INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

12 April 2024 (am)

Subject CS1 – Actuarial Statistics
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 An insurance company wants to study the association between the number of years their clients spent in education and their claim amounts.

Data from 25 randomly selected claims are contained in the file `AmountYears.RData` in the following two variables:

`ClaimAmount` – this is the claim amount (in £).

`EducationYears` – this is the number of years the client spent in education.

(i) Plot the claim amounts against the years of education. [3]

A statistician suggests fitting a simple linear model to the data.

(ii) (a) Fit a linear model to the data. [2]

(b) Plot the regression line by adding it to the graph in part (i). [2]

Another statistician looks at the plot in part (i) and suggests a non-linear relationship.

(iii) (a) Fit a model with a quadratic term added to the model fitted in part (ii). [3]

(b) State the equation of the fitted model. [1]

(iv) Comment on the suitability of the quadratic model in part (iii), compared to the model in part (ii), based on the output from part (iii). [3]

[Total 14]

2 A financial consultancy working with large firms wishes to model the relationship between a firm's assets and the number of senior management positions in the firm. The data file `firms.Rdata` contains the variables:

`assets` – this is the value of assets (in millions of £).

`sn_positions` – this is the number of senior managements positions.

You can load the file into R using `load("firms.Rdata")`.

(i) Plot the number of senior management positions as a function of assets. [3]

(ii) Plot the number of senior management positions as a function of log10_assets where log10_assets is the assets at log10 scale. [4]

(iii) Comment on your plots in parts (i) and (ii). [2]

An analyst wishes to check if the number of senior management positions follows a Poisson distribution.

(iv) Calculate the mean of the number of senior management positions. [1]

Use the command `set.seed(222)` to initialise the random number generator.

(v) Generate a sample of size equal to the number of firms from a Poisson distribution with parameter equal to the mean calculated in part (iv). [2]

(vi) Plot a histogram of the sample simulated in part (v) and a histogram of `sn_positions` on two separate graphs but on the same scale specifying appropriate axis limits and labels. [6]

(vii) Comment on your plots in part (vi). [3]

[Total 21]

3 A Multiple-Choice (MC) test with 20 questions requires a minimum of 16 correct answers for students to pass the test. A student prepares for the test using a mobile phone application that generates random practice tests with 20 questions per test.

Load the file `MCtestResults.Rdata` into R. This creates two variables:

`CorrectOutOf20Questions` – this contains the number of correct answers the student has achieved with the mobile phone application in each of 50 generated practice MC tests.

`TrialNumber` – this contains the corresponding test number from 1 to 50.

The student assumes that the test score, X, which is the number of correctly answered questions per test, has a binomial distribution, $X \sim \text{Bin}(n, p)$ with $n = 20$.

(i) Estimate the parameter p using the test scores in `MCtestResults.Rdata`, assuming that the test scores are independent of each other and identically distributed. [2]

(ii) Calculate the probability that the student will pass a test based on your estimate of p in part (i). [2]

(iii) Calculate the proportion of practice tests that the student has passed. [1]

(iv) Comment on the probability that the student will pass a test, based on your answers to parts (i), (ii) and (iii). [4]

(v) Plot the number of correct answers in each of the practice tests against the test number on the horizontal axis. [3]

(vi) Comment on the plot in part (v). [2]

A linear model is fitted to the data, which predicts that the number of correct answers in the next test (test number 51) will be 18.085.

(vii) Calculate the probability for the student to pass the next test (test number 51). [4]

[Total 18]

4 Consider a random variable, X, following a modified exponential distribution with Cumulative Distribution Function (CDF):

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp(-\lambda x^2), & x \geq 0 \end{cases}$$

(i) Plot the CDF $F(x)$ as a function of x for $x = 0, 0.1, 0.2, \ldots, 9.9, 10$ when $\lambda = 0.2$. [4]

(ii) Plot the density function $f(x)$ of X as a function of x for $x = 0, 0.1, 0.2, \ldots, 9.9, 10$ when $\lambda = 0.2$. [5]

A random sample of 100 values of X is provided in `randomSample.Rdata`. Loading the sample data into R will generate a vector x with 100 values representing the sample.

(iii) Calculate the value of the log likelihood function for the parameter λ at the point $\lambda = 0.2$ based on this random sample. [3]

(iv) Plot the values of the log likelihood function for the parameter λ based on the sample in `randomSample.Rdata`. Your plot of the log likelihood function must be for values of $\lambda = 0.01, 0.02, \ldots, 0.99, 1$. [7]

The maximum likelihood estimator for the parameter λ based on a random sample $X_1, \ldots, X_N$ is given by:

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} X_i^2}$$

(v) Estimate the value of λ using the maximum likelihood estimator given above and the sample in `randomSample.Rdata`. [3]

(vi) Comment on the plot in part (iv) and the estimate in part (v). [2]

[Total 24]

5 An insurance company, which currently only sells home insurance, is interested in entering the car insurance market. An underwriting manager at the company believes that the age and gender of the policyholder will be the most important factors in estimating the number of claims made under a car insurance policy.

The underwriting manager has commissioned a survey of its current home insurance customers who also have car insurance, choosing a male customer and a female customer for every age from 18 to 65, asking them how many car insurance claims they have made in the past 3 years. This dataset is saved in the file `ClaimsData.Rdata`. After loading this data into R, using the command `load("ClaimsData.Rdata")`, the data frame `ClaimsData` will be available, which contains the following three variables:

`age` – this is the age (in years) of the policyholder.

`gender` – this is either 'M' for male or 'F' for female.

`claim_count` – this is the number of car insurance claims reported by the policyholder over the past 3 years.

(i) Fit a normal linear regression model to the data using `claim_count` as the response variable and `age` as the explanatory variable. Your answer should include the estimated intercept and slope of the regression line. [3]

A colleague suggests that the response variable would be better modelled as having a Poisson distribution.

(ii) Fit a Generalised Linear Model (GLM) to the data using `claim_count` as the response variable and `age` as the explanatory variable, assuming a Poisson distribution for the response variable. Your answer should include the estimated coefficients and the Akaike's Information Criterion (AIC) of the fitted model. [4]

The underwriting manager wishes to compare the fit of the GLM in part (ii) against that of the normal linear regression model in part (i).

(iii) Explain why scaled deviances cannot be used to compare the fit of the models in parts (i) and (ii). [3]

(iv) Fit, by choosing a suitable argument for `family` in the `glm` command, a GLM to the data that is equivalent to the model fitted in part (i). Your answer should include the estimated coefficients and the AIC of this fitted model. [4]

(v) Compare the fit of the models fitted in parts (ii) and (iv). [2]

The underwriting manager believes the Poisson GLM would be improved by adding the explanatory variable `gender` as well as its interaction with `age`.

(vi) (a) Fit a Poisson GLM to the data of the form `age*gender`. Your answer should include the estimated coefficients and the AIC of this fitted model.

(b) Compare, using scaled deviances, the fit of this model to that in part (ii).

[7]
[Total 23]

END OF PAPER