INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

30 September 2020 (am)

Subject CS2B – Risk Modelling and Survival Analysis Core Principles

Time allowed: One hour and forty-five minutes

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

If you encounter any issues during the examination please contact the Examination Team on T. 0044 (0) 1865 268 873.

1 An actuary is investigating the asymptotic behaviour of the sample autocorrelation function for two time series models.

(i) Generate, using a random number generator seed of 967, a simulated sequence of $n = 200$ observations for a first-order moving average process with parameter $\beta_1 = 0.4$, assigning the simulated values to a vector called *YMA*. [2]

(ii) Generate, using a random number generator seed of 967, a simulated sequence of $n = 200$ observations for a first-order autoregressive model with parameter $\alpha_1 = 0.45$, assigning the simulated values to a vector called *YAR*. [2]

(iii) Plot, on four separate graphs, the sample autocorrelation function (sample ACF) and sample partial autocorrelation function (sample PACF) for each of the two time series models, *YMA* and *YAR*, generated in parts (i) and (ii). [6]

(iv) Comment on the general features of the plots in part (iii) with reference to whether they are consistent with the theoretical behaviour of the corresponding functions for the true models. [4]

The *acf()* function in R can also provide a vector output of the sample ACF values, with component *i* giving the sample ACF at lag $i - 1$, provided that the *plot* argument of the function is set to 'FALSE'.

(v) Determine the numerical values for the sample ACF at lag 2, for each of the two time series models, *YMA* and *YAR*, generated in parts (i) and (ii). [2]

(vi) Construct R code that

- first sets a random number generator seed of 967; and then
- generates 1,000 random vectors (of length $n = 200$) for each of the two models in parts (i) and (ii); and
- assigns the values of the sample ACF at lag 2 for each random vector to two vectors *ACF2MA* and *ACF2AR* (each of length 1,000).

[8]

(vii) Determine the mean and variance of the two vectors, *ACF2MA* and *ACF2AR*, generated in part (vi). [3]

(viii) Plot, on two separate graphs, the histograms of the two vectors, *ACF2MA* and *ACF2AR*, generated in part (vi). [4]

(ix) Comment on the results in parts (vii) and (viii), including whether they agree with the expected asymptotic behaviour. [9]

[Total 40]

2 Before answering this question, generate the vector, X, in R using the following code:

```
set.seed(1027); X = rexp(n=1000, rate=0.01)
```

The vector X represents the gross claim sizes of 1,000 claims. The payments are to be split between an insurance company and its reinsurer under an Excess of Loss reinsurance arrangement with a retention level $M = 400$.

(i) Determine the proportion of the claims that are fully covered by the insurer. [2]

(ii) Generate an additional vector, Y, which is of the same length as X, such that Y represents the amounts to be paid by the insurer for each component of X. [1]

(iii) Generate an additional vector, Z, which is of the same length as X, such that Z represents the amounts to be paid by the reinsurer for each component of X. [1]

An actuary assumes that the underlying gross claims distribution follows an exponential distribution of some unknown rate λ. The actuary needs to estimate λ using only the claim amounts recorded in vector Y.

(iv) Construct R code that calculates the log-likelihood, as a function of the parameter λ, given the claim amounts data in vector Y. [10]

(v) Determine the value of λ at which the log-likelihood function reaches its maximum. [6]

[Total 20]

3 An analyst is investigating a life insurance portfolio data set that comprises two variables, $x1$ and $x2$, for 200 policyholders. The analyst is exploring whether the 200 policyholders can be divided into two clusters (labelled A and B) based on the two variables, $x1$ and $x2$.

Before answering this question, generate the data set, *portfolio*, in R using the following code:

```
set.seed(2019);
portfolio = data.frame(x1=rnorm(200,3,1),
x2=scale(c(rnorm(70,4,1), rnorm(130,10,1))))
```

In the first stage of the investigation, the analyst decides to assign the first 100 policyholders in the data set to cluster A, and the remaining policyholders to cluster B.

(i) Construct a new column in the data set, *portfolio*, called *group_label_stage1*, containing the policyholder cluster labels, defined above. [4]

(ii) Determine the coordinates $(x1_A, x2_A)$ of the centre of cluster A and the coordinates $(x1_B, x2_B)$ of the centre of cluster B. [6]

(iii) Construct a new column in the data set, *portfolio*, called *dist_A* containing the Euclidean distances between the policyholders and the centre of cluster A. [4]

(iv) Construct a new column in the data set, *portfolio*, called *dist_B* containing the Euclidean distances between the policyholders and the centre of cluster B. [4]

The analyst decides to update the cluster labels by assigning to each policyholder the label of the cluster whose centre is nearest, according to the distances calculated in parts (iii) and (iv).

(v) Construct a new column in the data set, *portfolio*, called *group_label_stage2*, containing the updated policyholder cluster labels, defined above. [4]

(vi) Generate a 2x2 matrix showing the number of policyholders with each possible combination of values of *group_label_stage1* and *group_label_stage2*. [2]

(vii) Comment on the matrix generated in part (vi) with reference to how the cluster labels have changed between *group_label_stage1* and *group_label_stage2*. [4]

(viii) Plot the column, $x1$, of data set, *portfolio*, against column, $x2$, (with $x1$ on the x-axis and $x2$ on the y-axis), using two distinct colours to identify clusters A and B according to the label, *group_label_stage2*. [6]

The analyst decides to stop at this stage and to report *group_label_stage2* as the final set of cluster labels.

(ix) Comment on this decision. [6]

[Total 40]

END OF PAPER