

INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

May 2026

CS1B - Actuarial Statistics

Time allowed: 1 Hours 45 Minutes

Total Marks: 100

Q.1) Following table presents the data on the number of matches played and total runs scored in respect of three batsmen viz. Rohit, Virat and Shivam over four seasons of a premier cricketing league:

Number of matches	Rohit	Virat	Shivam
Season 1	14	16	18
Season 2	15	17	19
Season 3	16	18	20
Season 4	14	17	19

Runs Scored	Rohit	Virat	Shivam
Season 1	420	640	720
Season 2	450	680	760
Season 3	500	720	800
Season 4	430	660	750

Sponsoring partners of the premier cricketing league want to estimate the runs to be scored in respect of these players in the upcoming season 5 of the league. It is decided to use Empirical Bayes Credibility Theory Model 2 (EBCT Model 2) for this purpose.

Following code has been written in R for the same:

```

matches <- data.frame(
  Season1 = c(14, 16, 18),
  Season2 = c(15, 17, 19),
  Season3 = c(16, 18, 20),
  Season4 = c(14, 17, 19)
)

runs <- data.frame(
  Season1 = c(420, 640, 720),
  Season2 = c(450, 680, 760),
  Season3 = c(500, 720, 800),
  Season4 = c(430, 660, 750)
)

n <- ncol(runs)
N <- nrow(runs)

X <- runs / matches
Xibar <- rowSums(runs) / rowSums(matches)

Pibar <- rowSums(matches)
Pbar <- sum(Pibar)

Pstar <- sum(Pibar * (1 - Pibar / Pbar)) / (N * n - 1)

m <- sum(runs) / Pbar

s <- mean(rowSums(matches * (X - Xibar)^2) / (n - 1))

```

```
v <- (sum(rowSums(matches * (X - m)^2)) / (n * N - 1) - s) / Pstar
```

- i) Copy the above code lines and display the batting averages (i.e. X) for each player across all the four seasons of the premier cricketing league.

Note: You can directly copy these code lines from **Codes.docx** file in your R Script.

[1]

- ii) Calculate the estimates of $E[m(\theta)]$, $E[s^2(\theta)]$ and $\text{Var}[m(\theta)]$ under EBCT Model 2. [3]

- iii) Calculate the credibility factors Z_i . [2]

- iv) If all the three players are estimated to play 15 matches each in the upcoming season of the premier cricketing league, then estimate the runs to be scored during this season using the assumptions of EBCT Model 2. [3]

- v) Based on your calculations in part (iii), comment on whether we are giving more emphasis on the direct data or on the collateral data while calculating risk premiums. [1]

- vi) Instead of estimating the runs scored using the assumptions of EBCT Model 2, it is decided to estimate the runs to be scored using the assumptions of Empirical Bayes Credibility Theory Model 1 (EBCT Model 1). By making appropriate changes in the code given earlier or otherwise, determine the credibility factor under the assumptions of EBCT Model 1. [4]

- vii) Hence estimate the number of runs to be scored for Rohit, Virat and Shivam during the fifth season under the assumptions of EBCT Model 1. [3]

- viii) Briefly comment on the differences in the answers obtained in parts (iv) and (vii). Also, comment on which model is the most appropriate in this case. [3]

[20 marks]

Q.2) Researchers are studying ectoplasmic density in the brain – a substance believed to influence cognitive resonance.

Following data is available relating to Brain Resonance Index (BRI) and Ectoplasmic Load (EL) measured in grams for 10 individuals:

Individual	Brain Resonance Index	Ectoplasmic Load
	(X)	(Y)
A	169.6	71.2
B	166.8	58.2
C	157.1	56
D	181.1	64.5
E	158.4	53
F	165.6	52.4
G	166.7	56.8
H	156.5	49.2
I	168.1	55.6
J	165.3	77.8

Use the following R code to load the data in R.

```
BRI <- c(169.6,166.8,157.1,181.1,158.4,165.6,166.7,156.5,168.1,165.3)
```

```
EL <- c(71.2,58.2,56.0,64.5,53.0,52.4,56.8,49.2,55.6,77.8)
```

Note: You can directly copy these code lines from Codes.docx file in your R Script.

