Solving the probability regression models by using R programming environment
Case Study For patients Epilepsy in Basra

Assist. Prof. Nada. Badr. Jarah

University of Basra -Collage of management and economic - Statistics department

General Specialty: Computer Science

nadabadrjarah@yahoo.com

Abstract

The technology for dealing with accurate calculations progresses very quickly . and be a major impetus for scientific progress. and the perfect mastery of R in higher levels is necessary for many applications .

The R statistical programming environment provides an ideal stage to conduct the study because of its powerful programming capability , graphics and a comprehensive set of statistical functions , it contains more than 11164 packages .

The aim of this study is to solve the probability regression models represented by probit, tobit and logit ,as well as identifying the variables of the study (sex and age) that affect or increase the incidence epilepsy for the data of 2296 patients in Basra.

After estimating the three probable regression models for the effect of sex and age factors on the risk of developing epilepsy by using R programming language , the results showed that the regression of the probabilistic functions is that both the age factor and the sex factor have a significant effect in the case of the disease in terms of being inpatient or outpatient. This is confirmed by the Wold test. Also, these functions represented the best representation, which was confirmed by the coefficient of selection as it reached 0.96.

Key words : R language , Statistical Programming , probability regression , Epilepsy

1.Introduction

R is an open source program ,it can be downloaded within few minutes from R site ,it combines between the programming language and statistical analysis and a number of additional packages used to solve probability regression models represented by probit, tobit and logit .So with a few steps and simple programming phrases we obtained results and statistical analysis.[1]

Epilepsy is a neurological condition , which means that it affects the brain and also is a physical condition because the body is affected during the attacks of seizures and is studied in our research and comes in fourth place for the most common neurological problems , Migraine , stroke and Alzheimer's are the most frequent. [2]

To show the number of times the epilepsy occurs , the number of disease is used which varies from one patient to another ,some have several attacks end after a period of live and the others suffer from it for the rest of their lives.[3]

The programming phrases written by R language was applied to the study and analysis of the cases of epilepsy and the effect of age and sex variables on the incidence of the disease directly or indirectly . the study relied on the data available in basra healthy department for patients with epilepsy in 2008 and their total number were 2296 patients .

This study contributes to a measure of the progress of societies free from disease, because of the spread of epilepsy disease, which draws attention and calls for the need to study and analyze the case of infection and complications, which constitute a problem in terms of health, social and economic. It is becoming more widespread day after day, and its relationship with civilized progress is a direct relationship, unlike many diseases that scientific progress has been able to reduce or eliminate some of them permanently. So it has become a problem worthy of study to determine the factors that affect the incidence of epilepsy in the province of Basra in order to obtain indicators that help us to describe and identify the most important variables affecting the occurrence of this disease. It must be emphasized that it must be addressed at various levels of health, information and economics. As well as the lack of studies related to Basra in particular and in the Arab world in general, whether those studies were medical or statistical.

The aim of this study is to analyze probabilistic regression models of Probit, Tobit and Logit models, as well as to determine the study variables (age and sex) that affect or increase the incidence of epilepsy through data for 2296 patients in Basrah governorate .

2. Literature review

This is a new study. The statistical of the research is to solve probabilistic regression models of probit, Tobit and Logit using R program and epilepsy data, and may be found in one study is HSAUR.

Data for epilepsy patients were in a clinical trial. Randomly selected 59 patients who were divided into groups according to patient age, number of bouts and treatment period. Progabide anti-epilepsy was given and a number of episodes that occurred for two weeks were compared with another drug. Then use the R program to draw the comparison results for the study. [4]

3. The theoretical side :

3.1 programming language R

The use of the computer because of the high capacity to store information and retrieval and the existence of a number of statistical programs that have high accuracy in drawing forms and processing digital data at tremendous speed. And provide an opportunities for experimentation, discovery and processing, we choose the programming language R, because it has a wide potential and several sources to resd data, as well as the availability of statistical packages ready to simulates the same condition and the use of one of the probability models to study data specific to a particular region.[5]

