## Question 1

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

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

(b) Specify the significant variables on the **output** variable at the level of $\alpha$=0.05 and explore the related hypotheses test. Estimate the parameters of your model.

(c) Predict the output of the test dataset using the trained model. Provide the functional form of the optimal predictive model.

(d) Provide the confusion matrix and obtain the probability of correctness of predictions.

**Solution:**

To tackle this question, I have considered a relational dataset from UCI Machine Learning Repository which deals with the real estate valuation Sindian Dist., New Taipei City, Taiwan

Attribute Information:

The inputs are as follows

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

X2=the house age (unit: year)

X3=the distance to the nearest MRT station (unit: meter)

X4=the number of convenience stores in the living circle on foot (integer)

X5=the geographic coordinate, latitude. (unit: degree)

X6=the geographic coordinate, longitude. (unit: degree)

The output is as follow

Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Link for the dataset: https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

The dependent variable in our analysis is: Y= house price of unit area

The independent variable in our analysis are: X5=the geographic coordinate, latitude. (Unit: degree), X6=the geographic coordinate, longitude. (Unit: degree), X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.), X3=the distance to the nearest MRT station (unit: meter) and X4=the number of convenience stores in the living circle on foot (integer)

I have selected the above-mentioned variables are independent because they are stand alone and other variables in the model do not influence them

Since the code has been executed in RStudio, I will append the respective parts of Code as per the question along with its results and explanation of each line of code

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

For the dataset at hand, since the data points are continuous, I have opted for **Linear Regression** in **GLM**
Since y is continuous data and that falls under Linear Regression, the family associated would be '**gaussian**'

**CODE:**

```
install.packages('gmodels')
library(gmodels)#This package is used for bulding the conufusion matrix
Real.estate.valuation.data.set<-read.csv(file.choose(),header=TRUE)
#Uploading the dataset on the RStudio ecosystem
View(Real.estate.valuation.data.set)
#Viewing the dataset on the Rstudio setup
dataset <- Real.estate.valuation.data.set
#Equating the real estate to the word dataset for the further use in the code
set.seed(32)
#set seed is used to specify the initial value of the random number seed, over here 32 is set as the random number value
n=nrow(dataset)
#This is used to give the number of rows present in the dataset and we equate it to n
indexes = sample(n,n*(80/100)) # the ratio of trainset is 80% and testset is 20%
trainset = dataset[indexes,]
testset = dataset[-indexes,]
fit =
glm(Y.house.price.of.unit.area~X6.longitude+X5.latitude+X1.transaction.date+X
```

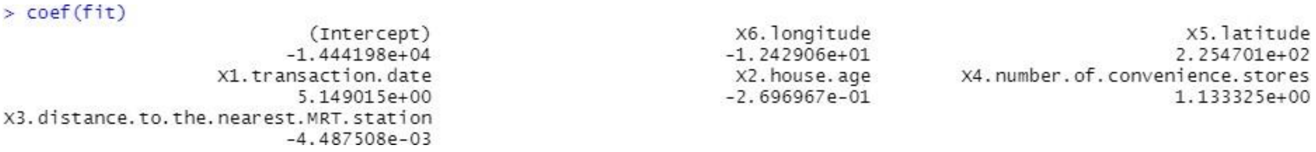
2.house.age+X4.number.of.convenience.stores+X3.distance.to.the.nearest.MRT.
station
,dataset, family='gaussian') **#Fitting the Model**
**#Where the dependent variable is Y.house.price.of.unit.area and the rest are independent variables**

2) **Specify the significant variables on the output variable at the level of $\alpha$=0.05 and explore the related hypotheses test. Estimate the parameters of your model.**

