**Lab 8**

Name: Jesto Peter

Title: Logistic Regression

Date: 28/01/2023

Class: 2MSTAT

* **Objective**
* Select any two datasets that consist of at least one binomial variable as a dependent and fit a **logistic regression model**.
* To prepare a report that consists of an introduction, analysis, and conclusion.
* **Procedure**

library("olsrr")

library("pscl")

library(MASS)

**#Dataset 1: We consider surgical dataset**  
data=surgical  
attach(data)  
head(data)

## bcs pindex enzyme\_test liver\_test age gender alc\_mod alc\_heavy y  
## 1 6.7 62 81 2.59 50 0 1 0 695  
## 2 5.1 59 66 1.70 39 0 0 0 403  
## 3 7.4 57 83 2.16 55 0 0 0 710  
## 4 6.5 73 41 2.01 48 0 0 0 349  
## 5 7.8 65 115 4.30 45 0 0 1 2343  
## 6 5.8 38 72 1.42 65 1 1 0 348

#We consider the dependent variable as alc\_heavy which indicates if the alcohol use was heavy or not(categorical variable)  
#Fitting model  
  
log\_model=glm(alc\_heavy~.,data,family = 'binomial')

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(log\_model)

##   
## Call:  
## glm(formula = alc\_heavy ~ ., family = "binomial", data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.072 0.000 0.000 0.000 1.236   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 94.49894 67.79583 1.394 0.163  
## bcs 0.78018 1.14601 0.681 0.496  
## pindex -1.15654 0.84471 -1.369 0.171  
## enzyme\_test -1.09427 0.79032 -1.385 0.166  
## liver\_test -4.62364 3.88231 -1.191 0.234  
## age -0.29989 0.21790 -1.376 0.169  
## gender -5.62454 3.70007 -1.520 0.128  
## alc\_mod -35.85588 4697.65897 -0.008 0.994  
## y 0.11423 0.08065 1.416 0.157  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 51.750 on 53 degrees of freedom  
## Residual deviance: 11.415 on 45 degrees of freedom  
## AIC: 29.415  
##   
## Number of Fisher Scoring iterations: 21

**#Conclusion:** In logistic regression ,we consider the best model as one whose diffrence between Null Deviance(Total SUm of squares) and Residual Decviance(Residual Sum of Squares) is the highest. Another way to check goodness of fit is to do the McFadden Test.

#Checking goodness of Fit  
  
pscl::pR2(log\_model)["McFadden"]

## fitting null model for pseudo-r2

## McFadden   
## 0.7794288

**#Conclusion:** We see that the test value is 0.7 ,hence our model is a very good fit.

#Predicting  
prob=predict(log\_model,data)  
head(cbind(alc\_heavy,round(prob,2)))

## alc\_heavy   
## 1 0 -44.05  
## 2 0 -15.50  
## 3 0 -1.85  
## 4 0 -13.55  
## 5 1 133.83  
## 6 0 -51.50

#Creating the confusion matrix   
cutoff=ifelse(prob<0.5,0,1)  
table(alc\_heavy,cutoff)

## cutoff  
## alc\_heavy 0 1  
## 0 43 1  
## 1 1 9

**#Accuracy**=TN+TP/TP+TN+FP+FN  
  
accuracy=(43+9)/(43+9+1+1)  
accuracy

## [1] 0.962963

**#Conclusion**: We have an accuracy of 96%

**#Dataset 2: We consider Epileptics Dataset**

df=epil  
data1=df[-2]  
  
  
attach(data1)

## The following objects are masked from data:  
##   
## age, y

(data1)

## y base age V4 subject period lbase lage  
## 1 5 11 31 0 1 1 -0.75635379 0.11420370  
## 2 3 11 31 0 1 2 -0.75635379 0.11420370  
## 3 3 11 31 0 1 3 -0.75635379 0.11420370  
## 4 3 11 31 1 1 4 -0.75635379 0.11420370  
## 5 3 11 30 0 2 1 -0.75635379 0.08141387  
## 6 5 11 30 0 2 2 -0.75635379 0.08141387

#We consider the dependent variable as V4 which indicates the group type to be assigned to  
#Fitting model  
  
log\_model1=glm(V4~.,data1,family = 'binomial')

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(log\_model1)

##   
## Call:  
## glm(formula = V4 ~ ., family = "binomial", data = data1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -9.097e-06 -1.901e-06 -2.110e-08 1.636e-06 7.064e-06   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.692e+02 4.204e+05 0.000 1.000  
## y -4.838e-03 1.275e+03 0.000 1.000  
## base 2.297e-03 1.077e+03 0.000 1.000  
## age -6.424e-03 1.463e+04 0.000 1.000  
## subject 1.761e-04 5.991e+02 0.000 1.000  
## period 4.845e+01 2.065e+04 0.002 0.998  
## lbase -2.547e-02 3.263e+04 0.000 1.000  
## lage 2.051e-01 4.135e+05 0.000 1.000  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2.6542e+02 on 235 degrees of freedom  
## Residual deviance: 7.2908e-09 on 228 degrees of freedom  
## AIC: 16  
##   
## Number of Fisher Scoring iterations: 25

#Conclusion: In logistic regression ,we consider the best model as one whose diffrence between Null Deviance(Total Sum of squares) and Residual Decviance(Residual Sum of Squares) is the highest. Another way to check goodness of fit is to do the McFadden Test.

#Checking goodness of Fit  
  
pscl::pR2(log\_model1)["McFadden"]

## fitting null model for pseudo-r2

## McFadden   
## 1

#Conclusion: We see that the test value is 1 ,hence our model is a very good fit.

#Predicting  
prob1=predict(log\_model1,data1)  
head(cbind(V4,round(prob1,2)))

## V4   
## 1 0 -120.91  
## 2 0 -72.45  
## 3 0 -23.99  
## 4 1 24.46  
## 5 0 -120.90  
## 6 0 -72.46

#Creating the confusion matrix   
cutoff1=ifelse(prob1<0.5,0,1)  
table(V4,cutoff1)

## cutoff1  
## V4 0 1  
## 0 177 0  
## 1 0 59

#Accuracy=TN+TP/TP+TN+FP+FN  
  
accuracy1=(177+59)/(177+59+0+0)  
accuracy1

## [1] 1

#Conclusion: We have an accuracy of 100%

* **CONCLUSION**

Hence we did LOGISTIC REGRESSION for 2 datasets, namely Surgical and Epiletic with binomial family. And found a model to predict the dependent variable in each dataset. We found the accuracy of the models after finding the confusion matrix.