INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

23 September 2021 (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 'CS2B_Sept_21_Qu_1_Data.csv' file contains mortality data for a particular population of females in the year 1921, taken from the Human Mortality Database. The file contains the following three variables:

Age	Age, x, in single years from 25 to 90
mu_x	Crude values of the forces of mortality at age x, μ_x
Exposed_x	Exposed to risk at age x nearest birthday

Before answering this question, the 'CS2B_Sept_21_Qu_1_Data.csv' file should be loaded into R and assigned to a data frame called *mortalitydata*.

(i) Plot the crude values of μ_x against age, x, using a line graph. [6]

(ii) Comment on the age pattern of the crude values of μ_x shown in the graph in part (i). [4]

The Gompertz law of mortality can be expressed as follows:

$$\mu_x = Bc^x$$

(iii) Perform a graduation of the crude values of μ_x using the Gompertz formula, specifying the values of B and c in your answer script. [7]

(iv) Perform a chi-square goodness-of-fit test of the graduated values of μ_x from part (iii) using a 95% confidence level. [9]

[Total 26]

2 Before answering this question, the R package for copulas should be loaded into R using the following code:

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

A copula model is to be constructed of the dependence between annual average temperature X, in degrees Celsius, and annual rainfall Y, in millimetres, in a region at risk of desertification. The marginal distributions of X and Y are:

$$X \sim N(20, 3^2), \quad Y \sim N(200, 50^2).$$

Two copulas are under consideration to model the dependence of X and Y:

Copula 1: A Gaussian copula with $\rho = -0.5$.

Copula 2: A Student's t copula with 3 degrees of freedom and $\rho = -0.5$.

The following R code generates n simulations of jointly distributed random variables U and V, where the marginal distributions of U and V are the uniform distribution on [0, 1] and the dependency between U and V is given by a Gaussian copula with correlation ρ (*rho*):

```
rCopula(n, normalCopula(param = rho))
```

The corresponding code for the Student's *t*-copula with *nu* degrees of freedom is:

```
rCopula(n, tCopula(param = rho, df = nu))
```

The output from the function `rCopula` is in the format of an $n \times 2$ matrix whose rows are the observations of (U, V).

Before answering part (i), generate the matrix, *GaussXY*, in R using the following code:

```
set.seed(3)
GaussUV = rCopula(200000, normalCopula(param = -0.5))
GaussXY = matrix(nrow = 200000, ncol = 2)
GaussXY[,1] = qnorm(GaussUV[,1], mean = 20, sd = 3)
GaussXY[,2] = qnorm(GaussUV[,2], mean = 200, sd = 50)
```

The 200,000 × 2 matrix, *GaussXY*, represents 200,000 observations of (X, Y) under Copula 1.

(i) Generate 200,000 observations of (X, Y) using Copula 2 and the marginal distributions of X and Y, assigning the observations to a 200,000 × 2 matrix named *t3XY*. You should set a random number generator seed of 3 before generating the observations and use the R function, `head()`, to display the first six rows of *t3XY* in your answer script. [5]

(ii) Plot, on a single graph, line graphs of the estimated conditional mean of Y given that $X \geq x$ against x for values of x at intervals of 1 from 10 to 30 inclusive for each of the two generated copulas. You should use separate colours to identify each of the two line graphs and set the y-axis range from 50 to 250. [16]

(iii) Comment on the key features of your plot in part (ii). [6]

(iv) Comment on the implications of your plot in part (ii) if the distribution of temperatures, X, is expected to become more weighted towards higher values as a result of climate change. [3]

[Total 30]

