Name	**Robin Roy** **https://github.com/lamrobinroy/**

Project: The selected dataset is *diabetes.csv* https://www.kaggle.com/johndasilva/diabetes from Kaggle

Input variable are X1, X2, X3 and Y is the output variable with Binomial data.

```
> mydata=read.csv("/cloud/project/diabetes.csv")
> head(mydata)
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
1           2     138            62            35       0 33.6
2           0      84            82            31     125 38.2
3           0     145             0             0       0 44.2
4           0     135            68            42     250 42.3
5           1     139            62            41     480 40.7
6           0     173            78            32     265 46.5
  DiabetesPedigreeFunction Age Outcome
1                    0.127  47       1
2                    0.233  23       0
3                    0.630  31       1
4                    0.365  24       1
5                    0.536  21       0
6                    1.159  58       0
```

```
> x1=mydata$Glucose
> x2=mydata$BloodPressure
> x3=mydata$BMI
> y=mydata$Outcome
>
```


A)

We select the GLM, logistic regression model as the output variable is discrete and has only two possible outcomes which are either 0 or 1

```
> #Split the datset in 80% trainset and 20% testset
> set.seed(3000)
> n=nrow(dataset)
> indexes=sample(n,n*(80/100))
> trainset=dataset[indexes,]
> testset=dataset[-indexes,]
```

B)

```
> #Data Cleaning
> dataset=na.omit(data.frame(x1,x2,x3,y))
> #fiting the model
> fit=glm(y~.,data=dataset,family='binomial') #this is logistic regression
> summary(fit)
```

Output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.908332   0.384400 -17.972   <2e-16 ***
x1           0.034191   0.002003  17.070   <2e-16 ***
x2          -0.003635   0.003018  -1.204    0.228
x3           0.067508   0.007969   8.471   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2569.4  on 1999  degrees of freedom
Residual deviance: 2028.1  on 1996  degrees of freedom
AIC: 2036.1
```

Alpha=0.05. From summary (fit), we observe that all values in P possess 0.05, and we understand that all the input variables are significant

C)

```
> lt=length(pred)
> predictedval=rep(0,lt)
> predictedval
> predictedval[pred>0.5]=1 #probability of outcome being 1, if p<
0.5 then outcome=0
> predictedval
> df=data.frame(testset[,4],predictedval)
> View(df)
```

```
> #Model Prediction
> pred=predict(fit,testset,type="response")
> pred
> #convert phat tp yhat
> predictedval=rep(0,nrow(dataset))
> predictedval
```

```
predictedval[pred>0.5]=1
preeictedval
rror: object 'preeictedval' not found
predictedval
  [1] 0 1 1 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0
 [38] 0 0 0 0 1 0 0 0 1 0 1 0 0 0 1 1 1 1 0 0 0 1 1 0 1 0 0 0 1 1 0 0 1 0 0 0 0
 [75] 0 1 0 0 0 1 0 0 0 1 0 0 0 1 1 1 1 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0
[112] 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
[149] 0 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0
[186] 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0
[223] 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
[260] 0 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0
[297] 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1 1 1 0 0 1
[334] 0 0 0 0 0 0 1 1 1 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
[371] 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0 0
[408] 0 1 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0
[445] 0 1 0 1 0 0 0 1 1 1 1 0 0 0 1 1 0 1 0 0 0 1 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0
[482] 0 0 1 0 0 0 1 1 1 1 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0
[519] 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0
[556] 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0
[593] 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0
[630] 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 1 0 0 1
[667] 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 1
[704] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1 1 1 0 0 1 0 0 0 0 0 0 1
[741] 1 1 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0
[778] 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 1 1 0 0 0 0 1 0 0 1 0 0
[815] 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0
[852] 1 1 1 1 0 0 0 1 1 0 1 0 0 0 1 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1
[889] 1 1 1 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0
[926] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0
[963] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0
1000] 0
[ reached getOption("max.print") -- omitted 1000 entries ]
```


D) Confusion Matrix

```
> #Confusion Matrix
> conf_Matrix=table(predictedval,actualvalues=testset[,4])
> conf_Matrix
           actualvalues
predictedval   0   1
           0 249  56
           1  29  66
```

Accuracy

