SAVEETHA SCHOOL OF ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

ITA 0443 - STATISTICS WITH R PROGRAMMING FOR REAL TIME PROBLEM

DAY 4– LAB MANUAL Part 2

Reg No:192124100

Name:X.M.Mary Sushmija

LOGISTIC REGRESSION ANALYSIS IN R

Exercise

5. Create a logistic regression model using the “mtcars” data set with the information given below.

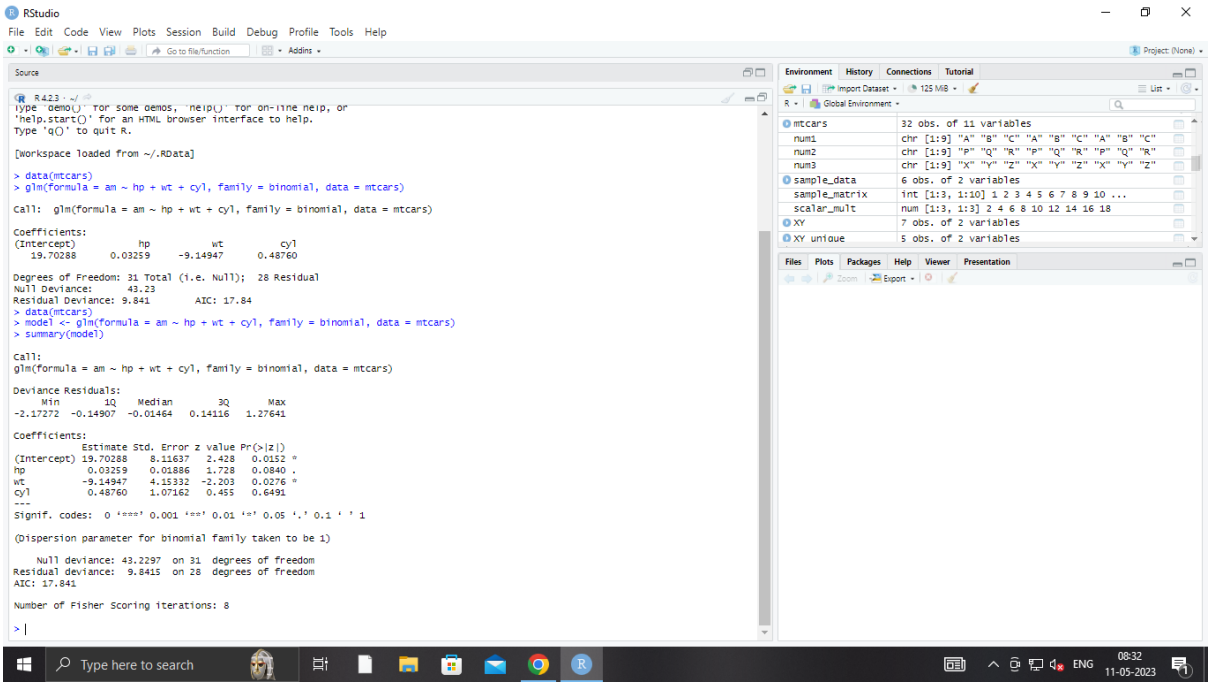
The in-built data set "mtcars" describes different models of a car with their various engine specifications. In "mtcars" data set, the transmission mode (automatic or manual) is described by the column am which is a binary value (0 or 1). Create a logistic regression model between the columns "am" and 3 other columns - hp, wt and cyl.

PROGRAM:

```
> data(mtcars)
> glm(formula = am ~ hp + wt + cyl, family = binomial, data = mtcars)

Call:  glm(formula = am ~ hp + wt + cyl, family = binomial, data = mtcars)

Coefficients:
(Intercept)           hp           wt          cyl
   19.70288      0.03259     -9.14947      0.48760

Degrees of Freedom: 31 Total (i.e. Null);  28 Residual
Null Deviance:         43.23
Residual Deviance: 9.841       AIC: 17.84
> data(mtcars)
> model <- glm(formula = am ~ hp + wt + cyl, family = binomial, data =
mtcars)
```

```
> summary(model)

Call:
glm(formula = am ~ hp + wt + cyl, family = binomial, data = mtcars)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.17272  -0.14907  -0.01464   0.14116   1.27641

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 19.70288    8.11637   2.428   0.0152 *
hp           0.03259    0.01886   1.728   0.0840 .
wt          -9.14947    4.15332  -2.203   0.0276 *
cyl          0.48760    1.07162   0.455   0.6491
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.2297  on 31  degrees of freedom
Residual deviance:  9.8415  on 28  degrees of freedom
AIC: 17.841
```

```
Number of Fisher Scoring iterations: 8
```