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)
```

A new start-up bank is considering which product it should first launch to market. One of the products it is considering is a personal loan. An Actuary employed by the bank wants to construct a simple decision tree machine learning model to determine which loan applications to approve.

The Actuary has decided to construct the decision tree model to classify each potential customer as either expected to default (default = 1) or not expected to default (default = 0) based on the following inputs:

- a categorical feature, f_1, that takes values –1, 0 or +1
- a continuous feature, f_2, that takes values between –1 and 1.

With no previous customer data to draw on, the Actuary decides to construct a specimen data set to train the decision tree model containing the following information for each of n hypothetical customers, where n = 10,000:

- feature f_1 taking the values –1, 0 and +1 with equal probability
- feature f_2 being uniformly distributed on [–1,1]
- a field containing the value 1 if the customer defaults on the loan and 0 otherwise.

To determine whether a hypothetical customer will default, the Actuary decides to model the probability of default as:

$$\frac{\exp(f)}{1+\exp(f)}$$

where $f = a + bf_1 + cf_2$ and a, b and c are parameters.

Before constructing the specimen data set, the Actuary first needs to generate three sets of 10,000 observations from the uniform distribution on [0, 1].

(i) Generate a 10,000 × 3 matrix named U of observations from the uniform distribution on [0, 1]. Each column of U should contain 10,000 observations that have been generated together but separately from the observations in other columns. You should set a random number generator seed of 12345 before generating the observations in the first column and use the R function, `head()`, to display the first six rows of U in your answer script. [4]

Let U_{ij} represent the (i, j) entry of matrix U.

The Actuary proceeds with creating the specimen data set by constructing:

- an n-component vector f_1 whose ith component is equal to –1 if $U_{i1} \leq \frac{1}{3}$, to 0 if $\frac{1}{3} < U_{i1} \leq \frac{2}{3}$ and to +1 if $U_{i1} > \frac{2}{3}$
- an n-component vector f_2 whose ith component is equal to $2(U_{i2} – 0.5)$
- an n-component vector f whose ith component is the value of f given by the formula above for customer i
- an n-component vector *defprob* whose ith component is the probability of default given by the formula above for customer i
- an n-component vector *default* whose ith component is equal to 1 if U_{i3} is less than or equal to the probability from the vector *defprob* for customer i and 0 otherwise
- a data frame in the format required to construct decision trees using the R package `rpart`, incorporating the vectors f_1, f_2 and *default*.

The R code to create the data frame in the format required, given the vectors f_1, f_2 and *default*, is:

```
data.frame("f1" = f1, "f2" = f2, "default" = default)
```

(ii) Generate a data frame named *specimen*, in line with the Actuary's construction rules, corresponding to the parameter values n = 10,000, a = 0 and $b = c$ = 0.5. Use the R function, `head()`, to display the first six rows of *specimen* in your answer script. [15]

(iii) Determine the expected probability of customer default based on the specimen data. [2]

(iv) Comment on how realistic your answer to part (iii) is in terms of what the bank may expect in practice, suggesting how the parameter a could be changed to make it more realistic. [4]

The R code for fitting and plotting a decision tree object called *tree*, that predicts whether a customer will default or not, is as follows:

```
tree = rpart(default ~ ., data = specimen, method = "class")
rpart.plot(tree, digits = 3, type = 5, extra = 106)
```

(v) Plot the decision tree object, called *tree*, using the R code above. [1]

(vi) State, using the decision tree plot, whether a customer with features $f_1 = 0$ and $f_2 = –0.425$ is expected to default or not. [1]

Before launching the product, the Actuary gathers actual customer data from another bank, that is available in the public domain, to assess the predictive power of the decision tree model. The 'CS2B_Sept_21_Qu_3_Data.csv' file contains the actual data for 100 customers in the required format for the decision tree model.

Before answering the remainder of this question, the 'CS2B_Sept_21_Qu_3_Data.csv' file should be loaded into R and assigned to a data frame called *actual*.

(vii) Show the first six rows of the *actual* data frame in your answer script using the R function, `head()`. [2]

The R code for generating the predicted default classifications of the policyholders in the *actual* data frame using the *tree* decision tree object is as follows:

```
predict(tree, actual, type = 'class')
```

(viii) Generate the predicted default classifications of the 100 customers in the *actual* data frame using the R code above, assigning the predictions to a vector called *predict_defaults*. Use the R function, `head()`, to display the first six values of *predict_defaults* in your answer script. [2]

(ix) Generate a confusion matrix of actual defaults versus predicted defaults for the 100 actual customers, displaying the matrix in your answer script. [2]

(x) Determine, using R, the values of the precision and recall percentages for the decision tree model's predictions for the 100 actual customers, where the true positive is defined as the case where the model predicts a default for a customer that has actually defaulted. [5]

Based on the value of the recall percentage for the 100 actual customers, the bank's Finance Director concludes that the decision tree model needs no further refinement and can be used to determine which loan applications to approve when the product is launched.

(xi) Comment on the Finance Director's conclusion. [6]

[Total 44]

END OF PAPER