**CODE:**

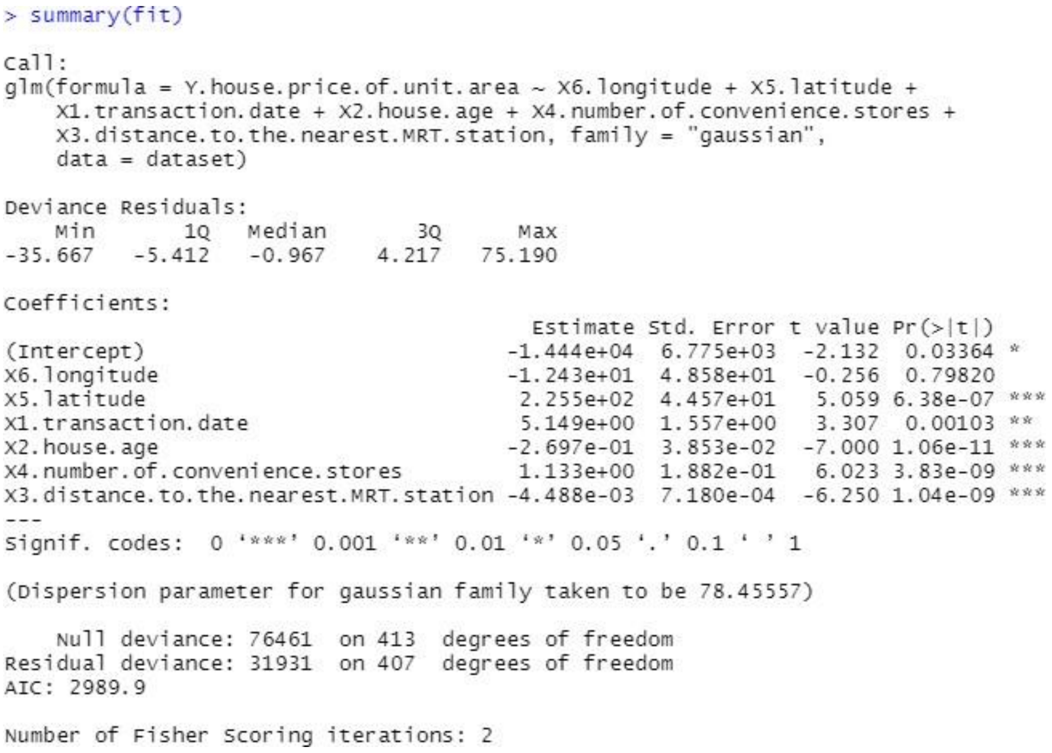
coef(fit)**#Co-efficient's of model**

summary(fit)**# show results**



```
> coef(fit)
                      (Intercept)                      X6.longitude                       X5.latitude 
                    -1.444198e+04                     -1.242906e+01                      2.254701e+02 
               X1.transaction.date                      X2.house.age     X4.number.of.convenience.stores 
                     5.149015e+00                     -2.696967e-01                      1.133325e+00 
X3.distance.to.the.nearest.MRT.station 
                    -4.487508e-03 
```

Paramaters of the model are as described below



```
> summary(fit)

Call:
glm(formula = Y.house.price.of.unit.area ~ X6.longitude + X5.latitude + 
    X1.transaction.date + X2.house.age + X4.number.of.convenience.stores + 
    X3.distance.to.the.nearest.MRT.station, family = "gaussian", 
    data = dataset)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-35.667   -5.412   -0.967    4.217   75.190  

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            -1.444e+04  6.775e+03  -2.132  0.03364 *  
X6.longitude                           -1.243e+01  4.858e+01  -0.256  0.79820    
X5.latitude                             2.255e+02  4.457e+01   5.059 6.38e-07 ***
X1.transaction.date                     5.149e+00  1.557e+00   3.307  0.00103 ** 
X2.house.age                           -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
X4.number.of.convenience.stores         1.133e+00  1.882e-01   6.023 3.83e-09 ***
X3.distance.to.the.nearest.MRT.station -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 78.45557)

    Null deviance: 76461  on 413  degrees of freedom
Residual deviance: 31931  on 407  degrees of freedom
AIC: 2989.9

Number of Fisher Scoring iterations: 2
```

As we can observe from the summary of the model, except for **x6.longitude** all the other variables are significant as they have a value lesser than 0.05

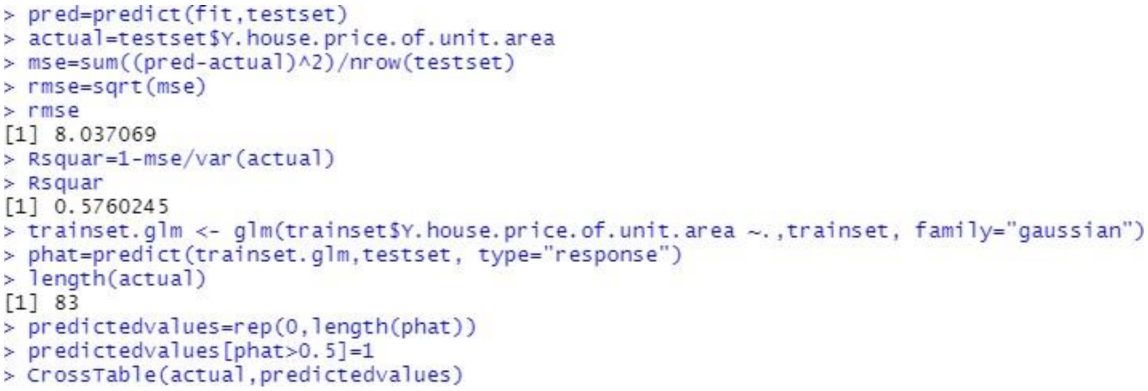
3) Predict the output of the test dataset using the trained model. Provide the functional form of the optimal predictive model.