POISSON REGRESSION ANALYSIS IN R

Exercise :

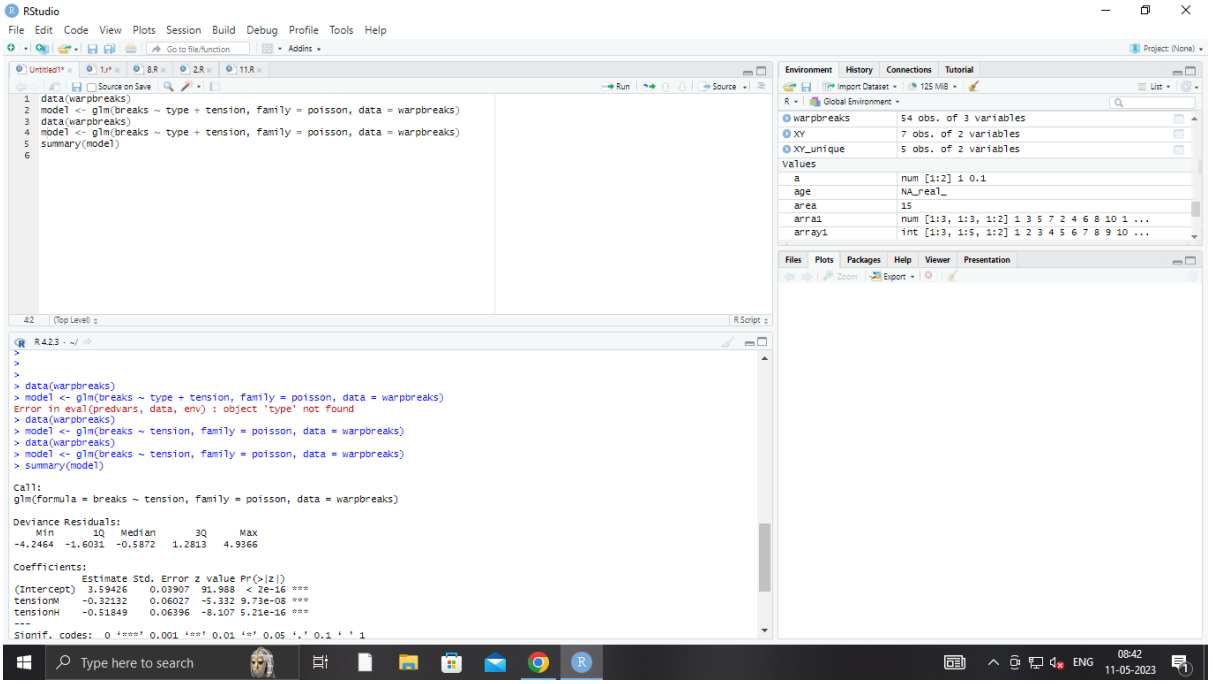
6. Create a Poisson regression model using the in-built data set “warpbreaks” with information given below.

In-built data set "warpbreaks” describes the effect of wool type (A or B) and tension (low, medium or high) on the number of warp breaks per loom. Consider "breaks" as the response variable which is a count of number of breaks. The wool "type" and "tension" are taken as predictor variables.

PROGRAM:

```
> data(warpbreaks)

> model <- glm(breaks ~ tension, family = poisson, data = warpbreaks)

> data(warpbreaks)

> model <- glm(breaks ~ tension, family = poisson, data = warpbreaks)

> summary(model)
```

```
Call:

glm(formula = breaks ~ tension, family = poisson, data = warpbreaks)


Deviance Residuals:

    Min       1Q   Median       3Q      Max

-4.2464  -1.6031  -0.5872   1.2813   4.9366


Coefficients:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)  3.59426    0.03907  91.988  < 2e-16 ***

tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***

tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for poisson family taken to be 1)


    Null deviance: 297.37  on 53  degrees of freedom

Residual deviance: 226.43  on 51  degrees of freedom

AIC: 507.09


Number of Fisher Scoring iterations: 4
```



1.Randomly Sample the iris dataset such as 80% data for training and 20% for test and create Logistics regression with train data, use species as target and petals width and length as feature variables , Predict the probability of the model using test data, Create Confusion matrix for above test model

2. (i)Write suitable R code to compute the mean, median ,mode of the following values
c(90, 50, 70, 80, 70, 60, 20, 30, 80, 90, 20)
(ii) Write R code to find 2nd highest and 3rd Lowest value of above problem.