```
> #Accuracy
> accuracy_val=mean(predictedval==testset[,4]) #Correctness Prediction
> accuracy_val
[1] 0.7875
```

2)

a) A likelihood function is the probability density of the data, viewed as a function of the parameters

Let $x_1, x_2 \dots x_n$ have a joint density function $f(x_1, x_2 \dots, x_n | \theta)$. Given $x = x_1, x_2 = x_2 \dots, x_n = x_n$ is observed, the function of θ defined by:

$$L(\theta) = L(\theta | x_1, x_2, \dots, x_n) = f(x_1, x_2 \dots x_n | \theta)$$

~~[illegible]~~

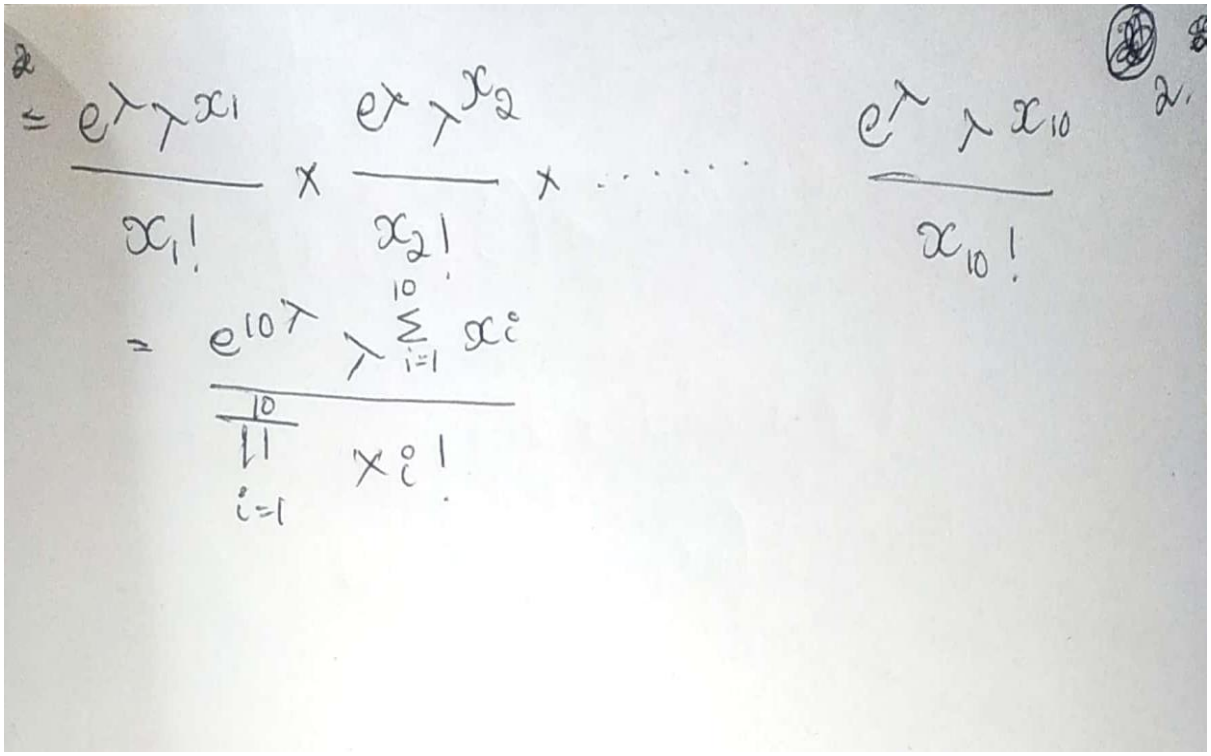
for poisson distribution, $f(x_i | \lambda)$

$$= \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \text{ where}$$

$x_i = 0, 1, 2 \dots$ and $\lambda > 0$

$$L(\lambda | x_1, x_2 \dots, x_{10}) = \prod_{i=1}^{10} f(x_i | \lambda)$$

$$= \prod_{i=1}^{10} \left(\frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right)$$



b) we adapt Gamma model as a conjugate prior model to the parameter λ

$F(\lambda) = Ga(\alpha, \beta)$ where $\alpha > 0$ and $\beta > 0$

$= \frac{\beta^\alpha}{\Gamma(\beta)} \lambda^{\alpha-1} e^{-\beta\lambda}$ $\frac{\beta^\alpha}{\Gamma(\beta)}$ is free of λ

ie, $F(\lambda) \propto e^{-\beta\lambda} \lambda^{\alpha-1}$

if $\alpha = 2, \beta = 5$, $F(\lambda) \propto e^{-5\lambda} \lambda$

c) Posterior distribution = Prior distribution * likelihood function

Posterior probability $\propto$ likelihood * prior probability

$F(\lambda | x_1, x_2 \cdots x_{10}) \propto F(\lambda) * L(\lambda | x_1, x_2 \cdots, x_{10})$

$= e^{-10\lambda} \lambda^{\sum_{i=1}^{10} x_i} * e^{-10\lambda} \lambda$

$= e^{-10\lambda} \lambda^{\sum_{i=1}^{10} x_i + 1}$

$\alpha^* = \sum_{i=1}^{10} x_i + \alpha$ $\beta^* = \beta + n$

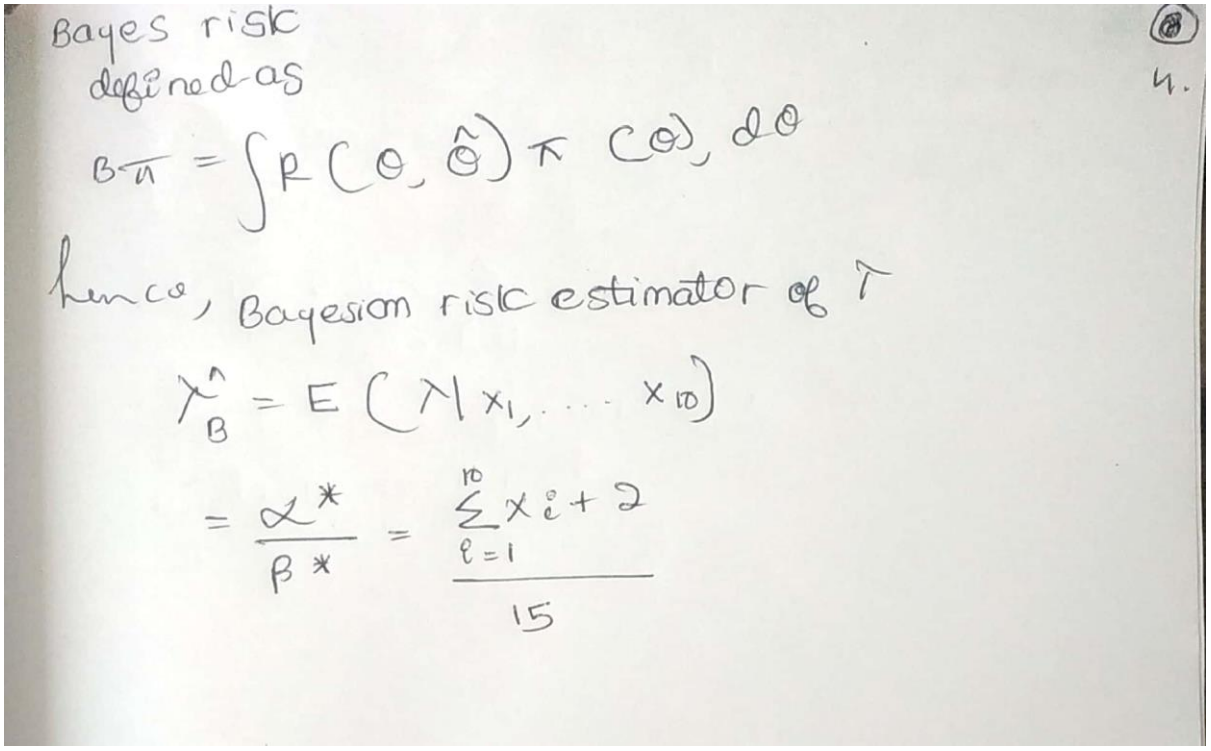
where we know $n = 10$, $\alpha = 2$, $\beta = 5$

ie, $F(\lambda | x_1, x_2 \cdots x_{10}) \propto Ga(\alpha^* = \sum_{i=1}^{10} x_i + 2, \beta^* = 15)$

