INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

28th November 2023

Subject CS1B – Actuarial Statistics (Paper B)

Time allowed: 1 Hour 45 Minutes (10.15 – 12.00 Hours)

Total Marks: 100

INSTRUCTIONS TO THE CANDIDATES

1. ***Mark allocations are shown in brackets.***
2. ***Attempt all questions beginning your answer to each question on a new page.***
3. ***Attempt all sub-parts of the question in one document only, unless otherwise instructed to do so.***
4. ***All the detailed guidelines are available on exam screen.***
5. ***Do save your work in solution template on a regular basis.***
6. ***If Any, Data set file(s) accompanying the question paper is available for download on the exam screen.***
7. ***You need to import the same into R studio as soon as you begin the exam.***
8. ***Ensure to copy and paste R codes and output at regular intervals onto the solution template.***
9. ***Please check if you have received complete Question Paper and no page is missing. If so, kindly get new set of Question Paper from the Invigilator.***

AT THE END OF THE EXAMINATION

Please return this question paper to the supervisor separately. You are not allowed to carry the question paper in any form with you. You are requested to save and submit the work before leaving the examination premises.

Q. 1)

i) Hotel room prices in Mumbai are normally distributed with a mean of Rs. 5400 and standard deviation of Rs. 900, whereas prices in Kolkata are normally distributed with a mean of Rs. 3600 and standard deviation of Rs. 1500. Compute the probability that a hotel room price in Mumbai is atleast twice the price in Kolkata. Note that the hotel prices in Mumbai and Kolkata are independent of each other.

Store mean and sd values as x.mean, x.sd, y.mean,y.sd,etc.
*Hint: p**** is a useful function to ascertain CDF.* (5)

ii)

a) Show that the population mean and standard deviation of "differences in hotel prices between Mumbai and Kolkata" (i.e., Mumbai hotel prices – Kolkata hotel Prices) are 1800 and 1749.86. (3)

b) Generate a sample of size 50 for the " differences in hotel prices between Mumbai and Kolkata". (2)

c) Draw qqplot and qqline for the sample generated and comment on the results. (4)

Store population mean and standard deviation of differences as dif.mean and dif.sd
Make sure to set the seed value to 1234 i.e use set.seed(1234) before generating the sample.
Store the sample as dif.sample
Don't paste the sample

iii) Test using 5% significance level whether the mean of *"difference in hotel prices"* is less than 1375 based on the sample generated in part (ii).

a) Perform z-test and comment. (5)

b) Perform t-test (assuming population is not known) and comment. (3)

Please ensure to include test conclusion.

iv)

a) Generate another sample but with size=1000 and store as dif.sample2. (1)

b) Draw qqplot and qqline for the sample generated in (iv.a) and comment on the results. (3)

[26]

Q. 2) You are a Research Analyst and working on a short term assignment. Your professor provided dance data (dance.csv) containing scores of 20 contestants of a famous dance show and asked you to perform the following tasks.

i) Using read.csv load the data and use *head* command to view the first few rows of the data. (2)
No need to paste the output of head command.

dance.csv contains following fields:

- Judges: It indicates the score provided by judges
- Audience: It indicates the sum of scores provided by audience

- Final: It indicates the final score of the contestant

ii) Plot a scatterplot for each pair of data. (4)
Make sure to paste the scatterplot in your answer scripts.

iii) Using scatterplot of part (ii), comment on the relationship between the pairs of data. (3)

iv) Fit a multiple linear regression model with

- Final score as response variable and
- Judges and Audience score as explanatory variables.

a) Store the model as m1. (2)

b) Show summary of the output and also write the equation of the fitted model. (4)

c) Compute confidence interval of all parameters. (3)

d) Comment on the significance of the each explanatory variable either referring to above confidence interval or other statistics. (3)

You published the report showing the above results and reference data on your research website. After few months, a journalist accused the dance show of ignoring audience scores and cited your report.