The functional form of the optimal predictive model is:

**Pred Y.house.price.of.unit.area=**

**t=-1.444*10^$^{4}$+2.255*10$^{2}$*x5.latitude+5.149*10^$^{00}$X1.transaction.date -2.697*10^$^{-01}$* X2.house.age +1.133*10^$^{00}$* X4.number.of.convenience.stores-4.488*10$^{-03}$* X3.distance.to.the.nearest.MRT.station**

The output metrics of the test dataset are as stated in the screengrab below



```
> pred=predict(fit,testset)
> actual=testset$Y.house.price.of.unit.area
> mse=sum((pred-actual)^2)/nrow(testset)
> rmse=sqrt(mse)
> rmse
[1] 8.037069
> Rsquar=1-mse/var(actual)
> Rsquar
[1] 0.5760245
> trainset.glm <- glm(trainset$Y.house.price.of.unit.area ~.,trainset, family="gaussian")
> phat=predict(trainset.glm,testset, type="response")
> length(actual)
[1] 83
> predictedvalues=rep(0,length(phat))
> predictedvalues[phat>0.5]=1
> CrossTable(actual,predictedvalues)
```

I have calculated both the Root mean squared error and r squared, but to quantify the fit of the model and to tell us how well the plotted regression line fits the actual data we have implemented I have opted for Rsquared

The rsquared percentage is 57.6%

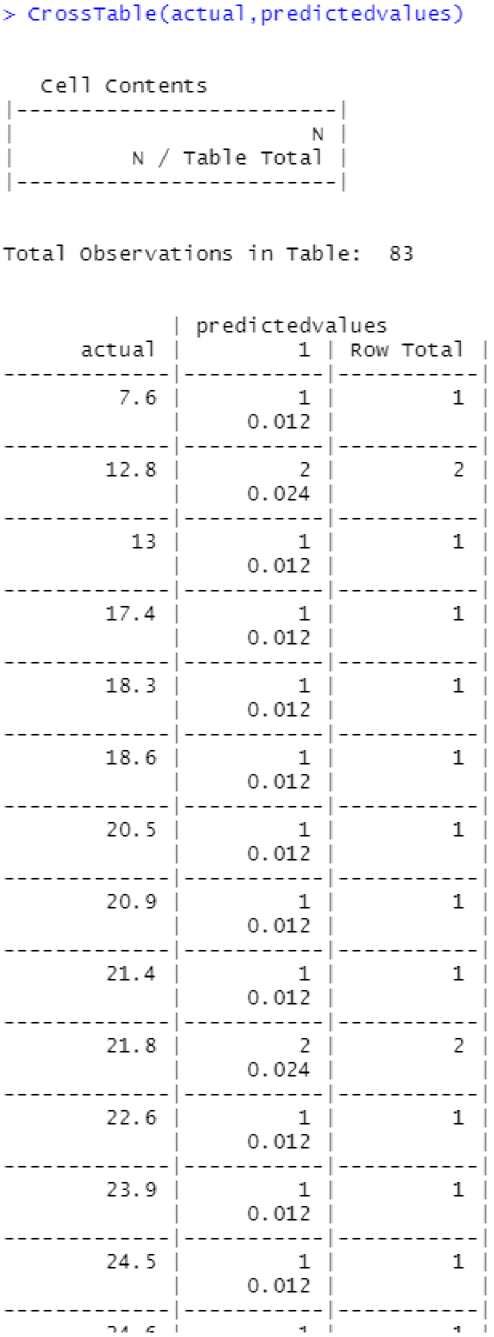
**4) Provide the confusion matrix and obtain the probability of correctness of predictions.**

CrossTable(actual,predictedvalues)

table(actual,predictedvalues)

