**B9DA101-Statistics for Data Analytics - Assignment 2**

**Amol Wakde – 10543430**

**Question 1**

Consider a relational dataset and specify your input and output variables, then:

* 1. Train the model using 80% of this dataset and suggest an appropriate GLM to model **output** to **input** variables.

**(10 Marks)**

* 1. Specify the significant variables on the **output** variable at the level of 𝛼=0.05 and explore the related hypotheses test. Estimate the parameters of your model.

**(10 Marks)**

* 1. Predict the output of the test dataset using the trained model. Provide the functional form of the optimal predictive model.

**(10 Marks)**

* 1. Provide the confusion matrix and obtain the probability of correctness of predictions.

**(5 Marks)**

**(Total: 35 Marks)**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Question1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Answer**: -

From the below dataset, we are using height and weight as an input variable and fisherman as a target/response variable.

Downloaded and Imported the fishermen dataset using R studio.

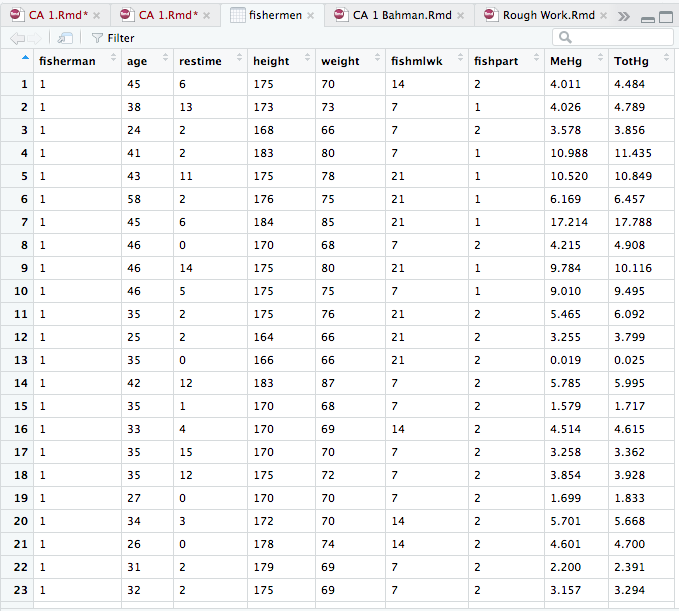
**Description**: mercury levels among fishermen and a control group of non-fishermen and various other factors related to it.

**Variables/names**

Fisherman indicator (fisherman) (Fisherman=1 and non-fisherman=0)

Height in cm (height)

Weight in kg (weight)



Since output variable fisherman is binary, we will use logistic regression (Binary).

Prepare the dataset accordingly

x1=fishermen$height #input variable 1 (independent)

x2=fishermen$weight #input variable 2 (independent)

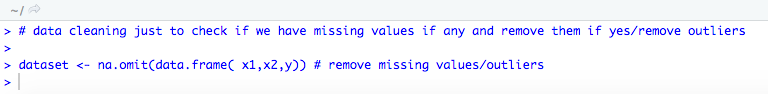
y=fishermen$fisherman #output variable (dependent- binary variable)

A picture containing bird, flower

Description automatically generated

#Data cleaning just to check if we have missing values if any and remove them if yes/remove outliers

dataset <- na.omit(data.frame(x1,x2,y)) #remove missing values/outliers if any



#split into training and test sets

set.seed(3578)

n=nrow(fishermen)

indexes = sample(n,n\*(80/100))

A screenshot of a cell phone

Description automatically generated

#separate training and test sets

trainset = dataset[indexes,]

testset = dataset[-indexes,]

A picture containing bird, flower

Description automatically generated

#fit the logistic regression

trainset.glm <- glm(y ~.,trainset, family="binomial") # ~. shows that we include all ind. variables

summary(trainset.glm)

A screenshot of a social media post

Description automatically generated

From the above output we can see that all the P-values > α=0.05 so there is no significant input variable to include.

Where,

= 5.04

= -0.01 (~ -0.0063)

= -0.04

But we will exclude all the attributes because those are not significant.

#Therefor predictive model will be as follows: -

Step1)

phat\_i = 1/ (1+exp -( + \* x\_i1 + \* x\_i2)) #Link Function – logistic/logit

phat\_i = 1/ (1+exp -(0+(-0.01) \* 0 + (-0.04) \* 0)) #excluding intercept and Xi\*

phat\_i = 1/1+ = ½ = 0.5

step2)

yhat\_i = 1 when phat\_i >= 0.5

0 when phat\_i < 0.5

So here, predicted value, yhat\_i = 1.

#predict phat

res=predict(trainset.glm,testset, type="response") # prediction

res

A screenshot of a cell phone

Description automatically generated

#convert phat to yhat

Hypothesis

H\_0 : Beta\_0 = 0

H\_1 : Beta\_0 != 0

predictedvalues=rep(0,nrow(testset))

predictedvalues[res>0.5]=1 #probability of Gender being 1, if p<0.5 then gender=0

df=data.frame(testset$y,predictedvalues )

View(df)

A picture containing dog, white, building, sitting

Description automatically generated

#confusion matrix (Contingency table of predicted Vs Actual values)

table (predictedvalues, actualvalues=testset [,3])

A picture containing bird

Description automatically generated

#accuracy

accuracy=mean(predictedvalues == testset[,3]) # correctness of prediction

accuracy

A screenshot of a cell phone

Description automatically generated

Hence the accuracy is 0.81.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Question 2\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Question 2**

Let are identically independently distributed (iid) with Poisson().