3.2 probability regression models [6]

The forms of regression relationships can range from simplicity (where the value of y is calculated in the term of single variable x1) to more complex forms ,where we use more than one x transformer to calculate the value of y for example (y = a + bx + c x2 + d x3), this regression relationship can only be linear (showing x in the first degree) and can be extended to quadratic or cubical formulas like (y = a + bx + c x2 + dx3), to calculate a linear regression in R, we use the following formula : lm (y ~ x)

First : Probit model

Is a type of regression test for the case of binomial data in which the response either 0 or 1, is used in the case of data that follows normal distribution , and depends on the cumulative distribution function of normal distribution**.**

Second : Tobit model

Is close to the former one, called approximation regression model where the dependent variable is shown if certain conditions are met .This method is used the dependent variable contains zero and other continuous observations and has more than one form taken by the logarithmic probability function according to the dependent variable method.[7]

Third : Logit model

Is another form of binary data conversion to linear one and is very similar to probit. It is used when the dependent variable takes only two values, zero and one. And the descriptive variables are descriptive or quantitative. It deals with the variance of the dependent variable, which changes with the variable of the explanatory variable and depends on the cumulative distribution function and the relationship between the probability and the explanatory variable is a nonlinear relationship.

Till now, the relationship between Probit and Logit is almost indistinguishable, and some find the difference between Probit and Logit is that probit is preferred when the data is usually distributed in its first form or binary code. But when data not follows the natural distribution, the use of the Probit model, even if the data does not normal distribution then best model is logit regression **.**

3.3 Epilepsy

Due to the spread of epilepsy disease in the province of Basra, which draws attention and calls for the need to know all facts about this disease, as well as the lack of studies on this subject in Basra in particular and in Iraq in general, whether those studies, medical , statistical or information.

Epilepsy is one of the oldest and the most common neurological diseases among children and adults, it is one of the most common diseases that have misconceptions and that disability due to this disease is the result of these misconceptions surrounding the disease and the way patients are treated by their families, teachers and colleagues at work. There is a wrong idea that patients with epilepsy are less able and less intelligent than their peers, and that the disease lasts for a life time despite the ability to curve the disease through treatment as well as the sufferers can exercise their normal lives even before they fully recover from the disease .[8]

Hence, it is clear that epilepsy and its complications is actually a problem in terms of health, social . And economic and it is becoming more widespread day after day, and its relationship with civilized progress is a positive relationship, unlike many diseases which scientific progress has been able to reduce or eliminate some of them.[9]

3.4. Packages in R

Is the studies and research and data contributed by thousands of users of the language R, which amounted to writing these lines to (11871), and it represented in words of software or programs ready to solve most of the statistical problems and written by researchers from all over the world, and when we need certain packages we first install, loaded R contains a set of packages that enable us to read , analyze data and perform simple statistical tests.[10]

Since it is a possible that someone has already solved the problem we are working on, and can access and benefit from their work by uploading that package on the Internet we use the term install.packages ()

If you do not have a package installed, run: install.packages ("packagename"), or if you see the version is out of date, run: update.packages ().

In the CRAN package store, we have both a robustbase [11] package and a HSAUR [12] package that uses epilepsy data in randomized clinical trial to investigate the effect of an antiretroviral drug. The data frame consists of 236 cases on the following six variables:

1) treatment: The treatment group is a factor of placebo and progabide

2) base: The number of seizures before the test

3) age: Age of the patient

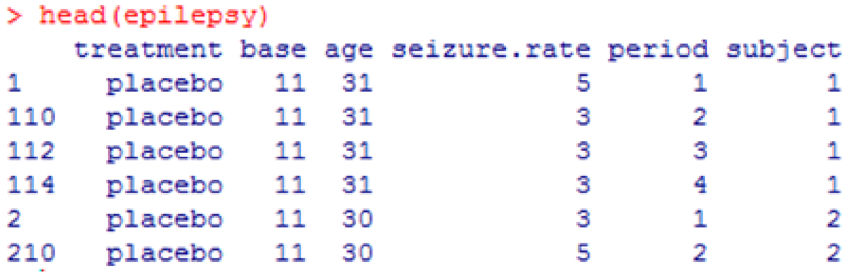
4) seizure. rate: pathological seizure rate

5) period: Treatment period of 1-4

6) subject: the patient's identity of 1-59

This data can be accessed as follows :

```
> install.packages('HSAUR')
> data(packages = 'epliepsy')
```



```
> head(epilepsy)
    treatment base age seizure.rate period subject
1     placebo   11  31            5      1       1
110   placebo   11  31            3      2       1
112   placebo   11  31            3      3       1
114   placebo   11  31            3      4       1
2     placebo   11  30            3      1       2
210   placebo   11  30            5      2       2
```

3.5 library in R

You can call the packages and functions in the programming language R and activate them in work space and using the instruction library() , you can get a list of all the installed packages and identify the specific probability of the program R and the extent of the absorption of this version in the application of the functions and instructions you may need in the program, some important function has been called to apply all aspect of the study.[13]

4. Application side

4.1 The data

For the purpose of studying , the effect of sex and age on the people with epilepsy ,the data available in Basra Health Department in Basra Governorate in 2008 reaching 2296 patients :

```
> df<-read.table("f:/dat.txt",header=TRUE)
```

Where :

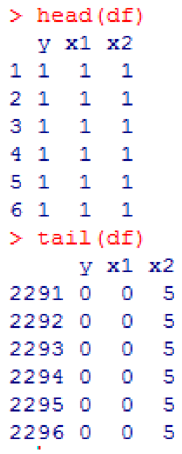
df : it is designated as a variable name to read and store data.

read.table : a function use to read a file type txt.

dat : is the name of a file that exists in the memory pane f contains a spread sheet consisting of row and columns,the row represent the number of cases which are 2296 and the columns are y,x1 and x2

header=TRUE : it means that the first row in the file contains the names of column and not the data.

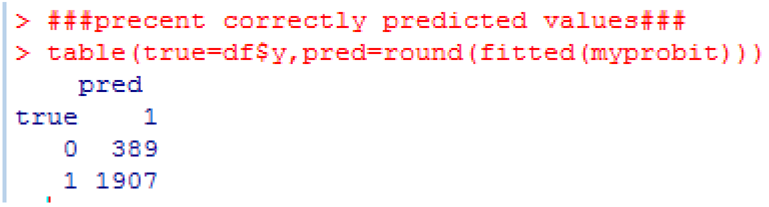
Because of the large data size we display using the functions head() and tail()for the beginning and the end of the data in the files as follows :



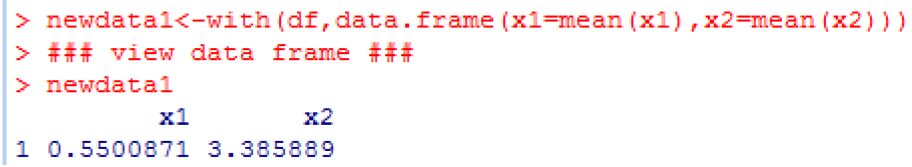
```
> head(df)
  y x1 x2
1 1  1  1
2 1  1  1
3 1  1  1
4 1  1  1
5 1  1  1
6 1  1  1
> tail(df)
     y x1 x2
2291 0  0  5
2292 0  0  5
2293 0  0  5
2294 0  0  5
2295 0  0  5
2296 0  0  5
```