**#The above steps are used give us the confusion matrix**

I have created the confusion matrix of the given dataset and it as depicted below



```
> CrossTable(actual,predictedvalues)

   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  83

             | predictedvalues
      actual |         1 | Row Total |
-------------|-----------|-----------|
         7.6 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        12.8 |         2 |         2 |
             |     0.024 |           |
-------------|-----------|-----------|
          13 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        17.4 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        18.3 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        18.6 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        20.5 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        20.9 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        21.4 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        21.8 |         2 |         2 |
             |     0.024 |           |
-------------|-----------|-----------|
        22.6 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        23.9 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        24.5 |         1 |         1 |
             |     0.012 |           |
-------------|-----------|-----------|
        24.6 |         1 |         1 |
```

For the probability of correctness of predictions, Outliers have a serious influence on the regression line derived from linear regression. Thus, classifying two classes will not be successful. Thus, by passing the predicted value through the sigmoid function and changing it to probability [1]

## Question 2

Let $x_1, \dots , x_{10}$ are identically independently distributed (iid) with Poisson ($\lambda$).

**a)** Compute the likelihood function (LF**).**

**b)** Adopt the appropriate conjugate prior to the parameter $\lambda$ (Hint: Choose hyperparameters optionally within the support of distribution).

**c)** Using (a) and (b), find the posterior distribution of $\lambda$.

**d)** Compute the minimum Bayesian risk estimator of $\lambda$.

**SOLUTION:**

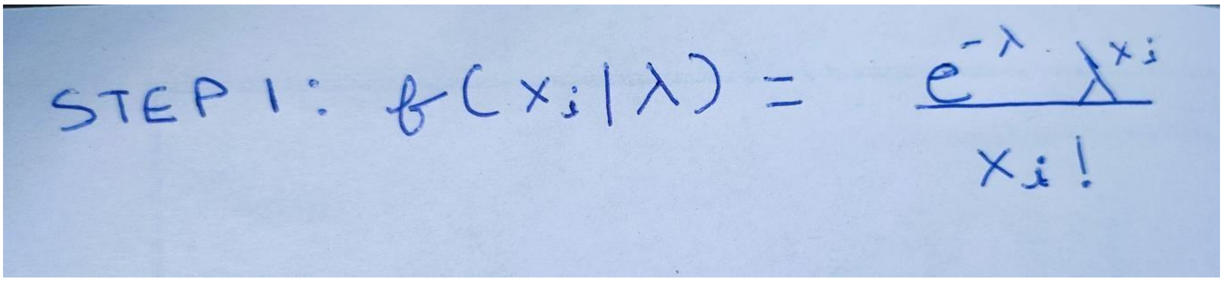
**FOR ALL THE EQUATIONS I HAVE WRITTEN IT ON PAPER AND I HAVE UPLOADED THE SAME OVER HERE SO THAT IT WOULD LOOK NEATER AND MORE COMPACT**

### a) Compute the likelihood function (LF).

Considering the model is Poisson distribution, we have to calculate the likelihood function keeping the Poisson distribution under consideration

Likelihood function = Lf

x1 .....x10 are distributed through the Poisson distribution (which is represented by lambda)



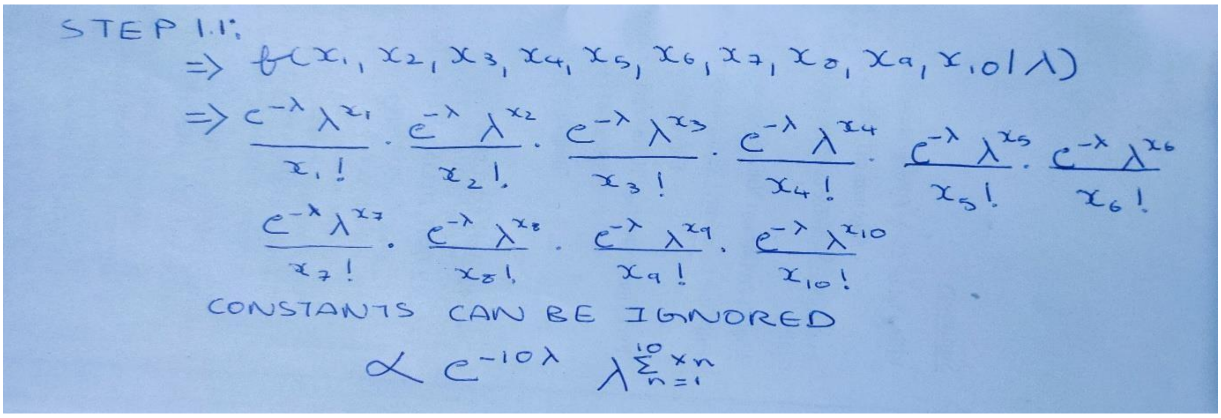
STEP 1: $f(x_i | \lambda) = \dfrac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i!}$

