INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

20 September 2023 (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 The data files 'CS2B_S23_Q1_Deaths.csv' and 'CS2B_S23_Q1_Exposures.csv' contain the deaths and central exposed to risk, respectively, for English and Welsh males for each age last birthday from 20 to 100 inclusive, in each year from 1961 to 2020 inclusive. The files contain both row and column headings.

(i) Construct R code to load the data files into R and to assign them to two 81 × 60 matrices called 'Deaths' and 'Exposures' respectively. [4]

(ii) Generate an 81 × 60 matrix called 'm_xt', consisting of the crude central mortality rates calculated by dividing the deaths by the exposures. [2]

Gompertz Law is fitted to the crude central mortality rates, m_{xt}, for each calendar year t separately using linear regression of the log central mortality rates on age x such that:

$$m_{xt} = \exp\left(\alpha_t + \beta_t(x - \bar{x})\right)$$

(iii) Generate a 60 × 2 matrix called 'Gompertz', consisting of the values of α_t and β_t for each calendar year t. Display the first six rows of Gompertz in your answer script. [8]

(iv) Plot two line graphs showing the values of α_t and β_t you fitted in part (iii) for each calendar year t, displaying the graphs side-by-side in your answer script. [7]

(v) Comment on the key features of your graphs in part (iv). [3]

(vi) Plot a graph showing the value of the chi-square goodness of fit statistic for the Gompertz graduation for each calendar year t. [8]

(vii) Suggest with reasons, what further investigations should be carried out in view of your graph in part (vi). [3]

[Total 35]

2 The data file 'CS2B_S23_Q2_Data.csv' contains historical data of two time series over the past 1,000 months.

(i) Extract the data into a table named 'data' assigning the second and third column to X and Y, respectively. Display the first five entries of X. [3]

(ii) Plot X and Y in a single chart suitably naming the variables. [5]

(iii) Plot the sample autocorrelation and sample partial autocorrelation functions, for both X and Y. [3]

(iv) Comment on the stationarity of both X and Y with reference to the plots in (iii) above. [1]

(v) Determine the number of times X and Y have to be differenced (parameter d) in order to convert them into a stationary series by calculating the sample variance of the differenced process at $d = 0$, 1, 2 and 3. [14]

An analyst infers that X and Y are cointegrated. They define:

$$Z_t = Y - bX$$

where $[1, -b]$ is called the cointegrating vector.

The relationship between X and Y is then described by a linear regression:

$$Y = a + bX + e_t$$

where e_t is the residual.

(vi) Determine the values of a and b in the equation for Y above. [3]

(vii) Comment on the analyst's inference, without using R, if we assume e_t is a white noise. [4]

(viii) Comment on the analyst's inference, without using R, if we assume e_t is not a white noise. [3]

[Total 36]

3 An insurance company is looking to launch an investigation into the mortality patterns of a group of pensioners aged from 55 to 90. The data is contained in the file called 'CS2B_S23_Q3_Pensioners.csv'.

(i) Construct R code to load the dataset into R, assign it to a dataframe called 'dataset1', and display the last eight rows. [2]

Mortality rates are calculated as the ratio of 'DeathCount' by 'PopulationSize'.

(ii) Construct R code to add a new column called 'logRate' containing the logarithm of the mortality rates, to the dataframe dataset1. [1]

The company would like to use this data to estimate mortality rates beyond age 90. They decide to start with a simple model of the form:

$$lq_x = \alpha + \beta x + \varepsilon_x$$

where lq_x is the logarithm of mortality rate at age x, α and β are the intercept and slope parameters, respectively and ε_x indicates the error term.

The Residual Sum of Squares (RSS) is given by:

$$\text{RSS} = \sum_{x=55}^{90} (\text{logRate} - \alpha - \beta x)^2$$

(iii) Construct R code to determine, to five decimal places, the estimate of the intercept and slope parameters that minimises the RSS. [3]

(iv) Plot the fitted model together with the logarithm of the observed mortality rates as a function of age on the same graph. [4]

(v) Comment on the plot in (iv) above. [1]

The company would like to look at ways to improve the model and decides to investigate the following family of five models with polynomial terms in age x:

$$lq_x = \sum_{k=0}^{p} \alpha_k x^k + \varepsilon_x$$

for p = 1,2,3,4,5 and where α_k are parameters to be estimated.

(vi) Construct R code to fit these five models to the data and store the corresponding RSS into a dataframe called 'output_rss' such that the first column of output_rss contains the values of p and the second column contains the values of the RSS. [8]

(vii) State which is the best model (among the five models) according to RSS. [1]

(viii) Construct R code that uses the best model to forecast the logarithm of the mortality rates from age 91 to 110 and print your forecast values. [7]

(ix) Comment on this forecast and the appropriateness of using RSS for model selection. [2]

[Total 29]

END OF PAPER