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. Attempt All Questions

A) **5 Marks**

A random variable X follows a lognormal distribution with $\mu=10$ and $\sigma=2$.

i. Set the seed as 2021 and simulate 1000 standard normally distributed random values.

ii. Apply a suitable transformation to the simulated values above to obtain 1000 random values with the lognormal distribution (as X).

iii. Calculate the empirical mean and median of the lognormally distributed simulated values, and comment on them.

B) **5 Marks**

A life insurer has operations of similar size in three different regions (X, Y and Z). Last year, the region X, Y and Z saw 255, 296 and 322 claims losses respectively).

i. Given that a Poisson distribution is reasonable for the number of claims, test the hypothesis (separately for each region) that the average number of losses per month is 25.

ii. What number can replace 25, for the hypothesis to stand (not be rejected) for all the regions?

C) **5 Marks**

A die is rolled 60 times and yields 11 ones, 14 twos, 11 threes, 8 fours, 9 fives and 7 sixes.

i. Test the hypothesis that the die is fair using the chi-squared goodness of fit test.

ii. How would your answer differ if the die were rolled 300 times and still landed in the same proportion as above? Repeat the same hypothesis testing as above and comment.

Q2. Attempt All Questions

A) **5 Marks**

From the package 'datasets', procure the 'sleep' database, which shows the increase in hours of sleep (column named 'extra') under the influence of two soporific drugs (group column specifies the drug) to the 10 different persons (ID column being their unique identifier). Determine whether one of the drugs is more effective than the other.

B) **5 Marks**

From the package 'datasets', procure the 'quakes' database.

i. Fit a linear model of mag (magnitude of earthquake) over stations (number of stations reporting the earthquake).

ii. Also compute the correlation between mag and stations and ascertain that it is consistent with the R-squared metric reported by the linear model fitting.

iii. Finally, by way of plotting residuals or otherwise, visually ascertain if this is a fit case for a linear model

C) **5 Marks**

From the package 'datasets', procure the 'USArrests' database, which contains the number of arrests for murder, assault and rape per 100000 residents of the state. (There is also a column for % of urban population, but we don't need it for now.)

i. First verify that there is a moderate to high pairwise correlation between all the crime rates.

ii. Next, compute the principal components (PC).

iii. How much variance is captured by the first PC? How many PCs are required to be able to capture more than 90% variance?

Q3. Attempt Q.3A or 3B

A) **30 Marks**

Puromycin or no puromycin?
From the package 'datasets', procure the 'Puromycin' database. Please cleanse the data before use, if necessary.
In an enzymatic reaction, the higher is the substrate concentration (conc), the higher will be the reaction velocity (rate). In the dataset, there is an additional column (state) which specifies whether or not the reaction is additionally treated with puromycin.

i. Compute the Pearson correlation between conc and rate. Also compute the Spearman correlation between them. What does this difference suggest?

ii. Also plot the conc and the rate as a scatter plot. What are your observations?

iii. Fit a linear model first for rate over conc, and comment on the fit. How does the fit improve when the state is added as a factor?

iv. Analyze the residuals in the (better of the two) model in part (iii).

v. Based on your observations in part i, ii, and iv, transform the conc variable suitably and then fit a linear model (including or excluding state as appropriate), such that you get a better fit.

vi. Would you say that the puromycin treatment affects the reaction rate significantly?

vii. Can the model be improved by adding an interaction term between conc (or its transformed version) and state?

viii. An alternative approach would be to fit a GLM instead of a simple linear model. Fit an appropriate GLM and compare the models / conclusions.

B) **30 Marks**

Economy and Employment

From the package 'datasets', procure the 'longley' database. Please cleanse the data before use, if necessary. The data contains the following columns for the period from 1947 to 1962:

GNP.deflator: GNP implicit price deflator
GNP: Gross National Product
Unemployed: number of unemployed
Armed.Forces: number of population in the armed forces
Population: 'non-institutionalized' population above 14 years of age
Year: year of data
Employed: number of people employed

i. Compute the correlation matrix and state your observations.

ii. Also plot the columns of the dataset on a suitable chart and comment on the same.

iii. Fit a linear model for Employed over all the other variates. How do you view the quality of fit? How is its predictive power likely to be?

iv. Then try to select the appropriate explanatory variates by backward selection approach.

v. Then try to select the appropriate explanatory variates by forward selection approach. How does your model compare with the previous one? Comment on issues if any.

vi. Your friend suggests that Principal Component Analysis (PCA) may be helpful to tackle the issues faced in this modelling. Investigate using PCA if some of these issues can be resolved. Are there still any other issues?

vii. Another potential technique is to work with normalized data, which means the raw (absolute) numbers like Employed (or Unemployed and other such variates) can be divided by the total population for the respective year. Will that approach yield a significant modelling improvement?