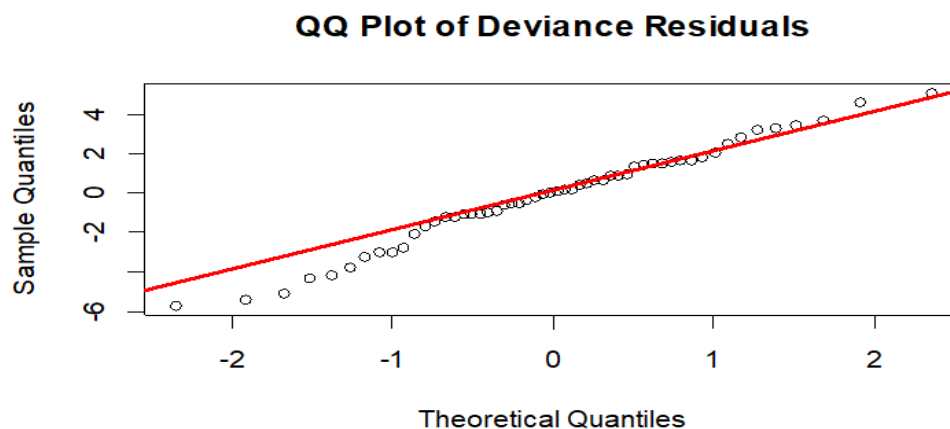
- i) Draw a labelled scatterplot of ectoplasmic load on the vertical axis versus brain resonance index on the horizontal axis. Comment on the scatterplot. [3]
 - ii) Determine the sample correlation coefficient using the Karl Pearson's Method. [2]
 - iii) Fit a regression of Y on X. Print the summary of the regression model in R. Also state the estimated standard errors for slope and intercept parameters. [4]
 - iv) Draw the fitted line on your scatterplot. [2]
 - v) Predict the ectoplasmic load of an individual whose brain resonance index is 160.4. [2]
 - vi) Estimate the covariance between the slope and intercept parameter of the regression model.
Hint: [Use vcov function in R.] [2]
 - vii) Some researchers still believe that ectoplasmic load does not depend on brain resonance index. By calculating a two-sided 95% confidence interval for the slope parameter of the regression line, test this claim of the researchers. You are required to clearly state the null and alternative hypothesis, the confidence interval and the conclusion of the test. [4]
 - viii) Some others believe that instead of testing for the slope coefficient, one should test whether there exists any non-zero correlation between the two variables. Perform the test at 5% level of significance. Clearly state the null and alternative hypotheses, p-value and the conclusion of the test. [4]
 - ix) "If a particular dataset passes the test in part (vii), then passing the test in part (viii) is a foregone conclusion." Briefly comment on the statement giving suitable reasons to justify your answer. [2]
- [25 marks]**

Q.3) An insurance company wants to analyse its motor insurance portfolio over a period of 7 years. For each policyholder, the number of insurance claims during the observation period is recorded and is available in the file **MotorClaims.csv**.

The drivers in the data are classified according to the following risk factors:

- Sex - of the driver (1= Male, 2=Female)
- Region – of residence (1 = Urban, 2 = Semi-Urban, 3 = Rural))
- Car_type – type of car insured (1 = Hatchback, 2 = Sedan, 3 = SUV)

