INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

19 April 2023 (am)

Subject CS2B – Risk Modelling and Survival Analysis Core Principles

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 Before answering this question, the ‘`markovchain`’ R package should be loaded into R using the following code:

```
install.packages("markovchain")

library(markovchain)
```

A three-state Markov chain model consisting of healthy (‘H’), sick (‘S’), and dead (‘D’) states has the following weekly transition probabilities:

$p_{HH} = 0.98$
$p_{HS} = 0.019$
$p_{HD} = 0.001$
$p_{SS} = 0.75$
$p_{SH} = 0.2$
$p_{SD} = 0.05$

(i) Construct a ‘markovchain’ object for the transition matrix for the above Markov chain model, calling it ‘`transitions_markov`’. [2]

(ii) Calculate the probability that, given a life is healthy now, the life will be sick at time 2 weeks. [2]

(iii) Calculate the probability that, given a life is healthy now, the life will be sick at some point in the next 52 weeks. [3]

(iv) Calculate the probability that, given a life is healthy now, the life will remain healthy for the entire year. [1]

A new, highly contagious disease has become prevalent in the country and it is proposed that the HSD model above be amended to model the effects of this disease. Lives become ill for a period of time and then either recover or die. Once recovered, a life is deemed immune and cannot become ill again.

It has been decided to use a Markov jump model instead of a chain model. The Markov model consists of four states: healthy (‘H’), sick (‘S’), recovered (‘R’) and a dead (‘D’) state. The following daily transition rates have been estimated from a recent investigation:

$\mu_{SD} = 4\%$
$\mu_{SR} = 13\%$

The rate from healthy to infected, μ_{HS}, is equal to b multiplied by i_t where b is a constant and i_t is the proportion of lives in the sick state at time t. All other transition rates are zero.

(v) Explain why it may be preferable to set $\mu_{HS} = b\ i_t$ (as above), rather than using a constant value (as has been done for the other rates). [1]

(vi) Give an example of a scenario that would be expected to result in a particularly high value of b. [2]

The proposed value of b is 0.3, and it is assumed that at time 0, 1% of the population are in the sick state and the remaining 99% are in the healthy state.

(vii) Using this revised model and a step length of 0.01 days, calculate the occupancy probabilities in each of the four states from $t = 0.01$ to $t = 100$ days inclusive. You should output your answers for each value of t to successive rows of a matrix called 'mat_ans'. [11]

(viii) Plot a graph showing the probabilities in part (vii) with suitably labelled axes and making clear which plot components correspond to which state. [5]

(ix) Calculate the probability that a life healthy at time 0 is sick after:

(a) 5 days.

(b) 30 days.

[3]

(x) Calculate the expected present value of a daily rate of £1 payable while sick using a force of interest of 6% p.a. [4]

[Total 34]

2 A life insurance company has issued a 1-year group insurance policy covering 10,000 employees categorized into four age bands of 2,500 employees each.

The following amounts are paid upon each death under the policy:

- Death from pandemic diseases mentioned in the policy document: £80,000.
- Death from cause other than above: £40,000.

The premium charged for each group was estimated as the mean of a compound binomial distribution.

The 'CS2B_A23_Qu_2_Data.csv' file contains the following data for each group:

group:	name of the group
age_band:	age range in the group
Mort:	expected probability of death
p_dth:	proportion of deaths due to pandemic diseases

Before answering this question, the 'CS2B_A23_Qu_2_Data.csv' file should be loaded into R and assigned to a data frame called 'Prem'.

(i) Calculate the premium that company would have charged under each of the groups, ignoring expenses and profit margin, assigning your output to a vector called 'Prem_charged' and displaying 'Prem_charged' in your answer script. [6]

The head of the department wishes to project the actual claims over the next year using random numbers.

(ii) Generate a sample of size 2,500 under each of the above groups from a Bernoulli distribution, simulating whether death occurs for each of the employees using the expected probability of death from the table 'Prem'. You should output your results to a data frame called 'Tab_R' consisting of four column vectors 'G1_R'to 'G4_R'. The random number generator seed to be used is 123.

Use the `summary()` function to display your results. [6]

(iii) Generate a data frame 'Tab_U' of observations from the Uniform(0,1) distribution, consisting of four column vectors, 'G1_U' to 'G4_U', with 2,500 components each. The random number generator seed to be used is 300. Use the R function `head()`, to display the first six rows of 'Tab_U'. [4]

(iv) Generate a data frame ‘Tab_V’, consisting of four column vectors ‘G1_V’ to ‘G4_V’, with 2,500 components each, representing claim amounts under each group.

If death has occurred, the relevant entry of ‘Tab_V’ should be set to 80,000 if the random variable generated in part (iii) is less than or equal to the proportion of pandemic deaths specified in the table ‘Prem’ for the relevant group, or 40,000 otherwise.

