INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

19 April 2021 (am)

Subject CS2 – Risk Modelling and Survival Analysis Core Principles

Paper A

Time allowed: Three hours and fifteen 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.

 © Institute and Faculty of Actuaries

1 The Frank copula, C_F, for a bivariate distribution is defined as:

$$C_F(u, v) = -\frac{1}{\alpha} ln\left(1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{(e^{-\alpha} - 1)}\right), \alpha > 0$$

(i) Determine the probability that two jointly distributed random variables, X and Y, are both less than or equal to their median values where X and Y follow the Frank copula, C_F, with $\alpha = 1$. [2]

(ii) Determine the revised value of the probability in part (i) when $\alpha = 0.1$. [1]

(iii) Determine the probability that two jointly distributed random variables, X and Y, are both less than or equal to their median values where X and Y follow the product copula. [1]

(iv) Comment on your answers to parts (i), (ii) and (iii) with reference to the sign and level of dependence exhibited by the Frank copula. [2]

[Total 6]

2 A second-order moving average process is defined by the following equation:

$$X_t = \mu + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2}$$

where e_t is a white noise process with variance σ^2.

(i) Determine $E(X_t)$ and $\mathrm{Var}(X_t)$. [3]

(ii) Determine the autocovariance function, $\{\gamma_k\}$, of X_t for $k \geq 0$. [4]

[Total 7]

3 Consider the following stochastic process:

$$Y_t = X_1 + X_2 + \cdots + X_t$$

where $t = 1, 2, 3, \ldots$ and X_i are random variables that are independently and identically distributed as $N(0, \sigma^2)$.

(i) Explain whether Y_t is a Markov process. [2]

(ii) Derive an expression, in terms of t, for the value of the correlation coefficient $\mathrm{Corr}(Y_t, Y_{t+20})$, simplifying your answer as far as possible. [5]

(iii) Comment on how the value of the correlation coefficient derived in part (ii) behaves as t increases. [2]

[Total 9]

4 A duration-dependent Markov jump process model is constructed to model the recovery and mortality rates of individuals who are sick with a particular medical condition. The model includes the following three states:

- sick
- recovered
- dead.

Under this model, an individual who has been sick for duration t experiences:

- a transition rate to the recovered state of $a\exp(-bt)$
- a transition rate to the dead state of $a(1-\exp(-bt)) + c$

where a, b and c are parameters > 0 that depend on the characteristics of the individual.

(i) Comment on the reasonableness of this model. [2]

(ii) Demonstrate that the probability that an individual who has just become sick eventually recovers is:

$$\frac{a}{(a+b+c)}.$$

[7]
[Total 9]

5 A life insurance company has offices in Towns A and B. The company writes 25-year term assurance policies. Below are data from the two offices relating to policyholders of the same age, x. Policies in force and deaths are on an 'age last birthday' basis.

	Town A	*Town B*
Policies in force on 1 January 2020	3,000	1,770
Policies in force on 1 January 2021	3,300	1,674
Deaths in calendar year 2020	63	26

(i) Estimate the force of mortality for the calendar year 2020 in respect of each of the offices in Towns A and B. [2]

A detailed examination of the records shows that 50% of the policies in force in Town A at both dates were in respect of smokers, and 20% of policies in force in Town B at both dates were in respect of smokers.

The national forces of mortality at age x for smokers in 2020 were 50% higher than those for non-smokers.

(ii) Estimate the force of mortality for smokers and for non-smokers in each of the Towns A and B, clearly stating any assumptions that you make. [6]

The life insurance company charges policyholders in Towns A and B the same premiums. It charges smokers in both towns 50% more than non-smokers.

(iii) Comment on the company's pricing structure in light of your results from parts (i) and (ii). [3]

[Total 11]

6 Claims on a particular type of insurance policy follow a compound Poisson process with an annual claim rate per policy of 0.4. Individual claim amounts are Exponentially distributed with mean 120. In addition, for a given claim and independent of its size, there is a probability of 20% that an extra claim handling expense of 30 is incurred. The insurer charges an annual premium of 60 per policy.

Estimate, using a Normal approximation, the minimum number of policies to be sold so that the insurer has at least a 99% probability of making a profit. [11]

7 An Actuarial Analyst is investigating the forces of mortality for males aged 65 and over. The Analyst has studied a group of 100 male lives, all of whom were exactly 65 years old at the beginning of the study, over a 10-year period and has estimated the following forces of mortality based on the lives observed in the study:

Duration from age 65 (years)	*Force of mortality (p.a.)*
0–1	0.040
1–8	0.005
8+	0.080

(i) Determine, to six decimal places, the probability that a 65-year-old male is alive at age 75, using the estimated forces of mortality in the table above. [3]

A colleague has suggested that the Analyst use the following formula to model the forces of mortality at age x:

$$\mu_x = 0.0020291 + 0.0001000 \times 1.0793496^x$$

(ii) Verify that the probability that a 65-year-old male is alive at age 75 using the suggested model matches the result calculated in part (i) to six decimal places. [4]

The Analyst decides, on the basis of the matching probabilities in parts (i) and (ii), to use the suggested formula to graduate the estimated forces of mortality in the table above.

