INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

20 September 2022 (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 Before answering this question, the following R code should be run:

```
set.seed(912)

y=arima.sim(list(order=c(0,1,0)),n=400)
fit=arima(y,order=c(1,0,0))
fit
```

(i) Comment briefly, in your own words, on each line of R code above. [2]

(ii) (a) State the standard error of the ar1 parameter estimate in the `fit` object created by the R code above.

(b) Determine the corresponding 95% confidence interval.
[2]

(iii) Comment on your answer to part (ii). [2]

(iv) Calculate the predicted values using the model `fit`, the future values of y for ten steps ahead. [2]

(v) Generate, and display in your answer script, a matrix, A, of dimension 10×2, which contains the predicted values in part (iv) together with the corresponding standard errors. [2]

(vi) Construct R code to generate a plot that contains the time series data y, together with the 'ten steps ahead' predictions from part (iv) and their 95% prediction intervals. [4]

(vii) Construct R code to display, next to each other, the sample AutoCorrelation Function (sample ACF) and sample Partial AutoCorrelation Function (sample PACF) for the data set y. [2]

(viii) Construct R code to display, next to each other, the sample ACF and sample PACF for the residuals of the model `fit`. [2]

(ix) Comment on the graphical output of parts (vii) and (viii). [4]

(x) Perform the Ljung and Box portmanteau test for the residuals of the model `fit` with four, six and twelve lags. [4]

(xi) Comment, based on your answers to parts (ix) and (x), on whether there is enough evidence to conclude that the model `fit` is appropriate. [4]
[Total 30]

2 Before answering this question, the survival package should be loaded into R with the following code:

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

The government of Country U has asked a non-profit organization to study possible adverse effects of a new vaccine administered to individuals, with particular reference to the possibility of blood clots within the first 28 days of receipt of a vaccine.

Before answering this question, the 'CS2B_S22_Qu_2_Data.csv' file should be loaded into R and assigned to a data frame called 'data'. This .csv file contains the data from an investigation for 2,400 individuals. The file contains the following six variables:

`Life`: patient identifier (integers 1, 2, … 2,400)

`Drug`: indicator (1 = received vaccine, 0 = did not receive vaccine)

`Age`: indicator (0 = age less than or equal to 50, 1 = age greater than 50)

`co_morbidity`: indicator (1 = individual has another chronic disease at the time of receipt of vaccination, 0 = no chronic disease)

`Status`: indicator (0 = censoring due to the end of period, 1 = censoring due to death (reason unknown), 2 = admission to hospital due to blood clots within 28 days of receipt of vaccine, 3 = admission to hospital due to reasons other than blood clots within 28 days of receipt of vaccine)

`Time`: duration in days at which admission to hospital/censoring occurred (integers with a range of 0–28; 0 = day of vaccination).

(i) Comment on whether the censoring in this investigation is likely to be non-informative. [3]

(ii) Construct a table named 'data_main', which is the same as 'data' but with a new column added. The newly added column should be named 'ST' and should contain the values:

- 0 if 'Status' in 'data' is 0 or 1 or 3
- 1 if 'Status' in 'data' is 2.

Display the last 20 rows of 'data_main'. [6]

(iii) Plot the Kaplan–Meier survival functions required to analyse the effect of vaccination on blood clots assuming that censoring is non-informative. You should plot both survival functions on the same axes, using separate colours to identify each survival function. You should use a range from 0.97 to 1 on the y-axis. [9]

(iv) Comment on your plot from part (iii). [2]

Analysts in the organization have decided to analyse further by using Cox's proportional hazards model and by adding covariates into the investigation.

The following decisions were made:

- Significance of covariates would be tested with interactions.
- At least two covariates would be used.
- Two covariates to be compulsorily used are vaccine indicator and age.

They are now deciding to add one more covariate: co-morbidity.

(v) Test the hypothesis, using the likelihood ratio statistic, that co-morbidity has no effect on blood clots allowing for vaccine indicator and age, stating the null and alternative hypotheses and using the Breslow method for tie handling.

[14]

[Total 34]

3 The dataset 'CS2B_S22_Qu_3_Data.csv' contains the following four variables: `mpg`, `disp`, `qsec`, `hp`.

You are tasked to build a regression model to explain the response, `hp`, in terms of the features `mpg, disp` and `qsec`. Thus, the model must be in the form:

$$y_i = \beta_0 + \beta_1 * \texttt{mpg}_i + \beta_2 * \texttt{disp}_i + \beta_3 * \texttt{qsec}_i + \varepsilon_i$$

where β_0, β_1, β_2 and β_3 are the regression parameters and ε represents the error term.

This model can be written compactly as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is the vector containing the response data and $\boldsymbol{X}$ is the design matrix.

(i) Write R code to load the dataset 'CS2B_S22_Qu_3_Data.csv' into R, create the design matrix, $\boldsymbol{X}$, including column headings and display the first six rows of this design matrix.

[**Note**: The $\boldsymbol{X}$ you create must be of the type 'matrix'.] [5]

You would like to fit the above model by minimizing the following:

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \frac{\lambda}{2}||\boldsymbol{\beta}||^2$$

It can be shown that, for a given value of λ, the estimate $\boldsymbol{\beta}_\lambda$ of $\boldsymbol{\beta}$ that minimises the expression above is such that:

$$(\boldsymbol{X}^t\boldsymbol{X} + \lambda * \boldsymbol{I})\,\boldsymbol{\beta}_\lambda = \boldsymbol{X}^t\boldsymbol{y}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{X}^t$ is the transpose of $\boldsymbol{X}$.

(ii) State the name of this modelling approach. [1]

(iii) Construct an R function, `ridge_fit`, that takes as inputs the value of λ, the vector of response data and the design matrix, and then returns the value of $\boldsymbol{\beta}_\lambda = (\boldsymbol{X}^t\boldsymbol{X} + \lambda * \boldsymbol{I})^{-1}\boldsymbol{X}^t\boldsymbol{y}$.

[**Hint**: The R function `solve` can be used to compute the inverse of a matrix. That is, for a given invertible matrix, $\boldsymbol{M}$, its inverse, $\boldsymbol{M}^{-1}$, can be computed in R by running the following code: `solve(M)`.] [6]

(iv) Calculate and display the value of the vector $\boldsymbol{\beta}_2$ (i.e. the value of vector $\boldsymbol{\beta}_\lambda$ for $\lambda = 2$). [2]

(v) Construct R code to compute the values of the vectors $\boldsymbol{\beta}_\lambda$ and store them into successive rows of a matrix named matrix_LAMBDA, for $\lambda = i/10$, where $i = 0, 1, 2, ..., 10{,}000$. [5]

(vi) Display the top six rows of matrix_LAMBDA. [1]

You would like to select the best value of λ. A senior statistician suggests that you should base your selection on a statistical information criterion. However, most statistical information criteria depend on the so-called effective dimension of the model. For this task, it can be shown that the effective dimension is the sum of the diagonal elements of the following matrix: $\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \lambda * \boldsymbol{I})^{-1}\boldsymbol{X}^t$.

(vii) Construct an R function called `dim_fit` that takes as inputs the design matrix and the value of λ, and then returns the corresponding effective dimension. [6]

(viii) Construct R code to compute the values of the effective dimensions and store them in a vector called `vector_dim`, for $\lambda = i/10$, where $i = 0, 1, 2, ..., 10{,}000$. [4]

(ix) Plot the effective dimension as a function of λ. [4]

(x) Comment on your plot from part (ix). [2]

[Total 36]

END OF PAPER