INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

20 September 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 insurance company launched a new product over a year ago and the dataset 'S24_Q1_sales_data.csv' shows the number of products sold each day. The company is looking to use this data to forecast sales volumes for future months.

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

(ii) Construct R code to check if the values in the column labelled 'day' are in ascending order and state your conclusion. [3]

(iii) Construct R code to plot a suitably labelled line graph of the number of products sold over time. [3]

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

Let 'products_sold_k' denote the vector obtained by differencing the sales volume k times.

(v) Assess the stationarity of 'products_sold_k' for $k = 0, 1, 2$. [7]

The company would like to investigate models in the form ARIMA(p, 2, q) for the 'products_sold' data where p and q can take the values 0, 1, 2 or 3.

(vi) Construct R code to fit ARIMA(p, 2, q) models to the data for all possible combinations of p and q and store the values of the Akaike Information Criterion (AIC) in a 4×4 matrix named 'aic_matrix' with appropriate row and column names. [8]

(vii) Identify the most appropriate model for fitting to the data according to AIC. [1]

(viii) Construct R code that uses the model identified in part (vii) to forecast sales volume over the next 60 days, store the forecast values into a vector called 'forecast_volume', and display the first ten values of this forecast. [4]

(ix) Plot a graph showing the forecast sales from part (viii). [4]

(x) Comment on the plot in part (ix). [2]

[Total 35]

2 An insurer is analysing the price of their pet insurance policies. In this instance the analyst is assessing mortality rates of a particular breed of cat.

Mortality data in respect of the breed of cat has been collected from a recent 2-year investigation. The raw data in respect of each cat is found in the file 'S24_Q2_MortData1.csv'. This dataset includes the exact age of the cat when observation commenced and ended, together with the following indicator variable on whether the cat died during the observation or not:

- where the cat died: Died = 1
- where the cat survived: Died = 0.

The aggregate Exposed To Risk (ETR) (in years) and number of deaths, split by age last (in years), calculated from 'S24_Q2_MortData1.csv' are shown below:

Age last	*ETR*	*Deaths*
0.000	5,737.507	1,256.000
1.000	15,630.348	3,996.000
2.000	19,255.943	5,770.000
3.000	18,669.125	6,455.000
4.000	17,763.585	7,406.000
5.000	16,698.656	8,052.000
6.000	15,586.883	8,878.000
7.000	14,364.160	9,653.000
8.000	8,334.790	6,464.000
9.000	1,713.752	1,531.000

(i) Construct R code to load the dataset from 'S24_Q2_MortData1.csv' saving it as a dataframe. [1]

(ii) Construct R code to verify the ETR values above in respect of each integer age last from 0 years up to 9 years. Save the ETR values in a vector called 'etr1'. [7]

(iii) Construct R code to verify the above total number of deaths in respect of each integer age last from 0 years up to 9 years. Save the death values in a vector called 'deaths_summr'. [4]

(iv) Calculate the crude rates for each age. [2]

The files 'S24_Q2_Table1.csv' and 'S24_Q2_Table2.csv' are two established mortality tables that have been calculated using data from similar breeds of cat to that being analysed. Before completing the remainder of this question you will need to load 'S24_Q2_Table1.csv' and 'S24_Q2_Table2.csv' into R and save them as two datasets 'Table1' and 'Table2'.

(v) Plot the crude rates from part (iii) above and the rates from 'Table1' and 'Table2' together in one suitably labelled plot. You should show the crude rates as points and the two sets of table rates as different coloured lines. [4]

The insurance company first wishes to test the null hypothesis that the rates in 'Table1' represent the true mortality of the cats in this analysis.

(vi) Construct R code to complete the following three statistical tests of this null hypothesis and in each case state the conclusion of the test:

(a) Chi-squared test

(b) serial correlations test

(c) signs test.

[12]

(vii) Construct R code to repeat the three tests in part (v) for the null hypothesis that the rates in 'Table2' represent the true mortality of the cats in this analysis, again stating the conclusion of each test. [6]

[Total 36]

3 Alpha Advisors are a small firm of investment consultants advising on the selection of investment funds. An analyst at Alpha Advisors is conducting research on the investment return after fees of 50 funds that invest in emerging markets. The data collected on these funds is found in the file 'S24_Q3_Data.csv' that contains nine columns:

fund	An ID number for each fund from 1 to 50.
return	The % return after fees over the last year.
rating	Alpha Advisors' rating of the fund (a metric used for making fund selection recommendations) on a scale of 1 to 5 at the start of the year.
risk	A measure on a scale of 1 to 10 of the amount of risk taken in the fund.
size	The total value of assets in the fund in millions of US dollars.
experience	The number of years of industry experience of the lead fund manager.
inflow	The net of monetary flows in and out of the fund from investors in the last year.
holdings	The number of stocks owned by the fund.
turnover	The % turnover (stock purchases and sales divided by size) in the fund.

The analyst seeks to understand the main factors affecting investment returns in the last year.

(i) Construct R code to load the dataset from 'S24_Q3_Data.csv' into R saving it as a dataframe called 'fund_data' and display the first nine rows. [2]

(ii) Construct R code to build a multiple linear regression model for return as the response variable and rating, risk, size, experience, inflow, holdings and turnover as explanatory variables and display the results. [3]

Alpha Advisors claim in their own marketing materials to have skill in selecting better performing funds using their rating.

(iii) Comment briefly on this claim based on your output in part (ii). [3]

(iv) Explain why it can be useful to work with rescaled explanatory variables with zero mean and standard deviation of one. [1]

(v) Calculate standardised versions of each of the seven explanatory variables above and then build a multiple linear regression model for return with these standardised explanatory variables. [4]

(vi) Construct R code to calculate the sum of squares of the regression coefficients for the model in part (v) and store this value as 'penalty'. [2]

One way of applying ridge regression techniques to this scenario is to seek a linear regression model that minimises a penalised sum of squares given by:

$$RSS - \lambda \sum_{all\ i} \beta_i{}^2$$

where *RSS* is the residual sum of squares of the regression model, the β_i are the regression coefficients for that model and $\lambda \geq 0$ is some constant.

(vii) Determine the values of λ for which a simple linear regression model with only the rescaled Alpha Advisors ratings as the explanatory variable has lower penalised sum of squares than the full multiple linear regression model in part (v) by calculating the penalised sum of squares for these two models for λ = 0, 1, 2, … 10. [12]

The head of marketing at Alpha Advisors wants to run advertising with the headline 'All you need are our Alpha ratings' claiming that the rating alone can be used to determine better performing funds.

(viii) Comment on this marketing claim with reference to your results in part (vii). [2]

[Total 29]

END OF PAPER