INSTITUTE AND FACULTY OF ACTUARIES

CURRICULUM 2019

SPECIMEN EXAMINATION

Subject CS2B – Risk Modelling and Survival Analysis

Time allowed: One hour and forty-five minutes

INSTRUCTIONS TO THE CANDIDATE

1. *You have 1 hour and 45 minutes to complete this examination paper.*

2. *At the end of the examination you should upload a Microsoft Word file including your answers together with sufficient R code for the Examiners to work out how you arrived at your answers.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 3 questions, beginning your answer to each question on a new page of your Word file.*

5. *The CSV files machinelearn.csv and survival.csv accompany this exam paper.*

The filename of your Word file must include your ARN, and the paper sat (e.g. "9000000 CS1B" and each page of the file should contain your ARN as a header or footer.

Please note that the content of this booklet is confidential and you are not to discuss or reveal the content under any circumstances nor are they to be used in a further attempt at the examination.

If you encounter any issues during the examination please contact the Online Education team at online_exams@actuaries.org.uk T. 0044 (0) 1865 268

Question 1

The data file "machinelearn.csv" contains demographic data on 588 districts of England and Wales in the 1860s. The features included in the data set are defined as follows:

Name of area	Name of the district.
PopDensity	Population density in persons per acre.
SexRatio	Population sex ratio (males per female).
PropMAgric	Proportion of adult males working in agriculture.
PropMMining	Proportion of adult males working in mining.
PropFManuf	Proportion of adult females working in manufacturing.
PropFDomServ	Proportion of adult females working in domestic service.
DRTuberculosis	Death rate from tuberculosis.
DRLung	Death rate from other diseases of the lungs.

All the features except DRTuberculosis and DRLung were obtained from census data.

A medical historian is interested to know whether districts which were demographically different had different death rates from diseases of the lung. She suggests using the demographic features to divide the 588 districts into clusters, and then examining the distribution of death rates within each cluster.

(i) Explain why the raw data should be scaled before applying a clustering algorithm, illustrating you answer with examples from the "machinelearn.csv" data set. [6]

(ii) Perform a *k*-means cluster analysis on the data, normalised using *z*-scores, using the six features PopDensity, SexRatio, PropMAgric, PropMMining, PropFManuf, PropFDomServ, dividing the data into 6 clusters. Calculate the number of districts in each cluster and the mean values of the clusters on the six features. [8]

(iii) Briefly describe the characteristics of each of the clusters you identify. [6]

(iv) Calculate the mean death rates from tuberculosis and other diseases of the lungs for each cluster. [5]

(v) Comment on how successful the clusters are at identifying groups of districts with different death rates. [5]

[Total 30]

Question 2

The data file "survival.csv" contains data on the duration between the date of first cohabitation and the date of first birth for 4,091 women in Armenia in the late-twentieth and earl twenty-first century. They were obtained from a survey carried out in 2010. Women who had not given birth before the survey were treated as censored at the survey date. The variables are defined as follows:

DUR	Duration in months between first cohabitation and first birth (or censoring)
EVENT	Takes value 1 if DUR is duration until birth, and 0 of DUR is duration until censoring
AGE	Age of women in years at START
URBAN	Takes value 1 if woman lived in an urban area, and 0 if she lived in a rural area
POOREST	Takes value 1 if woman was in poorest wealth stratum, and 0 otherwise
POOR	Takes value 1 if woman was in second poorest wealth stratum, and 0 otherwise
MIDDLE	Takes value 1 if woman was in middle wealth stratum, and 0 otherwise
RICH	Takes value 1 if woman was in second richest wealth stratum, and 0 otherwise
YEAR	Date of first cohabitation (measured in years minus 1900, so 87.333 means April 1987)
LOWER	Takes value 1 if woman had a low level of education, and 0 otherwise.

(i) Plot the Kaplan-Meier estimate of the survival function of duration to first birth for all women. [6]

(ii) Estimate a Cox regression model of the duration between cohabitation and first birth using all available covariates, describing your results. [14]

(iii) By applying the likelihood ratio test, estimate a parsimonious Cox regression model including only statistically significant covariates. [15]

[Total 35]

Question 3

Let y_t be a sequence of random variables that follows the moving average model of order 1

$$y_t = y_{t-1} + b\varepsilon_t, \qquad t = 2,3,...,N$$

where:

- ε_t are Gaussian, independent and identically distributed $N\left(0,\sigma^2\right)$
- b is a known constant in the interval $[-1,1]$
- $y_1 = 0$

According to this model, y_t is equal to a noise term ε_t plus (or minus, depending on the sign of b) some fraction of the noise term ε_{t-1} at the previous time.

(i) Write an R function "`MA1()`", which generates a sequence y_t from the model above.

The function should have the following arguments:

- the final time N
- the parameters b and σ

The function should return a list with three components: b, σ and a vector **Y** containing the generated sequence.

The function should check that N and σ are positive and that b lies in the interval $[-1, 1]$ and return an error if not. [10]

(ii) Plot four generated sequences of **Y** using the function above, in a 2 × 2 display, for values $N = 200$, $b = -0.35$ and $\sigma = 0.4$. [4]

(iii) Run in the R-console window the following two lines:

```
set.seed(num)
x=arima.sim(model=list(ar=.3,ma=.6),n=200)
```

where `num` is replaced with the last four digits of your telephone number (or any other number of your choosing). [3]

(iv) Explain what actions the two lines in part (iii) are performing. [4]

(v) Plot the auto correlation function (ACF) and partial auto correlation function (PACF) for the object x in part (iii), and comment on this output. [7]

(vi) Fit an ARMA(2,2) model to the vector x, including diagnostics, and comment on the result. [7]

[Total 35]

END OF PAPER