contains the actual data for 125 customers in the required format for the decision tree model. Load this file into R and assign it to a data frame called *actual.*

vi) Use R code to generate first 10 rows of the *actual* data frame and paste it in your answer script. (2)

vii) Generate the predicted default classifications of the 125 customers in the *actual* data frame using the R code above, assigning the prediction to a vector called *predict_defaults;* display its first 20 values. When using Eqn. 3, use the 'sample' from (ii), i.e. using *n*=20,000, a=0, b=c=0.5. (4)

viii) Generate and display in your answer script a confusion matrix of actual defaults versus predicted defaults for the 125 actual customers. (2)

ix) Using R code determine the values of the precision and recall percentages for the decision tree model's predictions for the 125 actual customers, where the true is defined as the case where the model predicts a default for a customer that has actually defaulted. (4)

Based on the value of the recall percentage for the 125 actual customers, the Finance Director of the company concludes that the decision tree model needs no further refinement and can be used to determine which loan applications to approve when the product is launched.

x) Comment on the Finance Director's conclusion. (6)

[43]

Q. 3) It has been decided to analyse percentage change in quarterly personal consumption expenditure and personal disposable income from 1970 to 2010. This information is given in a time series *usconsumption* in R's '*fpp'* library.

Install packages '*fpp'* and rename *usconsumption[,1]* as *consumption.*

i) Plot this time series giving appropriate labels for each axis; paste the R code and the chart into your answer script. (3)

ii) Plot the ACF and PACF of this time series giving appropriate labels to each axis; paste the R code and charts into your answer script. (4)

iii) Fit ARIMA (3,0,0) model based on your answer in part (ii) stating the equation of the model, while also justifying your approach. (4)

The dataset *usconsumption* also includes quarterly personal disposable income from 1970 to 2010. Use following R code to load income data:

income <- usconsumption[,2]

iv) Compare the performance of the ARIMA model you have chosen in part (iii) with a linear regression model of consumption on income, by comparing the root mean square error (RMSE) for the fitted values of each model. Paste your R code and output into your answer script. (8)

An analyst has suggested that neither model in part (iv) is a good fit to the data and has asked you to propose an alternate model.

v)

a) Suggest a suitable alternative model to fit the data.
b) Fit the model you suggested in part (v) (a) to the data stating the equation used.
c) Compare the results of this model to the models fitted in part (iv). (12)
[31]

****