Lf p (x1, x2, x3, x4, x5, x6, x7, x8, x9, x10 | λ) = f (x1| λ) f (x2| λ) f (x3| λ) f (x4| λ) f (x5| λ) f (x6| λ) f (x7| λ) f (x8| λ) f (x9| λ) f (x10| λ)

This is essentially the multiplication of each conditional probability

Then we have to multiply it with the conditional probability and we have to compute and simplify the answer

The final LF (Likelihood Function) is as specified in the picture below, I have used the sigma notation to signify the summation of the terms used in the equation



STEP 1.1:

$$\Rightarrow f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} | \lambda)$$

$$\Rightarrow \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdot \frac{e^{-\lambda}\lambda^{x_3}}{x_3!} \cdot \frac{e^{-\lambda}\lambda^{x_4}}{x_4!} \cdot \frac{e^{-\lambda}\lambda^{x_5}}{x_5!} \cdot \frac{e^{-\lambda}\lambda^{x_6}}{x_6!}$$

$$\frac{e^{-\lambda}\lambda^{x_7}}{x_7!} \cdot \frac{e^{-\lambda}\lambda^{x_8}}{x_8!} \cdot \frac{e^{-\lambda}\lambda^{x_9}}{x_9!} \cdot \frac{e^{-\lambda}\lambda^{x_{10}}}{x_{10}!}$$

CONSTANTS CAN BE IGNORED

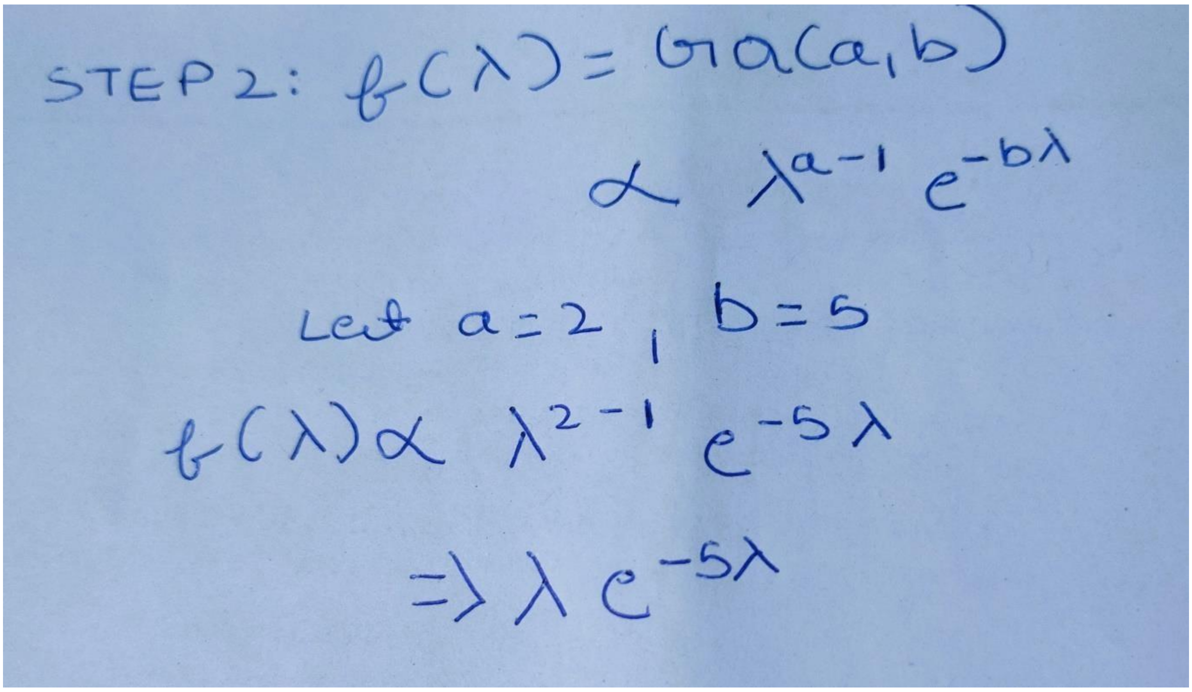
$$\propto e^{-10\lambda} \lambda^{\sum_{n=1}^{10} x_n}$$

## b) Adopt the appropriate conjugate prior to the parameter $\lambda$ (Hint: Choose hyperparameters optionally within the support of distribution).

