INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

18 April 2024 (am)

Subject CS2 – Risk Modelling and Survival Analysis Core Principles

Paper B

Time allowed: One hour and fifty minutes

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator.

If you encounter any issues during the examination please contact the Assessment Team on T. 0044 (0) 1865 268 873.

1 An insurer has written ten insurance policies under which, at most, one claim can be made from each policy. The probability of a claim from policy i is as follows:

$$P_i = 0.1 + 0.02i \quad \text{for } i = 1, 2, \ldots, 10$$

Claim amounts for all policies follow the same Gamma distribution with shape parameter $\alpha = 10$ and rate parameter $\lambda = 2$.

(i) Construct R code to generate and display a single vector setting out the probability of a claim for each policy using the `seq` function. [2]

(ii) Construct R code to calculate and display the theoretical mean and theoretical standard deviation of the total claim amounts. [5]

(iii) Construct R code to generate 1,000,000 simulations of claim amounts under each of the ten policies, using a random number generator seed of 500 and display the first five sets of simulations. [6]

(iv) Construct R code to generate and display a vector showing the mean and standard deviation of claim amounts for each policy from the simulations in part (iii). [3]

(v) Calculate the mean and standard deviation of the total claim amounts from the ten policies in this portfolio from the simulations in part (iii) and compare these with your answers in part (ii). [4]

(vi) Estimate, from the simulations in part (iii), the probability that there will be no claims incurred from all ten policies. [2]

(vii) Calculate the 99th percentile of total claim amounts from the simulations in part (iii). [2]

(viii) Plot a histogram of the simulated total claim amounts, showing density on the y-axis. [2]

(ix) Plot a density function of the Normal distribution using the parameter values from part (ii) on top of the histogram in part (viii). [2]

(x) Calculate the 99th percentile of total claim amounts using the Normal distribution in part (ix) to approximate the total claims distribution. [2]

(xi) Comment on the appropriateness of using the Normal distribution in this case. [2]

(xii) Plot a graph similar to the one in part (ix), but with the probability of a claim from policy i as follows:

$$P_i = 0.4 + 0.02i \quad \text{for } i = 1, 2, \ldots, 10$$

[3]

(xiii) Comment on the appropriateness of using the Normal distribution in this new case. [3]

[Total 38]

2 The dataset 'CS2B_A24_Q2.csv' contains four variables: Body Fat Index (BFI), age, weight (in pounds) and height (in inches).

An analyst is considering fitting the following linear regression model that predicts the BFI:

$$\text{BFI}_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Weight}_i + \beta_3 \text{Height}_i + \varepsilon_i$$

The parameters can be fitted based on the following penalty function:

$$L(\boldsymbol{B}, \alpha, \lambda) = \sum_{i=1}^{n} (\text{BFI}_i - \beta_0 - \beta_1 \text{Age}_i - \beta_2 \text{Weight}_i - \beta_3 \text{Height}_i)^2 + \lambda \left(\frac{1-\alpha}{2} (\beta_1^2 + \beta_2^2 + \beta_3^2) + \alpha(|\beta_1| + |\beta_2| + |\beta_3|) \right)$$

where the vector of regression parameters, $\boldsymbol{B} = (\beta_0, \beta_1, \beta_2, \beta_3)$ and α and λ are some parameters that are to be defined.

(i) State the type of the regression model if:

(a) $\alpha = 1$.

(b) $\alpha = 0$.

[2]

(ii) Construct R code to generate a dataframe named 'BFI' that includes the contents of the file 'CS2B_A24_Q2.csv'. [2]

Run the following two lines of code:

```
X=as.matrix(BFI[,-1])
Y=BFI[,1]
```

(iii) Comment briefly on each of these two lines of code. [2]

(iv) Construct R code to generate a function called 'Penalty' that calculates the penalty function $L(\boldsymbol{B}, \alpha, \lambda)$ as above on the given data set, and with three input variables:

$\boldsymbol{B} = (\beta_0, \beta_1, \beta_2, \beta_3)$ the vector of beta values
'alpha' as the parameter α
'lambda' as the value of λ.

[10]

(v) Derive the value of the 'Penalty' function above for values of $\boldsymbol{B} = \left(1, \frac{1}{2}, \frac{1}{3}, 1\right)$, alpha = 0.5 and lambda = 0.8. [2]

(vi) Determine the values of alpha and lambda for which the 'Penalty' function is minimised when $\boldsymbol{B} = \left(1, \frac{1}{2}, \frac{1}{3}, 1\right)$. You do not need to perform any additional calculations. [4]

(vii) Calculate the corresponding minimum value of the ‘Penalty’ function when $\boldsymbol{B} = \left(1, \frac{1}{2}, \frac{1}{3}, 1\right)$. [2]

(viii) Determine the values of parameters α, λ and $\boldsymbol{B}$ for which the function $L(\boldsymbol{B}, \alpha, \lambda)$ reaches its minimum based on the reasoning used in part (vi). [4]

(ix) Comment on the validity of using the values of parameters α, λ and $\boldsymbol{B}$ from part (viii). [5]

[Total 33]

3 A study has been made into how long in weeks it takes for long distance athletes who suffer ankle injuries to recover and be able to run again.

The datafile 'CS2B_A24_Q3.csv' contains data on a group of such athletes where the information recorded for each one is:

- time = the number of weeks before recovery or censoring occurs.
- leave = 0 if the person leaves the study before full recovery, or 1 if time to full recovery is recorded.
- route = 0 if the person was treated with pain relief drugs and told to rest, or 1 if they were on a program of physiotherapy instead of the drugs.
- prior = 1 if there had been another ankle injury in the last 36 months, or 0 otherwise.

Before you start this question, you will need to install the survival package in R:

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

(i) Construct R code to generate a dataframe named 'study' that includes the contents of the file 'CS2B_A24_Q3.csv' and display the first twelve rows. [2]

(ii) State an example of how censoring in this study may be informative or non-informative for each treatment route. [2]

(iii) Generate R code to calculate the Kaplan–Meier estimate of the survival function for the length of time to recovery separately for athletes who take pain relief drugs and those on a physiotherapy programme, and plot the resulting two survival functions using different coloured lines in a single plot. [8]

(iv) Explain what conclusions can be drawn from your results in part (iii). [3]

(v) Construct R code to generate a Cox's proportional hazard model with two covariates, 'route' and 'prior'. You should use the Breslow method for tie handling and display the resulting regression coefficients. [6]

(vi) Comment on what conclusions can be drawn from your answer to part (v) about whether treatment route or prior injury is more important in predicting length of time to recovery. [4]

(vii) State how you could test which of the two covariates is more important. [4]

[Total 29]

END OF PAPER