The sponsors of dance show stated that incomplete data is used for analysis as audience score depends upon the number of audiences provided scores. They shared supplement data providing the number of audiences for each participant.

v) Using the following code to load the number of audiences for 20 contestants

```
audience.count<-c(110,100,90,120,100,100,100,100,110,110,100,
                  100,110,90,100,110,120,120,100,100)
```

Also, verify total count is equal to 2090. (2)

vi) Compute the new audience score by dividing the sum of score provided by audience ("*Audience*") with audience count ("*audience.count*"). (2)

Store as Audience2 and make sure to attach to the dance data.
Please don't paste Audience2 values.

vii) Perform correlation test to check whether any correlation exits between final score and new audience score. (3)

viii) Fit a new multiple linear regression model with

- Final score as response variable and
- Judges and New Audience score as explanatory variables.

Store this model as m2 and show the summary of the output. (3)

ix) Using a suitable statistic from part (iv.b) and (viii). outputs, compare models m1 and m2 and suggest which model is better. (3)
Please write the figures of the statistic while answering.

[34]

Q. 3) An analyst fitted various models on the data containing claims information of 20 policies and shared the output to conduct analysis.

Output 1:

```
Call:
glm(formula = Claim ~ 1, family = poisson(lin = "log"), data = q3)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8439  -0.8975  -0.1791   0.3925   2.5561

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5306     0.1715   3.094  0.00197 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 30.147  on 19  degrees of freedom
Residual deviance: 30.147  on 19  degrees of freedom
AIC: 71.114

Number of Fisher Scoring iterations: 5
```

i) Write the equation of the fitted model using Output 1 as above and specify which distribution is used to model the response variable. (2)

ii) Using output 1, show that the sample mean of the response variable is 1.7 (when rounded to one decimal place). (2)

Output 2:

```
Call:
glm(formula = Claim ~ Gender + Health - 1, family = poisson(lin = "log"),
    data = q3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.22474  -0.81754  -0.07119   0.27453   1.44149

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
GenderF            1.1394     0.3490   3.265 0.001096 **
GenderM            1.1394     0.2216   5.143 2.71e-07 ***
HealthNonDiabetic  -1.4271     0.3939  -3.623 0.000291 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 38.229 on 20 degrees of freedom
Residual deviance: 14.436 on 17 degrees of freedom
AIC: 59.403

Number of Fisher Scoring iterations: 5

iii) Write the equation of the fitted model using Output 2. (4)
Make sure to define the explanatory (categorical) variables.

iv) Output 2 doesn't have an intercept as a coefficient unlike output1. Please provide the reason for the same by comparing glm R formulas. (2)

v) Compare output 1 and output 2, and suggest which model is better using suitable statistics. (2)

vi) Compute log likelihood of the model given in output 2. (3)

vii) Claim follows Poisson distribution. Test for the parameter to be equal to 1.5 at 1% level of significance. (5)
Hint: Use outputs and various sub-parts to determine x for poisson test.

[20]

Q. 4) The table below shows the total claim number (cancellations) per year, Xij, for 4 travel companies over last 4 years.

		Years,j			
		1	2	3	4
Travel Companies, i	Make	455	458	587	531
	Ease	251	322	292	340
	Go	309	246	217	120
	Clear	400	426	470	547

i) Using Empirical Bayes Credibility Theory (EBCT) Model 1, compute the following

a) *Copy the below code to load the data:* (1)

```
q4<-matrix(c(455,251,309,400,
 458,322,246,426,
 587,292,217,470,
 531,340,120,547),
ncol=4,nrow=4)
```

b) $E[m(\theta)]$ (2)

c) $E[s^2(\theta)]$ (2)

d) $Var[m(\theta)]$ (3)

e) Z (2)

ii) Using part (i), calculate the expected claim number for Go and Clear. (3)

iii) What additional information is required to use EBCT Model 2. (2)

iv) Travel company "Ease" launched a membership program last year providing full refund on cancellations. Number of cancellations believed to follow binomial distribution with parameters n=3 and 0.20.

Number of cancellations in last year on 150 memberships are as follows:

Cancellations	0	1	2	3
Members	61	71	15	3

Carry out goodness of fit test for the binomial model specified for number of cancellations on each membership. (5)

[20]

****