The conjugate prior for $\lambda$ is Gamma

We have the hyperparameters(a,b) as a = 2 and b = 5

The prior is represented in the end of the equation in the picture attached



$$\text{STEP 2: } f(\lambda) = Ga(a,b)$$

$$\propto \lambda^{a-1} e^{-b\lambda}$$

Let $a = 2$, $b = 5$

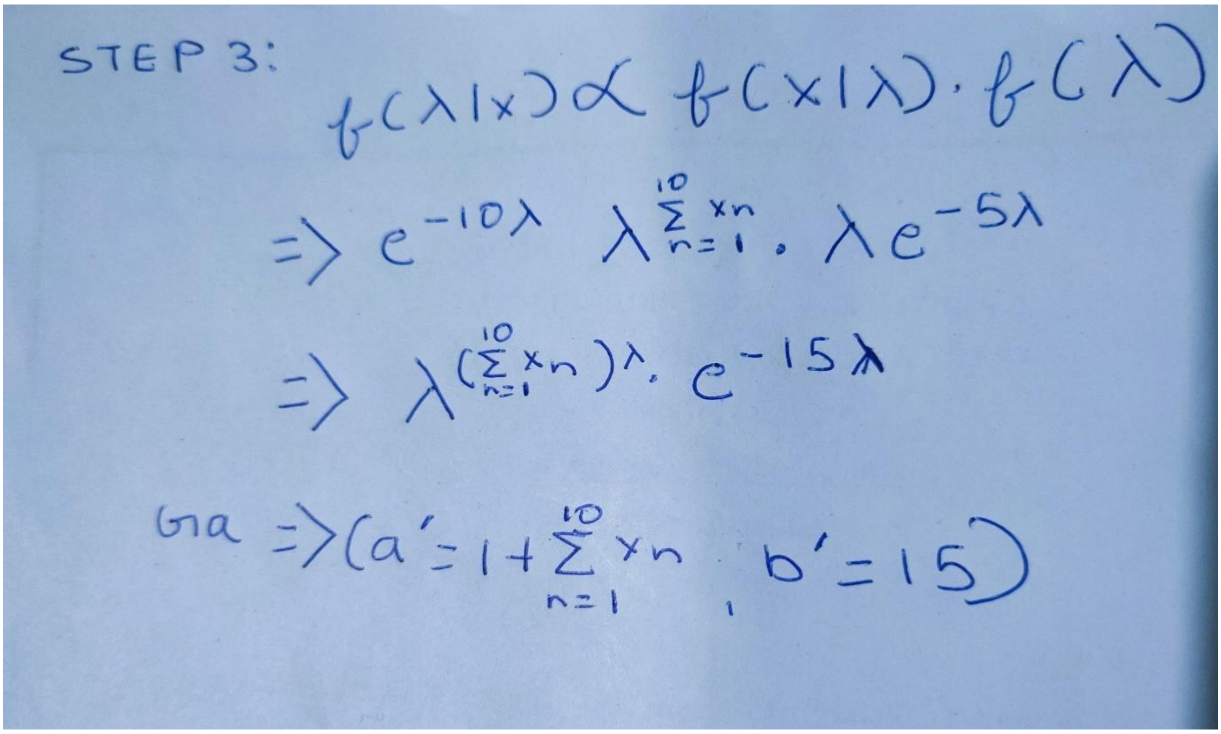
$$f(\lambda) \propto \lambda^{2-1} e^{-5\lambda}$$

$$\Rightarrow \lambda e^{-5\lambda}$$

## c)Using (a) and (b), find the posterior distribution of $\lambda$.

We have to multiply the prior with the likelihood function through the bayes rule to get posterior distribution of $\lambda$

This is signified in the picture attached below, in which we get the values of a` and b` (hyper parameters)



STEP 3:

$$f(\lambda|x) \propto f(x|\lambda)\cdot f(\lambda)$$

$$\Rightarrow e^{-10\lambda}\ \lambda^{\sum_{n=1}^{10} x_n} \cdot \lambda e^{-5\lambda}$$

