INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

12 September 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 For an investigation into the buying behaviour of consumers shopping at an online store, data about the age, the time per month spent on the online shop's website and the number of purchases from the shop per year are recorded.

This dataset is saved in the file 'onlineShop.Rdata'. After loading the data into R, using the command `load("onlineShop.RData")`, the data frame `onlineShop` will be available. The data frame `onlineShop` contains the following three variables:

- `age` – the age (in years) of the customer
- `time_spent_on_website` – the time per month spent on the shop's website (in hours)
- `purchases_per_year` – the number of purchases from the shop per year.

(i) Fit a Generalised Linear Model (GLM) to the data assuming a Poisson distribution with a log link function, using `purchases_per_year` as the response variable and `age` and `time_spent_on_website` as explanatory variables without an interaction term. Your answer should include the estimated coefficients, the deviance and the Akaike's Information Criterion (AIC) of the fitted model. [4]

(ii) State an expression for the expected number of purchases per year in terms of the variables `age` and `time_spent_on_website` using the fitted model from part (i). [1]

It is suggested that a log link function may not be appropriate, and a square root link function should be used instead, that is $g(\mu) = \sqrt{\mu}$. In R, a square root link function can be specified using `family = poisson(link = "sqrt")`.

(iii) Fit a GLM to the data assuming a Poisson distribution with a square root link function using `purchases_per_year` as the response variable and `age` and `time_spent_on_website` as explanatory variables without an interaction term. Your answer should include the estimated coefficients, the deviance and the AIC of the fitted model. [4]

(iv) State an expression for the expected number of purchases per year in terms of the variables `age` and `time_spent_on_website` using the fitted model from part (iii). [1]

(v) Compare the models in parts (i) and (iii) and decide which model should be preferred using the deviance. [1]

It is suggested that the deviance is not an appropriate measure to compare the two models in parts (i) and (iii), and that the AIC should be used instead.

(vi) Comment on this suggestion in the context of comparing the two models in parts (i) and (iii). In particular, identify any advantages that the AIC may have compared to the deviance in this context. [2]

[Total 13]

2 An investment analyst is interested in the relationship between the share price of a company and its Earnings-Per-Share (EPS).

A sample of data for eight companies has been collected. This is stored in the file 'CompanySample.RData'. After loading this data into R, the data frame `CompanySample` will be available that contains two columns:

- `price` – the company's share price (in $)
- `EPS` – the company's earnings-per-share (in $).

(i) Fit a linear model to the sample, with `price` as the dependent variable and `EPS` as the independent variable, and state the estimated coefficients of the model. [3]

(ii) State the value of the coefficient of determination, R^2, for the model fitted in part (i). [1]

(iii) Perform an analysis of variance, using the `anova(…)` command to test if there is a linear relationship between `price` and `EPS`. Your answer should include the F-statistic and p-value of the test. [4]

(iv) Comment on how well the linear model in part (i) fits the data, using your answers to parts (ii) and (iii). [2]

The analyst believes that the observation where price = 99 is an anomaly because of the large size of the value of price and should be removed from the sample.

(v) Fit a linear model to the sample without the observation price = 99 and state the value of R^2. [3]

[**Hint:** The code `CompanySample[-n,]` returns the data frame with the nth observation removed.]

(vi) Compare the fit of the models from parts (i) and (v). [3]

(vii) Discuss if removing the observation where price = 99 is appropriate. [3]

[Total 19]

3 A random variable X with values in the interval [0,1] has the following Cumulative Distribution Function (CDF):

$$F(x) = 1 - (1 - x^a)^b, \quad x \in [0,1]$$

and inverse CDF:

$$F^{-1}(u) = \left(1 - (1-u)^{\frac{1}{b}}\right)^{\frac{1}{a}}$$

where $a > 0$ and $b > 0$ are parameters. The Probability Density Function (PDF) is then given by:

$$f(x) = abx^{a-1}(1 - x^a)^{b-1}, \quad x \in [0,1]$$

You do not need to check the form of the inverse CDF or PDF.

(i) Plot the inverse CDF $F^{-1}(u)$ for an appropriate range of values of u when $a = 0.7$ and $b = 0.5$. [5]

(ii) Simulate 1,000 values of X for $a = 0.7$ and $b = 0.5$ using the inverse transform method and store the 1,000 simulated values in a vector in R for later use.

