PUSASQF206
Application of IT – Basics of R

Time: 2 hours
Total Marks: 60 marks

Note:

1) **The candidate has option to either attempt question 3A or question 3B. Rest all questions are mandatory.**
2) **Numbers to the right indicate full marks.**

Q1. **15 Marks**

Q1A. **5 Marks**

A random variable N is assumed to follow a Type 1 Negative Binomial distribution with k = 4 and p = 0.3

(a) Simulate 100 values from the above distribution , using the seed as the month of your birthdate. Show the empirical inter-quartile range. (2)

(b) Compare your answers to the population inter – quartile range. (2)

(c) Without any further work, how would answer in (a) change if number of simulated values increased. (1)

Q1B. **5 Marks**

Basket is a new age FinTech company allowing users to pay their credit card bills. In a bid to improve their recognition among the customers, the company ran a massive marketing company in the recently concluded Cricket league in India.

Prior to the campaign, a survey conducted on a random sample of 200 consumers , it was found that only 29 had heard of Basket.

Post the campaign, a similar survey found that 31 out of a random sample of 180 consumers were aware about the company.

(a) Carry out Fisher's exact test in order to ascertain if the marketing campaign was successful. *Your output should include an appropriate 99% confidence interval of the odds ratio*. (4)

(b) Comment on the effectiveness of the campaign using the test in (i) (1)

Q1C. **5 Marks**

From the package 'datasets' , procure the 'Orange' database

(a) Fit an intercept only linear model of Circumference (in mm) called *model0* (1.5)

(b) Update the model in part (i) , to use Age and Age^2 as explanatory variables. (1.5)

(c) Explain why R^2 is not an appropriate statistic to compare the above models. Use an appropriate statistic to compare the above models. (2)

Q2 **15 Marks**

Q2A. **5 Marks**

Actuary working in the field of climate change is working to model the rate at which massive storms happen in the world. He has collected data for the last 18 years and the numbers are as follows :

9, 6, 6, 3, 2, 5, 4, 6, 2, 1, 4, 6, 3, 7, 8, 10, 9, 7

Store the above in a vector called `storm`

(a) He has asked you to use R and calculate the exact 99% confidence interval for parameter λ , assuming the number of storms occur at a constant rate of λ per year. (3)

(b) Compare the exact confidence interval in part (i) with the approximate confidence interval given by :

```
mean(storm) +
c(-1,1)*qnorm(0.005,lower = F)*sqrt(mean(storm)/length(storm))
```
(2)

Q2B. **5 Marks**

From the package 'datasets' , procure the 'ChickWeight' database

A scientist has recently concluded an investigation into Chicks and is currently asking for your help in analysing the importance of Age and Diet in predicting the weight of a Chick. He has asked you to assume that weights are always positive and follow a Gamma distribution while modelling.

(a) Create a GLM model of weight with Time and Diet as the explanatory variables along with appropriate link function. (2)

He has now asked you to consider interaction effect between Time and Diet.

(b) Update the model in (a) , to include the interaction effect. (1)

(c) Conduct an appropriate test in R , in order to ascertain the significance of the interaction term in the model. (2)

Q2C. **[5 Marks]**

StarTree Insurance Co. Receives claims daily. The number of claims received for the past 365 days are as follows:

Claims	0	1	3	4	5	8
Freq	28	126	111	31	30	39

(a) Calculate average number of claims per day. (1)

(b) A new analyst suggests that the claims received daily follow a Poission distribution with λ.

Use chisq.test() to test the Goodness-of-Fit of the distribution suggested by the new analyst. (4)

Q3. Attempt any one from 3A and 3B. **30 Marks**

Q3A. **30 Marks**

Use the file ***PolicyData.csv*** file for this question.

Policy and claim information of 650 policies is provided to you, the data contains following fields :

Policy : Policy Number
Claim : Number of claims corresponding to each policy
Cust_Exp : Policyholder's final review of the company at the end of the policy tenure. *(VS = Very satisfied, SA= Satisfied, DS= Disappointed, VD= Very Disappointed)*
Amount : Claim Amount per policy. (*0 if no claim is reported*)

(a) Create a frequency table of number of claims and share how many policies made a claim. (2)

(b) Plot a histogram of ***Claim*** and suggest 2 distributions that can be a good fit for the random variable. (2)

(c) Given that claim count follows Poisson distribution with following two possible values for Poisson parameter :

- 0.35 and
- 0.30

Compute confidence interval at 95% confidence level to assess which value is more suitable for the given data. (3)

(d) Compute mean , variance and median of log of claim amount , by naming it *log_claim*. Make sure to exclude policies with no claim. (4)

(e) Obtain histogram and Normal QQ Plot of log amount. Add a line to the QQ Plot for Normal Distribution. (4)

(f) Indicate which distribution the claim amount might be following using evidence from (d) and (e) . (3)

(g) Assuming the log amount (i.e. log of claim amount) follows a Normal distribution , test if mean of log amount is greater than 10 at 90% level of confidence. State the hypothesis and conclusion clearly. (4)

(h) Assess whether the policyholder experience (i.e. Cust_Exp) changes with more number of claims. Create contingency table and perform test to check the above assertion. State the hypothesis clearly. (3)

(i) Amount is defined to be large if the amount is greater than 100,000. Calculate 95% confidence interval for proportion of large claim, and comment on the likelihood if more than 25% of claims are large. (5)

OR

Q3B. **30 Marks**

Refer to the data file *Indices_Returns.csv* and answer the following questions :

(a) Load the csv file into R and create a new column called ***Return_Direction*** . The value of this column will be "Positive" when the Sensex returns are positive and "Negative" when they are negative and convert the variable as a factor variable. (3)

(b) Fit an appropriate generalized linear model (GLM) to with a logit link function to relate the "Return_Direction" with the returns of 10 sectors as a multivariate model and display the summary of the model. (6)

(c) Identify which sectors have significantly impacted the direction of Index returns at 95% and 99% confidence level. (4)

(d) Verify the relationship between residual deviance of the model and the Akaike Information Criteria (AIC). (3)

(e) Plot the residuals of the fitted model and identify which month is the most significant outlier in the residuals. (4)

(f) Comment on the appropriateness of the model fitted. (2)
Your actuarial friend has suggested that the current model can be improved by removing the variables which do not impact the direction of index returns at 95% confidence level and refitting the GLM with 'logit' link function.

(g) Update the model fitted in (b) above, as suggested by your friend and display the summary of the model. (4)

(h) Compare the models in (b) and (g) using an appropriate test and comment on the difference in the residual deviances between the two models. (4)