Statistical Modelling in 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 in the right indicate full marks.
3. The candidate will be provided with the formula sheet and graph papers [if required] for the examination.
4. Use of approved scientific calculator is allowed.

Q1 A.

Use the file iris.csv and load it into R as data.

1. Name the species that has highest variance in Sepal.Length [1]
2. Add a new column to data named value such that the output is 1 if the variety is Setosa and 0 otherwise. [1]
3. Create a logistic regression model in R with value as predicted variable and other [2]

Columns as independent variables excluding variety.

1. Comment on values of the co-efficients [1]

Q1 B.

Use the freeny dataset. For loading the data use,

require(stats)

require(graphics)

1. Store the data in freeny\_data variable. Show first few rows in the data. [1]
2. Summarise the freeny\_data [1]
3. Create a linear regression model with variable y against other variables [2]
4. What is the value for R2? [1]

Q1 C. Create a Data Frame named COVID\_TESTS in R.

|  |  |  |  |
| --- | --- | --- | --- |
| Number | State | Active\_Cases | Positive |
| 1 | Maharashtra | 2000 | 36 |
| 2 | Karnataka | 1278 | 45 |
| 3 | Gujarat | 694 | 100 |
| 4 | Delhi | 235 | 150 |
| 5 | Madhya Pradesh | 548 | 65 |
| 6 | Kerala | 1836 | 87 |

1. Show the last 2 records in the data frame. [2]
2. While updating the data for number 4, it was observed that the Positive

was incorrectly noted. The correct number is 175. Make the necessary

changes. [2]

1. Plot a bar graph for Active Cases [1]

Q2 A.

1. Load data mtcars and display first few rows of the dataset. [1]
2. Create a variable input and store the columns am, cyl, hp and wt [1]
3. Create histograms for all the above mentioned (II) columns [1]
4. Create a logistic regression with am as target variable and cyl, hp

and wt as predictor variables. Provide the value of intercept. [2]

Q2 B. As a data scientist a confusion matrix was created by you using the predict() function.

The confusion matrix is given below:

|  |  |  |
| --- | --- | --- |
|  | 1(Predicted) | 0 |
| 1 (Actual) | 786 | 96 |
| 0 | 89 | 84 |

1. Create the above mentioned matrix in R [2]
2. Calculate Precision of the model [1]
3. Calculate Sensitivity of the model [1]
4. A competitive company provided the same model mentioning the precision

98% and sensitivity as 70%. Which model is better

1. Using precision as comparison
2. Using sensitivity as comparison [1]

Q2 C. Use the airquality dataset from R. Load library rpart.

1. Store the data in AQ and show first few rows. [1]
2. Examine the data AQ using str function. [1]
3. Use set.seed(888) and check if there are any na values in the Ozone column

Incase if there are NA values, run the below command:

AQ[is.na(AQ$Ozone),1]<-sample(AQ[!is.na(AQ$Ozone),1],37) [1]

1. Run the rpart model on Ozone as target variable and rest columns as

Independent variables. Print the summary of the model [1]

1. Print the model [1]

Q3 A.

CASE STUDY: Predicting the diabetes.

As a data scientist, a medical organisation has hired you. They want to understand the

significance of various factors of the diabetes. You are provided with the diabetes.csv dataset.

Perform the following steps:

1. Install/Load the libraries plotly, knitr, dplyr [1]
2. Read the dataset diabetes.csv and store it in db, display first few rows [1]
3. Check the structure of the dataset [1]
4. The column Outcome in the dataset is a categorical column, in order for the

data to be consistent, it needs to be standardised using as.factor [2]

1. Create a train-test split ratio with train test ration to be 80-20 [5]
2. Create a model to perform logistic regression with

Family = binomial(link=’logit’) [7]

1. Summarise the model and mention the AIC value [3]
2. Create a prediction model on the test data [5]
3. Create a train-test split with train test ration to be 60-40 and compare the

model with the above model using AIC value. [5]

Q3 B

Import the dataset Fish.csv, which contains the length, weight, width and height of a fish along

with the Species of the Fish. Import the file into the R session

I. Create a table such as below:

|  |  |  |  |
| --- | --- | --- | --- |
| Species | Count | Average Height | Average Weight |
|  |  |  |  |

[3]

II. Create a correlation matrix between the variables, leaving the species variable. Comment

on the relationship between the variables. [3]

III. Create a Linear Regression model for the Weight of the Fish given the lengths, height and

width. [10]

IV. State the assumptions on errors and create plots in R to test your assumptions and

comment on each plot [8]

V. Create a new model by removing all the variables that do not have a significant effect on

Weight at 10% level. [4]

VI. Create a graph of fitted values using the model in part E against the actual values.

Comment on your results. [2]