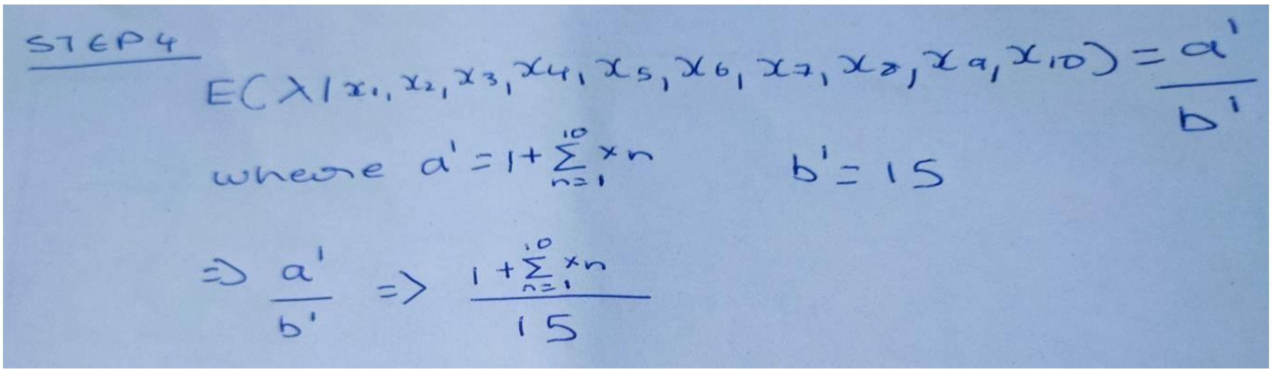
$$\Rightarrow \lambda^{(\sum_{n=1}^{10} x_n)^{\lambda}} \cdot e^{-15\lambda}$$

$$\text{Ga} \Rightarrow \left(a' = 1+\sum_{n=1}^{10} x_n\ ,\ b' = 15\right)$$

## d)Compute the minimum Bayesian risk estimator of $\lambda$.

To obtain the minimum Bayesian risk estimator, we need to calculate the expectation of lambda, which is a`/b`

The result is displayed through the picture attached below



STEP4

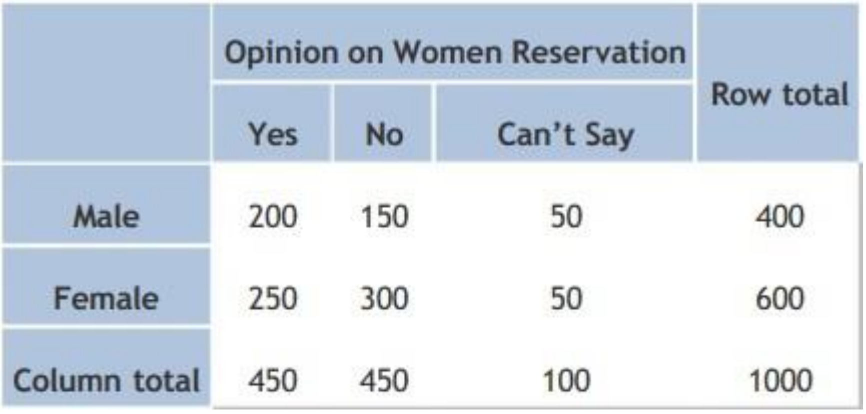
$$E(\lambda|x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}) = \frac{a'}{b'}$$

$$\text{where } a' = 1+\sum_{n=1}^{10} x_n \qquad b' = 15$$

$$\Rightarrow \frac{a'}{b'} \Rightarrow \frac{1+\sum_{n=1}^{10} x_n}{15}$$

## 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.



| | Opinion on Women Reservation | | | Row total |
|---|---|---|---|---|
| | Yes | No | Can't Say | |
| Male | 200 | 150 | 50 | 400 |
| Female | 250 | 300 | 50 | 600 |
| Column total | 450 | 450 | 100 | 1000 |

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

a) State the hypotheses.
b) Find the statistic and critical values.
c) Explain your decision and Interpret results

Since the code has been executed in RStudio, I will append the respective parts of Code as per the question along with its results and explanation of each line of code