(iii) Comment on the Analyst's decision. [4]
[Total 11]

8 An insurance company offering private medical insurance wants to build a machine learning model to predict the expected medical costs of policyholders based on their age and Body Mass Index (BMI). The data set that will be used to train the model consists of the actual medical costs, together with the age and BMI, of nine past policyholders and is set out in the table below:

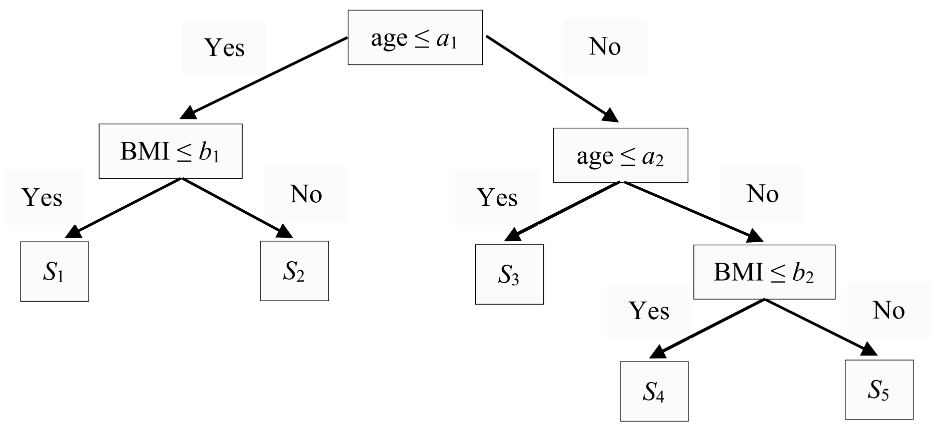
Age	*BMI*	*Actual medical costs (£)*
50	26.3	27,809
48	28.0	23,568
28	24.0	17,663
45	22.9	21,099
59	29.8	30,185
56	20.0	22,413
38	19.3	15,821
61	29.9	30,942
34	25.3	18,972

An Actuary has suggested using a recursive binary decision tree algorithm as the basis for the model. This involves splitting the space spanned by age and BMI into disjoint regions as follows:

1. Node 1 – Split at age a_1
2. Node 2 – Split the region ‘age $\leq a_1$’ at BMI = b_1
3. Node 3 – Split the region ‘age $> a_1$’ at age = a_2
4. Node 4 – Split the region ‘age $> a_2$’ at BMI = b_2

where a_1, a_2, b_1 and b_2 are constants with $a_1 < a_2$.

This algorithm splits the age and BMI space into five regions denoted $S_1, \ldots, S_5$, which are represented by the diagram below:



The resulting recursive binary decision tree model predicts medical costs that are a constant α_k in region S_k, $k = 1, \ldots, 5$.

(i) Derive the formula for the least squares estimator of α_k, $k = 1, \ldots, 5$. [3]

Based on past experience, the Actuary suggests setting the splitting points to $a_1 = 46$, $a_2 = 55$, $b_1 = 23$ and $b_2 = 28$.

(ii) Determine, using the training data set, the values of the least squares estimators of α_k, derived in part (i), for each value of $k = 1, \ldots, 5$. [4]

The data set that will be used to test the model consists of three past policyholders and is set out in the table below:

Policyholder reference number	*Age*	*BMI*	*Actual medical costs (£)*
1	57	27.9	27,768
2	60	28.1	30,023
3	40	21.1	18,524

(iii) Determine the predicted medical costs for each of the three policyholders in the test data set using the model fitted in part (ii). [2]

The Actuary decides to fit a linear regression model to the original training data set to compare the performance of the two machine learning models. The linear regression model takes the following form:

$$(\text{Medical costs})_i = \beta_0 + \beta_1 \times \text{age}_i + \beta_2 \times \text{BMI}_i + \varepsilon_i$$

where ε_i represents the error term for the ith policyholder and β_0, β_1 and β_2 are constants. After using least squares estimation on the original training data set, the Actuary estimates the values of the constants to be $\beta_0 = -8500$, $\beta_1 = 304$ and $\beta_2 = 698$.

(iv) Determine the predicted medical costs for each of the three policyholders in the test data set using the linear regression model. [2]

(v) Compare, qualitatively, using your answers to parts (iii) and (iv), the performance of the two machine learning models in predicting the medical costs of the policyholders in the test data set. [4]

[Total 15]

9 Consider the following time series process:

$$Y_t = 1 + 0.3\ Y_{t-1} + 0.1\ Y_{t-2} + e_t$$

where e_t is a white noise process with variance σ^2.

(i) Determine whether Y_t is stationary and identify the values of p, d and q for which the process is an ARIMA(p,d,q) process. [3]

Let ρ_k and ϕ_k denote the values at lag k of the autocorrelation and partial autocorrelation functions, respectively.

(ii) Determine the autocorrelation values ρ_1, ρ_2 and ρ_3. [4]

(iii) Determine the partial autocorrelation values ϕ_1, ϕ_2 and ϕ_3. [3]

A sample of the process Y_t is taken in which the sample autocorrelation values are equal to the theoretical values ρ_k.

(iv) Determine the minimum sample size, n, necessary to reject the null hypothesis of a white noise process, under the Ljung and Box 'portmanteau' test using three lags and a 5% significance level. [6]

(v) Discuss the relative merits of using a large or a small number of lags in the Ljung and Box 'portmanteau' test by considering how the value of n in part (iv) would vary if a different number of lags were used or if the sample autocorrelation values were not equal to the theoretical values. [5]

[Total 21]

END OF PAPER