3. Explore the airquality dataset. It contains daily air quality measurements from New York during a period of five months:
• Ozone: mean ozone concentration (ppb), • Solar.R: solar radiation (Langley),
• Wind: average wind speed (mph), • Temp: maximum daily temperature in degrees Fahrenheit,
• Month: numeric month (May=5, June=6, and so on),• Day: numeric day of the month (1 -4).
i. Compute the mean temperature(don't use build in function)
ii.Extract the first five rows from airquality.
iii.Extract all columns from airquality except Temp and Wind
iv.Which was the coldest day during the period?
v.How many days was the wind speed greater than 17 mph?

4. (i)Get the Summary Statistics of air quality dataset
(ii)Melt airquality data set and display as a long – format data?
(iii)Melt airquality data and specify month and day to be "ID variables"?
(iv)Cast the molten airquality data set with respect to month and date features
(v) Use cast function appropriately and compute the average of Ozone, Solar.R , Wind and temperature per month?

5.(i) Find any missing values(na) in features and drop the missing values if its less than 10% else replace that with mean of that feature.

(ii) Apply a linear regression algorithm using Least Squares Method on “Ozone” and “Solar.R”

(iii)Plot Scatter plot between Ozone and Solar and add regression line created by above model

6. Load dataset named ChickWeight,

(i).Order the data frame, in ascending order by feature name “weight” grouped by feature

“diet” and Extract the last 6 records from order data frame.

(ii).a Perform melting function based on “Chick", "Time", "Diet" features as ID variables

b. Perform cast function to display the mean value of weight grouped by Diet

c. Perform cast function to display the mode of weight grouped by Diet

7. a. Create Box plot for “weight” grouped by “Diet”

b. Create a Histogram for “weight” features belong to Diet- 1 category

c. Create Scatter plot for “ weight” vs “Time” grouped by Diet

8. a. Create multi regression model to find a weight of the chicken , by “Time” and “Diet” as as

predictor variables

b. Predict weight for Time=10 and Diet=1

c. Find the error in model for same

9 .For this exercise, use the (built-in) dataset Titanic.

a. Draw a Bar chart to show details of “Survived” on the Titanic based on passenger Class

b. Modify the above plot based on gender of people who survived

c. Draw histogram plot to show distribution of feature “Age”

10. Explore the USArrests dataset, contains the number of arrests for murder, assault, and rape for each of the 50 states in 1973. It also contains the percentage of people in the state who live in an urban area.

(i) a. Explore the summary of Data set, like number of Features and its type. Find the number of records for each feature. Print the statistical feature of data

b. Print the state which saw the largest total number of rape

c. Print the states with the max & min crime rates for murder

(ii).a. Find the correlation among the features

b. Print the states which have assault arrests more than median of the country

c. Print the states are in the bottom 25% of murder

(iii). a. Create a histogram and density plot of murder arrests by US stat

b. Create the plot that shows the relationship between murder arrest rate and proportion of the population that is urbanised by state. Then enrich the chart by adding assault

arrest rates (by colouring the points from blue (low) to red (high)).

c. Draw a bar graph to show the murder rate for each of the 50 states .

11. a. Create a data frame based on below table.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Spends	1000	4000	5000	4500	3000	4000	9000	11000	15000	12000	7000	3000
Sales	9914	40487	54324	50044	34719	42551	94871	118914	158484	131348	78504	36284

b. Create a regression model for that data frame table to show the amount of sales(Sales) based on the how much the company spends (Spends) in advertising

c. Predict the Sales if Spend=13500

12.(i) Write a R program to extract the five of the levels of factor created from a random sample from the LETTERS (Part of the base R distribution.)

(ii)Write R function to find the range of given vector. Range=Max-Min
Sample input, C<-(9,8,7,6,5,4,3,2,1),
output=8

(iii)Wirte the R function to find the number of vowels in given string
Sample input c<- “matrix”, output<-2

13.Load inbuild dataset “ChickWeight” in R

(i) Explore the summary of Data set, like number of Features and its type. Fins the number of records for each features

(ii)Extract last 6 records of dataset

(iii) order the data frame, in ascending order by feature name “weight” grouped by feature “diet”

(iv)Perform melting function based on “Chick","Time","Diet" features as ID variables

(v)Perform cast function to display the mean value of weight grouped by Diet

14.(i)Get the Statistical Summary of “ChickWeight” dataset

(ii)Create Box plot for “weight” grouped by “Diet”

(iii)Create a Histogram for “Weight” features belong to Diet- 1 category

(iv) Create a Histogram for “Weight” features belong to Diet- 4 category

(v) Create Scatter plot for weight vs Time grouped by Diet

15.(i) Create multi regression model to find a weight of the chicken , by “Time” and “Diet” as as predictor variables

(ii) Predict weight for Time=10 and Diet=1

(iii)Find the error in model for smae