**a) State the hypotheses.**

For this we have to consider the Null Hypothesis and the Alternative Hypothesis

Ho = Null Hypothesis

H1 = Alternative Hypothesis

Ho: Gender and Opinion on Women Reservation are independent

H1: Gender and Opinion on Women Reservation are not independent

**b) Find the statistic and critical values**

alpha=0.05

Since 0.05 is the level of significance, I have opted for Chi square test, as it is according to me the best method to analyse the fluctuations in the same set of variables

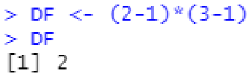
To move ahead with the chi squared test, we have to calculate the degree of freedom (DF)

DF = (R-1) * (C-1), where R is the no of rows and C is the no of columns in the table specified

In our table we have R = 2 and C = 3

DF <- (2-1) *(3-1)

DF



```
> DF <- (2-1)*(3-1)
> DF
[1] 2
```

So, from the above screengrab of the code from Rstudio it is clear that the value of DF is 2

After computing the value of DF, we need to calculate expected frequency count

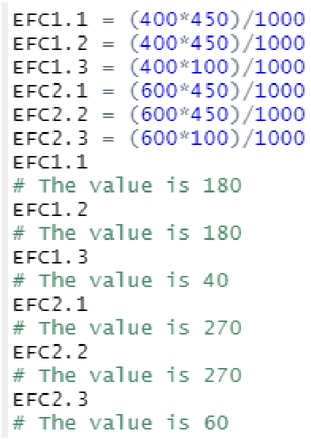
Expected frequency count is basically a probability count used in chi square test while tackling table calculations

To calculate the EFC (Expected Frequency count) we have to sum up the total of the 'a'th row and the total in the 'b'th column, then divide the result by the total

### c) Explain your decision and Interpret results

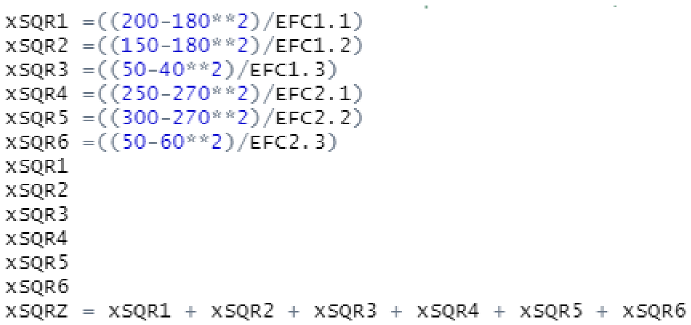
EFC = (aR * bC)/n where n represents the total population, 'a' represents row and 'b' represents column

n = 1000(No of Students)



```
EFC1.1 = (400*450)/1000
EFC1.2 = (400*450)/1000
EFC1.3 = (400*100)/1000
EFC2.1 = (600*450)/1000
EFC2.2 = (600*450)/1000
EFC2.3 = (600*100)/1000
EFC1.1
# The value is 180
EFC1.2
# The value is 180
EFC1.3
# The value is 40
EFC2.1
# The value is 270
EFC2.2
# The value is 270
EFC2.3
# The value is 60
```

Now we have to calculate Chi square statistic, for that the formula is the summation of the EFC minus the value under consideration divided by the EFC



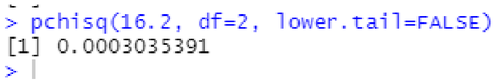
```
xSQR1 =((200-180**2)/EFC1.1)
xSQR2 =((150-180**2)/EFC1.2)
xSQR3 =((50-40**2)/EFC1.3)
xSQR4 =((250-270**2)/EFC2.1)
xSQR5 =((300-270**2)/EFC2.2)
xSQR6 =((50-60**2)/EFC2.3)
xSQR1
xSQR2
xSQR3
xSQR4
xSQR5
xSQR6
xSQRZ = xSQR1 + xSQR2 + xSQR3 + xSQR4 + xSQR5 + xSQR6
```

The above formula gives us chi square statistic value

**xSQRZ = 16.2**

The chi square value is 16.2

Now we have to use the degree of freedom and the chi square value to calculate the p square value



```
> pchisq(16.2, df=2, lower.tail=FALSE)
[1] 0.0003035391
> 
```

The p square value of 0.0003 is lesser than the value of significance level of 0.05, because of this we have to reject the null hypothesis

**So, after performing the chi square test, I have come to the conclusion that there is a relationship between opinion and gender on women reservation**

**REFERENCES:**

[1] https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/#:~:text=As%20this%20regression%20line%20is,value%20from%20the%20regression%20line.

[2] https://elearning.dbs.ie/my/