1. Compute the likelihood function (LF**). (10 Marks)**

1. Adopt the appropriate conjugate prior to the parameter (Hint: Choose hyperparameters optionally within the support of distribution). (**10 Marks)**
2. Using (a) and (b), find the posterior distribution of . **(10 Marks)**
3. Compute the minimum Bayesian risk estimator of . **(5 Marks)**

**(Total: 35 Marks)**

**Answer: -**

1. To find likelihood function-

Since, Probability mass function for poisson:

Where x=1,2,3,4…..10 and .

Since in above equation, we can write it as

----------------(a)

The parameter of interest is for Poisson distribution.

The likelihood distribution function is

=

**b)** To find Conjugate prior**-**

Since, Prior distribution:

*=*Gamma()

Since is constant in above equation.

We can write above equation as

consider α=2 and β=5

This is the conjugate prior model :-

-----------(b)

c) Using Baye’s Rule to find posterior model

|  |
| --- |
| **Posterior Model Observation Model \* Prior Model** |

\* (From (a) and (b))

Comparing above equation with

and

= and



Therefor

=Gamma()

= Gamma

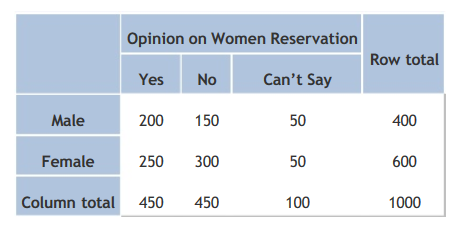
1. Baye’s estimater**-**

B = E= =

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Question3\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Question 3**

An opinion poll surveyed a simple random sample of 1000 students. Respondents were classified by gender (male or female) and by opinion (Reservation for women, No Reservation, or No Opinion). Results are shown in the observed contingency table below.



Does the gender and opinion on women reservation are independent? Use a 0.05 level of significance. To do so,

1. State the hypotheses. **(5 Marks)**
2. Find the statistic and critical values. **(10 Marks)**
3. Explain your decision and Interpret results. **(15 Marks)**

**(Total: 30 Marks)**

**Answer: -**

1. State the hypotheses.

[null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis).

**H0**: Gender and opinion on women reservation are independent.

and

an alternative hypothesis

**H1**: Gender and opinion on women reservation are not independent.

1. Find the statistical and critical values.

Given significance level is 0.05 (to do analysis). Apply a [chi-square test for independence](http://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20test%20for%20independence) by making use of sample data.

In order to analyze given sample data, we need to first find out below:

**1**.The degrees of freedom (DOF)

DOF = (r - 1) \* (c - 1) = (2 - 1) \* (3 - 1) = 2

Where,

r- no of rows (male, female =2)

c –no of columns (yes, no, can’t say = 3)

**2**.The expected frequency counts (EFC) and

EFC\_(row,col) = (nrow \* ncol) / n

For Male

EFC(1,1) = (400 \* 450) / 1000 = 180

EFC(1,2) = (400 \* 450) / 1000 = 180

EFC(1,3) =(400 \* 100) / 1000 = 40

For Female

EFC(2,1) =(600 \* 450) / 1000 = 270

EFC(2,2) =(600 \* 450) / 1000 = 270

EFC(2,3) = (600 \* 100) / 1000 = 60

Above data tabulated as follows: -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Opinion on Women Reservation** | | | **Row Total** |
| **Yes** | **No** | **Can‘t Say** |
| **Male** | 180 | 180 | 40 | 400 |
| **Female** | 270 | 270 | 60 | 600 |
| **Column Total** | 450 | 450 | 100 | 1000 |

**3**.The chi-square test statistics.

Calculate P-value, based on chi Square statistics and degree of freedom.

= Summation [ (OBS (row, col) –EXP(row, col)2] / E\_(row,col)]

= (200-180)2 / 180 + (150 -180)2 / 180 + (50-40)2 / 40 +

(250-270)2 / 270 + (300-270)2 / 270 +(50-60)2/40

= 400/180 + 900/180 +100/40 +400/270 + 900/270 +100/60

= 16.2

Chi Square critical value is 16.2 ( At 2 degrees of freedom and 0.05 significance level).

**4**.Next step is to find out the p-value using pchisq() function -

A screenshot of a cell phone

Description automatically generated

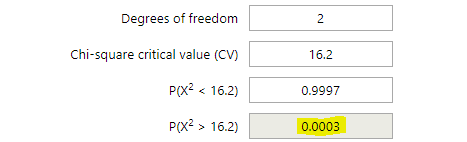
P-value belongs to the area on the left side, hence we want to know the area on the right-hand side, So P-value will become

A screenshot of a cell phone

Description automatically generated

Or Using

P(X2 > 16.2) Using Chi Square Distribution Calculator :-



**Conclusion** :->

Since P-value (0.0003) is < significance level (0.05), we will reject the null hypothesis (H0).

Hence, the gender and opinion on women reservation are dependent.