4.2 Descriptive Statistics

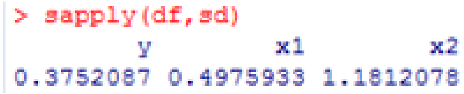
Below are some simple descriptive statistics such as percentages , averages and standard deviation for all variables under study , in R language we use the function mean() to calculate the mean and the function sd() to calculate the standard deviation , and by calling the library library('scales') , the percentage can be calculated by using the function percent().[14]



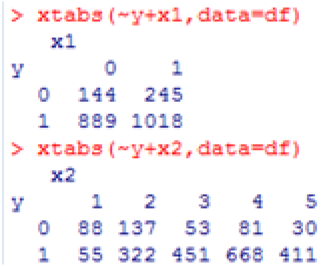
```
> ###precent correctly predicted values###
> table(true=df$y,pred=round(fitted(myprobit)))
    pred
true    1
   0  389
   1 1907
```



```
> newdata1<-with(df,data.frame(x1=mean(x1),x2=mean(x2)))
> ### view data frame ###
> newdata1
         x1       x2
1 0.5500871 3.385889
```



```
> sapply(df,sd)
        y        x1        x2
0.3752087 0.4975933 1.1812078
```



```
> xtabs(~y+x1,data=df)
   x1
y     0    1
  0 144  245
  1 889 1018
> xtabs(~y+x2,data=df)
   x2
y     1   2   3   4   5
  0  88 137  53  81  30
  1  55 322 451 668 411
```

Table(1) statistical indicators of variables

variable	No .of cases	percentages	Mean	S.d
y:dependent variable			0.831	0.375
(inpatient) :0	389	16.9		
(outpatient) :1	1907	83.1		

Total	2296	100.0		
x1 : sex			0.55	0.498
female :0 male :1	1033 1263	45.0 55.0		
x2 : Classes of age			3.386	1.181
1=less than 5 years 2=from 5 to 9 year 3=from10 to14 year 4=from15 to44 year 5= 45and more Total	143 459 504 749 441 2296	6.2 20.0 22.0 32.6 19.2 100.0		

Source : by the researcher based on the study data

Table(2)descriptive statistics of relationship of disease cases to factors affecting it

patient's condition factor influencing it	Y=0	Y=1
x1 : sex		
female :0 male :1	144 245	889 1018
x2 : Classes of age		
1= less than 5 years 2= from 5 to 9 year 3= from10 to14 year 4= from15 to44 year 5= 45and more Total	88 137 53 81 30 2296	55 322 451 668 411 100.0

Source : by the researcher based on the study data

Table (1) shows the percentage of cases of the disease for this sample, the percentage of the patient within the health unit (inpatient) (16.9) per 1000 cases, and the proportion of the patient outside the health unit (outpatient) (83.1) per 1000 cases. The table also shows the descriptive statistics of the variables, with the highest proportion of cases of the disease occurring in the fourth age group (44-15) is (32.6%), and shows that almost half of the infected sample of the study of both sexes, ie (45%) of the cases of the sample selected from Females, and (55) cases of males. Table (2) presents the factors affecting epilepsy ,both according to his condition (inpatient or outpatient).

