

INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

November 2025

**Subject CS2 - Risk Modelling and Survival
Analysis (Paper B)**

Time allowed: 1 Hour 45 Minutes

Total Marks: 100

- Q. 1)** A small general insurance company sells three different types of policies *viz.* motor insurance, property insurance and health insurance.

The claims received by the company over the investigation period from 1st November 2014 to 31st July 2025 have been given in the data file “Q1_Data.csv”.

- i) Import the library `dplyr` and write R code to load the dataset, assign it to a data frame named ‘claims’ and format the dates in the date format “DD-MM-YYYY”. Print the first 6 rows using the `head` function. (2)

- ii) Perform the following basic operations on the data:

- Check the data for missing values;
- Check the data for duplicate entries and remove the duplicate records from the data;
- Drop those records where the date of the claim is beyond the investigation period (i.e. before 1st November 2014 or after 31st July 2025). (5)

The company is investigating modelling its future claim numbers per annum as a Poisson process with parameter λ .

- iii) Re-arrange the ‘claims’ table in the ascending order of date of claim and display the first six rows using the `head` function. (1)

Observed waiting times (in years) for all the claims are to be calculated. Following function has been written in R for calculating the observed waiting time (in years):

```
add_waiting_period <- function (data) {
  temp <- cbind(data, c(NA, diff(as.numeric(data[,2])))/365.25)
  colnames(temp)[5] <- "waiting_Period"
  temp
}
```

- iv) Explain the above code lines in brief. (2)

- v) Using the function `add_waiting_period` as written above, determine observed waiting time (in years) and add it as a fifth column in the ‘claims’ table. Display the first six rows using the `head` function. Also determine the Poisson parameter for all the claims defining it as `Lambda_all`. (2)

- vi) Create a separate data frame for claims of ‘property’ line of business showing the date and size of each claim. Show the first six rows of property table using the `head` function. Also print the number of claims under the property line of business. (3)

- vii) For the ‘property’ table, rearrange the observations in the ascending order of the date of claim, show the first six rows using the `head` function, calculate the average waiting time (in years) using the `add_waiting_period` function and determine the Poisson parameter for property claims defining it as `Lambda_prop`. (2)

- viii) Using the Poisson parameter for all claims calculated in part (v) i.e. `Lambda_all` and using the proportion of property insurance claims to total insurance claims, determine the Poisson parameter for property claims using Poisson thinning. Define this as `Lambda_prop_thinning`. Also, compare the value of `Lambda_prop` from part (vii) with `Lambda_prop_thinning`. (2)

The claim amounts of each individual claim under all the three lines of business are assumed to be independent of each other and to follow a gamma distribution.

The company decides to model the total future claim amount for each type of policy as a compound Poisson process, and uses the method of moments to fit the parameters of the claim size distribution:

$$S_t = \sum_{j=0}^{N_t} X_j$$

Where $X_j \sim \text{Gamma}(\alpha, \beta)$.

ix) Calculate the sample mean and sample standard deviation for the claim amounts for “property insurance claims”. (2)

x) Estimate the values of the parameters alpha and beta for the Gamma distribution using the method of moments for “property insurance claims”. (2)

xi) Using the value of Poisson parameter determined in part (vii) and using the sample mean and sample standard deviations determined in part (ix), determine the mean and standard deviation of the total annual claim amounts for “property insurance claims”. (4)

Consider the Gamma distribution with parameter alpha and beta as determined in part (x). We want to check whether this distribution has a heavy tail.

xii) Calculate the mean residual life for this gamma distribution for $x = 500, 1000, 1500, \dots, 50000$ and define it as MRL_X . Plot a graph showing the curve of MRL_X plotted against the values of x and comment on whether the gamma distribution has a heavy tail. (6)

[33]

Q. 2) A friend of yours believes that the film industry is highly profitable and encourages you to invest in film production. To verify his claim, you decide to analyse the past data to assess the likelihood of a film becoming a box-office success. The file “Q2_Data.csv” contains information on the number of movies produced and the number of hits for each year.

i) Write R code to load the dataset, assign it to a data frame named ‘Movie_data’, and display the first five rows and last five rows in this dataset. (3)

ii) Add a new column to ‘Movie_data’ called ‘Success_rate’, consisting of the annual rates of success and display a scatterplot of the rates over time. (3)

iii) Comment on the scatterplot in part (ii). (1)

Based on the data, you decide to predict what the success rate would be in 2025 and 2026.