d) we define risk of an estimator as $\hat{\theta}(x)$ as

$R(\theta, \theta^n) = E_\theta(L(\theta, \hat{\theta}))$

$= \int L(\theta, \hat{\theta}(x)) P_\theta(x) dx$



Bayes risk defined as

$$B_\pi = \int R(\theta, \hat{\theta}) \pi(\theta) \, d\theta$$

hence, Bayesian risk estimator of λ

$$\hat{\lambda}_B = E(\lambda | x_1, \ldots x_{10})$$

$$= \frac{\alpha^*}{\beta^*} = \frac{\sum_{i=1}^{10} x_i + 2}{15}$$

3)

	Opinion on women reservation.			
	Yes	No	Can't Say	Row Total
Male	200	150	50	400
Female	250	300	50	600
Column Total	450	450	100	1000

Total no. of students = 1000.

Respondents = Male, Female

a) Opinion = Yes, No, Can't Say

b) H_0: women reservation are independent

H_1: women reservation are not independent

6.

(b) Level of significance = 0.05 = α

$$\varepsilon_{ij} = \frac{\text{total ith row} \times \text{total ith column}}{\text{total}}$$

$$\varepsilon_{11} = \frac{400 \times 450}{1000} = 180$$

$$\varepsilon_{10} = \frac{400 \times 450}{1000} = 180$$

$$\varepsilon_{13} = \frac{400 \times 100}{1000} = 40$$

$$\varepsilon_{21} = \frac{600 \times 450}{1000} = 270$$

$$\varepsilon_{22} = \frac{600 \times 450}{1000} = 270$$

$$\varepsilon_{23} = \frac{600 \times 100}{1000} = 60$$

$$\text{total value} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(o_{ij} - \varepsilon_{ij})^2}{\varepsilon_{ij}}$$

$$= \frac{(200-180)^2}{100} + \frac{(150-180)^2}{180} + \frac{(50-80)^2}{180}$$

$$+ \frac{(250-270)^2}{270} + \frac{(300-270)^2}{270}$$

$$+ \frac{(50-60)^2}{60}$$

$$= 2.23 + 5 + 2.5 + 1.48 + 3.39 + 1.67$$

$$= 16.22$$

$r = 2$

$c = 3$

degree of α =

$$(2-1)(c-1)$$

$$= (2-1) \times (3-1) = 2$$

Hence the critical value = 5.99

$$= \chi^2_{0.05}(r-1)(c-1) = \chi^2_{(2)} = 5.99$$

5) we have found the the critical value is less than the test value

Critical value < test value

5.99 < test value

H_0 is rejected

H_i is greater than H_0