- Job – exposure to driving based on job (1= Low, 2 = Medium, 3= High)
- i) Load the data and using lapply() or otherwise convert the risk factors into factors in R. [2]
 - ii) Fit a Generalised Linear Model (GLM) ‘model_1’ to the data assuming a Poisson distribution with a log link function, using claims as the response variable and sex, region, job, and car_type as explanatory variables, without interaction terms. Print the summary of this model in R. [4]
 - iii) Comment on the significance of each risk factor. [2]
 - iv) Estimate the expected number of claims for a male policy holder driving a Sedan in Urban area with high exposure to driving. [3]
 - v) Due to regulatory reasons, sex cannot be used as a risk factor. Refit a new model ‘model_2’ removing sex as risk factor and print the summary of this model. [3]
 - vi) State with reason which risk factor(s) can be further removed from the model_2. [1]
 - vii) Fit a new model ‘model_3’ removing the risk factor(s) identified in part (vi) and print the summary of the new model. [3]
 - viii) The analyst now wishes to investigate whether there is any interaction between the remaining significant risk factors.
Refit model_3 by including interaction between the significant risk factors. Label this as model_4. Print the summary of this model. [3]
 - ix) Following is a QQ-plot of the deviance residuals for model_4.



- Briefly comment on the above plot in the context of the goodness of fit of model_4. [2]
- x) Using Akaike Information Criterion (AIC), comment on which model is better fit among the four models. [2]

[25 marks]

Q.4) The number of car accidents at a fixed point on a road were recorded for 20 consecutive months. The results are as shown below, with x_i being the number of accidents at the fixed point in the i^{th} month

Note: You can directly copy these code lines from Codes.docx file in your R Script.

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
No of accidents	2	2	1	1	0	4	2	1	2	1	1	1	3	1	2	2	3	2	3	4

```
x <- c(2,2,1,1,0,4,2,1,2,1,1,1,3,1,2,2,3,2,3,4)
```

X is assumed to follow a Poisson distribution with parameter λ with probability function given as –

$$P(X = x) = \lambda^x * \exp(-\lambda) / x!$$

i) Calculate the sample mean and sample variance for the number of car accidents. [2]

ii) Create and store a sequence of 1000 values for the population parameter λ defining it as vector lambda assuming plausible values between 1 and 3.

You are NOT required to print these values. [1]

iii) Likelihood function for the Poisson distribution can be coded in R using the following code:

```
lik <- (lambda^(sum(x))*exp(-length(x)*lambda))/prod(factorial(x))
```

Note: You can directly copy these code lines from Codes.docx file in your R Script.

Calculate the maximum value of likelihood function based on the values of λ generated in part

(ii) and the value of λ corresponding to it. [2]

iv) Calculate the maximum value of the log likelihood function based on the values of λ generated in part (ii) and the value of λ corresponding to it. [3]

v) By first principles, state what will be the maximum likelihood estimate (MLE) of the Poisson parameter λ . Assume that the second order derivative is less than zero thus leading to maxima. [2]

vi) Plot suitable graph for the likelihood function against plausible range of λ . Also, add a line for MLE in the above graph. [2]

vii) Based on the plot in part (vi) and based on the maximum value calculated in part (iii), show that the maximum likelihood estimate of λ is the point where the likelihood function is maximized. [2]

viii) Cramér–Rao Lower Bound (CRLB) for the parameter λ of this distribution is given as λ_{MLE} / n .

Calculate an approximate 95% confidence interval for λ using the normal distribution with mean λ_{MLE} and variance given by the CRLB. [2]

ix) Calculate an **exact** two-sided 95% confidence interval for the Poisson parameter λ and compare with the confidence interval calculated in part (viii). [3]

x) Using the estimated value of λ from part (v), calculate the probability that:

- a) No accident occurs in a month
- b) More than 5 accidents happen in a month
- c) Exactly 7 accidents happen in a month

[3]

xi) Using the estimated value of the parameter λ from part (v), generate 1,000 random samples containing 100 values each from a Poisson distribution.

In order to initialise the random number generator, use `set.seed(2026)`.

You are NOT required to display the output.

[4]

xii) Calculate the values of $E(\bar{X})$ and $E(S^2)$ based on the random samples generated in part (xi).

Hence, show that $\bar{X}$ and S^2 are unbiased estimators of the population mean and population variance respectively.

[4]

[30 marks]