Let “Year” be represented by the random variable X (explanatory variable) and let “Success_rate” be represented by the random variable Y (response variable).

Following is the basic structure of the regression models which can be fitted to the data:

$$Y = \alpha + \sum_{i=1}^n \beta_i * X^i$$

Where 'n' can take the values 1, 2 or 3.

- iv) Fit a linear regression model "Model_1" with $n = 1$ and predict the success rates for 2025 and 2026. (2)
- v) Fit a quadratic polynomial regression model "Model_2" with $n=2$ and predict the success rates in 2025 and 2026. (4)
- vi) Plot the curves fitted in parts (iv) and (v) on top of the scatterplot plotted in part (iii). (2)
- vii) Comment on the suitability of the models 1 and 2. (2)

Your friend suggested that the data could be split into disjoint segments and separate curves can be fitted for each segment. You are considering splitting the data into the following three segments:

- Segment 1: 2000 to 2008,
- Segment 2: 2009 to 2016,
- Segment 3: 2017 to 2024.

- viii) Split the data frame 'Movie_data' into three new data frames 'Movie_data_segment1', 'Movie_data_segment2' and 'Movie_data_segment3'. (3)
- ix) Fit six different regression models as follows (print the models):

Segment	Linear Regression Model	Quadratic Polynomial Regression Model
Segment 1	"Model S11" with $n=1$	"Model S12" with $n=2$
Segment 2	"Model S21" with $n=1$	"Model S22" with $n=2$
Segment 3	"Model S31" with $n=1$	"Model S32" with $n=2$

(6)

- x) For each of these models print the BIC i.e. Bayesian Information Criterion.

Hint: [Use `BIC` function in R.]

Also determine which model is the best model based on BIC for each of the segments. (5)

- xi) Using the model identified as best for Segment 3 in part (x), predict the success rates for 2025 and 2026. Comment on the predictions. (3)

[34]

Q. 3) The time series 'mdeaths' and 'fdeaths' are available in R. These time series relate to the monthly deaths from bronchitis, emphysema and asthma in the UK, 1974–1979, for males (mdeaths) and females (fdeaths).

- i) You are initially required to –
 - a) Print both the time series.
 - b) Print the times of first and last observation of both the time series (3)

- ii) Plot the sample ACF and sample PACF for both the datasets, up to lag $k = 60$ months. Your graphs specific to a time series should be side by side, and should have appropriate labels. (5)
- iii) Plot the sample ACF and sample PACF for both the datasets after removing seasonality by seasonal differencing, up to lag $k = 60$ months. Your graphs specific to a time series should be side by side, and should have appropriate labels. (4)
- iv) Comment on the graphs as determined in parts (ii) and (iii). (2)
- v) Write a for loop to find AIC (Akaike's Information Criterion) for all ARIMA model fits for `fdeaths` with $p = 0, 1$, $q = 0, 1$ and $d = 0, 1$.

Seasonality is required to be adjusted and hence, set the 'seasonal' argument of the `arma` function to `'list (order = c(0,1,1), period = 12)'`

Show that the model with least AIC is ARIMA (0, 1, 1) for the `fdeaths` data set. (6)

- vi) Fit a ARIMA (0, 1, 1) model to both the datasets.

Seasonality is required to be adjusted and hence, set the 'seasonal' argument of the `arma` function to `'list(order = c(0,1,1), period = 12)'`. (2)

- vii) Perform the Ljung and Box portmanteau test on both the ARIMA models in part (vi).

Set the arguments `lag` to 14 and `fitdf` to 2.

Clearly state the null and alternative hypothesis, p-values and the outcome of the tests in respect of both the models at 5% level of significance. (5)

- viii) Predict the death counts of each gender for the 12 months in the next year 1980 using the above model fitted in part (vi). (2)

- ix) These two time series `mdeaths` and `fdeaths` are to be co-integrated using the cointegrating vector (0.90, 1.09) i.e. $X = 0.9 * mdeaths + 1.09 * fdeaths$. Perform the Phillips-Perron Unit Test in R using the `PP.test` function and check whether the cointegrated time series `X` with the vector (0.90, 1.09) is stationary or not. Clearly state the null hypothesis, alternative hypothesis, p-value and the outcome of the test at 5% level of significance.

Hint: [Phillips-Perron Unit Test tests whether the co-integrated time series can be differenced. If p-value is less than 5%, we conclude that the co-integrated time series cannot be differenced.] (4)

[33]