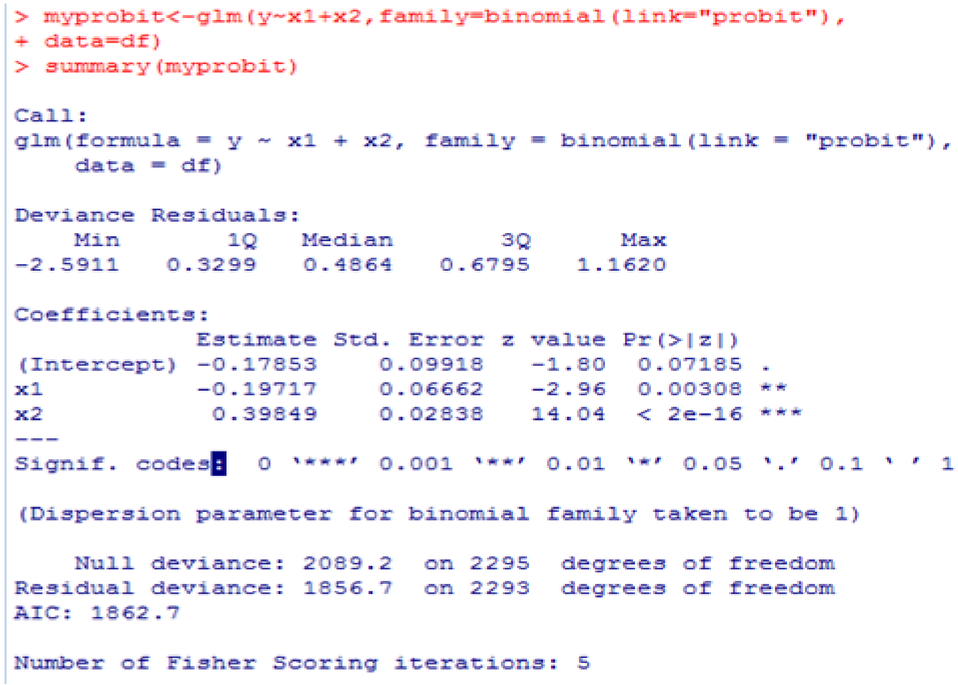
4.3 Analysis using R

R language in which the written commands are implemented directly without the used to build a complete program as in the case in most programming languages such as (C, Fortran , pascal), in addition , the syntax in R is very simple and intuitive. The following is a regression function for each probability regression models (Probit , tobit , and logit)and as follows:[15]

First : Probit regression[16]

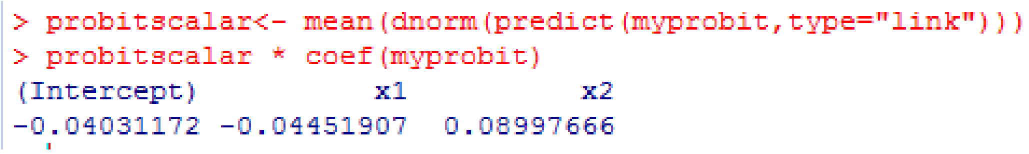
In this model using the function glm() with family=binomial(link="probit") the code below estimates the probit regression model using the function glm (generalized linear model), where the model output is stored in "myprobit" object ,then we can display the details of the result of this calculation using the function summary() which is a general one obtain a summary of the model and for all estimates.

> ### build the probit model ###



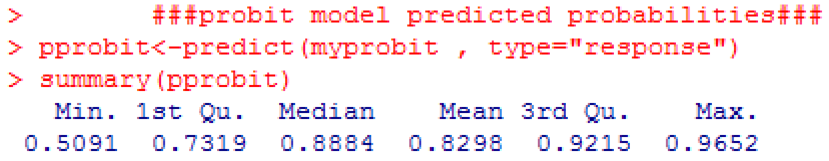
```
> myprobit<-glm(y~x1+x2,family=binomial(link="probit"),
+ data=df)
> summary(myprobit)

Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "probit"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5911   0.3299   0.4864   0.6795   1.1620

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.17853    0.09918   -1.80  0.07185 .
x1          -0.19717    0.06662   -2.96  0.00308 **
x2           0.39849    0.02838   14.04  < 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: 2089.2  on 2295  degrees of freedom
Residual deviance: 1856.7  on 2293  degrees of freedom
AIC: 1862.7

Number of Fisher Scoring iterations: 5
```