You must use the command `set.seed(123)` to initialise the random number generator before you start the simulation. [3]

(iii) Plot a histogram of the 1,000 simulated values obtained in part (ii) using relative frequencies. [2]

(iv) Plot the PDF $f(x)$ of X for $a = 0.7$ and $b = 0.5$ for values of $x \in [0,1]$. The graph of the PDF should be superimposed on the histogram produced in part (iii). [4]

(v) Comment on the plot in part (iv) comparing the simulated values of X with the PDF, also taking into account the size of the simulated sample. [2]

(vi) Plot the PDF $f(x)$ of X for $a = 0.7$ and $b = 0.5$ for values of $x \in [0,1]$ in a new plot, and add the PDF $f(x)$ for $a = 3$ and $b = 2$ to the same plot. [4]

A statistician wants to use one of the PDFs in part (vi) as a prior density for a Bayesian analysis.

(vii) Comment on the prior beliefs about the underlying parameter that the two PDFs represent. [3]

[Total 23]

4 Two members of an interview panel (A and B) have scored a large number of applicants (on a scale from 1 to 40), with the purpose of ranking the applicants according to their suitability for a job position. The scores for a sample of twelve applicants are given below, with `SA` and `SB` denoting the scores of interview panel member A and B, respectively:

- `SA = c(35,14,28,33,29,22,19,36,21,30,15,18)`
- `SB = c(38,18,25,30,22,17,23,29,32,31,15,19).`

(i) Plot a suitable graph for assessing the agreement in the two panel members' scores. [2]

(ii) Comment on the agreement in the two panel members' scores, based on the plot in part (i). [2]

(iii) Calculate Pearson's correlation coefficient r for these data. [1]

(iv) Calculate Spearman's rank correlation coefficient r_s for these data. [1]

(v) Calculate Kendall's rank correlation coefficient τ for these data. [1]

(vi) Comment on the suitability of the three correlation coefficients used in parts (iii)–(v), with respect to the purpose of the scoring. [2]

(vii) Perform three statistical tests for a suitable hypothesis of no association between the scores of the two panel members, using one of the correlation coefficients in parts (iii)–(v) for each test.

For each test, your answer should include the p-value and a conclusion. [11]

(viii) Comment on the validity of the tests in part (vii) for testing the hypothesis of no association between the scores of the two panel members. [2]

[Total 22]

5 A sample of 100 daily sales amounts at a store is collected, with the aim to investigate if daily sales amounts are affected by whether or not the store has substantial discount offers available on a particular day. The data are available in the file 'sales_data.RData', which contains the following variables for 100 days:

- `sales.amount` – the daily sales amount
- `discount` – an indicator, showing whether or not the store has substantial discount offers available on each day (discount = 1 if discounts are available; discount = 0 otherwise).

(i) Plot appropriate boxplots for comparing the sales amounts on discount and non-discount days. [3]

[**Hint:** You may find the `boxplot(formula, …)` R command useful.]

(ii) Comment on the effect of discount offers on sales amounts using the boxplots produced in part (i). [2]

A sales analyst considers two distributional assumptions: a normal and a gamma model for the daily sales amounts.

(iii) Fit two separate Generalised Linear Models (GLMs) to the data, one for each of the two distributional assumptions, to investigate the dependence of the daily sales amounts on availability of substantial discount offers on the day. For each model you should use the canonical link function. [4]

(iv) Comment on the impact of the availability of substantial discount offers on daily sales amounts based on the output of the two models fitted in part (iii).

Your answer should include a relevant p-value and interpretation of the estimate of the coefficient of the discount variable for each fitted model. [6]

(v) Determine which of the two models fitted in part (iii) should be preferred for investigating the dependence of daily sales amounts on the availability of substantial discount offers on the day. [3]

(vi) Determine the expected sales amount on a discount day under each of the GLMs fitted in part (iii), explicitly using the estimated coefficients from each model. [4]

(vii) Comment on the comparison of the answers in part (vi) taking into account your answer in part (v). [1]

[Total 23]

END OF PAPER