Use the `summary()` function to display your results. [8]

(v) Calculate the aggregate claim amount under each group, displaying your output as a vector. [3]

(vi) Calculate the projected profit or loss under each group and the policy as a whole. [2]

One of the employees suggests that the premium rate charged should be based on the actual claims obtained from part (v) above.

(vii) Comment on the validity of the employee’s suggestion, assuming that the distribution used, the parameters used and the methodology adopted are reasonable. [4]

[Total 33]

3 Before answering this question, the R packages for calculating and plotting Recursive Partitioning and Regression Trees should be loaded into R using the following code:

```
install.packages("rpart")

install.packages("rpart.plot")

library(rpart)

library(rpart.plot)
```

An insurance company is considering introducing a simplified underwriting process for a health insurance product. The company is considering using monthly benefit amount and another feature, 'Feature1', as criteria for determining which customers are eligible for the simplified process. There is a further feature, 'Feature2', which the company is prohibited by local regulations from using as a criterion.

In order to illustrate the issues involved in setting the eligibility criteria for the simplified process, an Actuary engaged by the company generates a specimen data set by making the following assumptions about the business mix:

- For a given customer, Feature1 takes the values 0 and 1 with equal probability.
- For a given customer, independently of Feature1, Feature2 takes the values 0 and 1 with equal probability.
- For a given customer, independently of Feature1, the benefit amount, 'Benefit', is such that its logarithm is Normally distributed with mean 7.5 and standard deviation 0.5 if Feature2 = 0, and Normally distributed with mean 7 and standard deviation 0.5 if Feature2 = 1.
- The variable 'Outcome' indicates whether full medical underwriting would reveal issues requiring the proposal to be declined or accepted with a premium loading. For a given customer, this variable takes the value 0, representing a proposal that would be accepted at standard premium rates, with probability:
    - o 0.95 if Feature1 and Feature2 are both 0.
    - o 0.8 if Feature1 and Feature2 are both 1.
    - o 0.9 if one of Feature1 and Feature2 is 0 and the other is 1.

  Otherwise, this variable takes the value 1, representing a proposal that would be declined or accepted with a premium loading.

(i) Generate a 100,000 × 5 matrix, A, in which:

- the first, second and fifth columns consist of independent observations from the uniform distribution on [0,1]. You should set a random number generator seed of 123 before generating the observations in the first column.
- the third column consists of independent observations from the lognormal distribution of Benefit given that Feature2 = 0.
- the fourth column consists of independent observations from the lognormal distribution of Benefit given that Feature2 = 1.

Use the R function `head` to display the first six rows of A in your answer script. [6]

Let A_{ij} denote the (i,j) entry of A.

(ii) Generate a matrix, B, with 100,000 rows and four columns, Feature1, Feature2, Benefit and Outcome, in which, in the ith row:

- Feature1, Feature2 and Outcome are equal to 0 if A_{i1}, A_{i2} or A_{i5}, respectively are less than the appropriate probability, and 1 otherwise.
- Benefit is equal to either A_{i3} or A_{i4} as appropriate.

Use the R function `head` to display the first six rows of data in your answer script. [9]

The Actuary uses a regression tree to split the hypothetical customers in the matrix B by Feature1 and Benefit, with the view that those customers in nodes with low probabilities of an adverse outcome under full medical underwriting, would be eligible for the simplified process. The Actuary uses the following R code to plot the tree:

```
tree = rpart(formula = Outcome ~ Feature1 + Benefit,

data = data.frame(B), control = rpart.control(cp = 2e-4,

maxdepth = 2, minbucket = 5000))

rpart.plot(tree)
```

(iii) Plot the regression tree object, called 'tree', using the R code above. [1]

(iv) Comment on the conclusions the Actuary should draw from the regression tree 'tree'. [5]

The Actuary creates a further regression tree by weighting each customer, i, in the matrix B as follows:

$$\frac{\frac{1}{2}(w_{0i}+w_{1i})}{w_i}$$

where:

- w_{0i} is the probability density from the lognormal distribution for the actual value of Benefit for the ith customer and Feature2 = 0.
- w_{1i} is the probability density from the lognormal distribution for the actual value of Benefit for the ith customer and Feature2 = 1.
- w_i is the probability density from the lognormal distribution for the actual values of Benefit and Feature2 for the ith customer.

(v) Explain the rationale for this choice of weights. [2]

(vi) Generate a vector, ‘Weight’, containing the weights as defined above. Use the R function `head` to display the first six entries of Weight in your answer script. [5]

The Actuary uses the following R code to plot the revised tree, ‘tree2’:

```
tree2 = rpart(formula = Outcome ~ Feature1 + Benefit,

data = data.frame(B), weights = Weight,

control = rpart.control(cp = 2e-4, maxdepth = 2,

minbucket = 5000))

rpart.plot(tree2)
```

(vii) Plot the revised regression tree object, called ‘tree2’, using the R code above. [1]

(viii) Comment on the practical suitability of this method of determining the eligibility criteria for the simplified underwriting process for this product. [4]

[Total 33]

END OF PAPER