> ### probit model average marginal effects ###

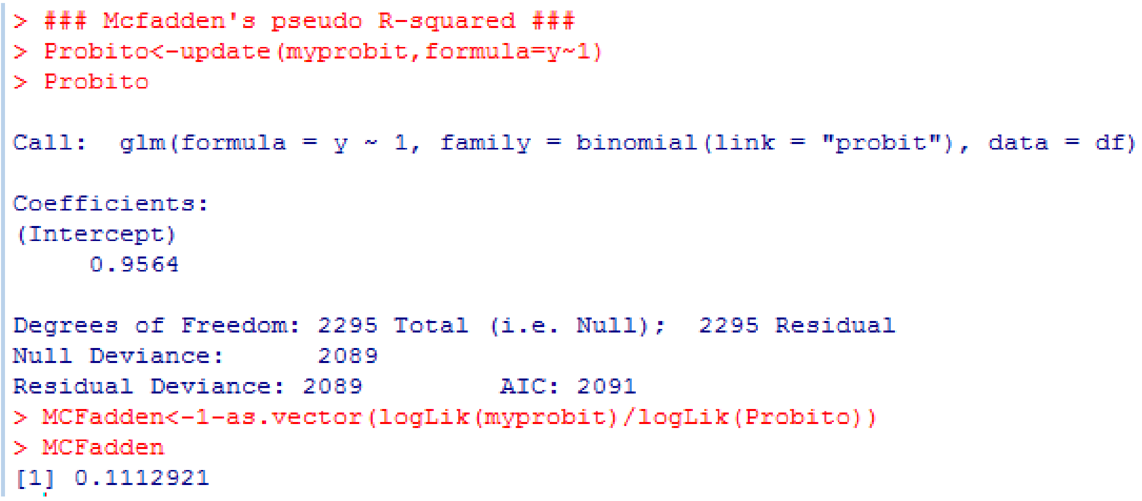


```
> probitscalar<- mean(dnorm(predict(myprobit,type="link")))
> probitscalar * coef(myprobit)
(Intercept)          x1          x2
-0.04031172 -0.04451907  0.08997666
```

> ### probit model predicted probabilities###

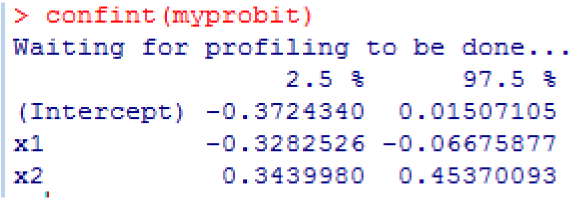


```
>        ###probit model predicted probabilities###
> pprobit<-predict(myprobit , type="response")
> summary(pprobit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5091  0.7319  0.8884  0.8298  0.9215  0.9652
```



```
> ### Mcfadden's pseudo R-squared ###
> Probito<-update(myprobit,formula=y~1)
> Probito

Call:  glm(formula = y ~ 1, family = binomial(link = "probit"), data = df)

Coefficients:
(Intercept)
     0.9564

Degrees of Freedom: 2295 Total (i.e. Null);  2295 Residual
Null Deviance:      2089
Residual Deviance: 2089         AIC: 2091
> MCFadden<-1-as.vector(logLik(myprobit)/logLik(Probito))
> MCFadden
[1] 0.1112921
```

To obtain confidence intervals for Parameter estimates



```
> confint(myprobit)
Waiting for profiling to be done...
                 2.5 %      97.5 %
(Intercept) -0.3724340  0.01507105
x1          -0.3282526 -0.06675877
x2           0.3439980  0.45370093
```

The results of probit indicate the high morale of the parameters of the model. This is confirmed by the z test according to the level of significance of the values, which is less than (0.05) or (0.01), clearly indicating the importance of these factors (age and sex) in the case of the patient (inpatient or outpatient) The value of R^2 is 0.956. This indicates that the variance in the patient's condition has been explained by the age and gender variable, meaning that the Probit model is suitable for the data.As for the confidence limits, the upper limit of the age parameter is -0.067and the minimum is -0.328 , which is limited between (-0.328, -0.067), while the gender parameter is limited between 0.344 and 0.454.

```
> ### The graph ###
> plot(myprobit,which=1)
```

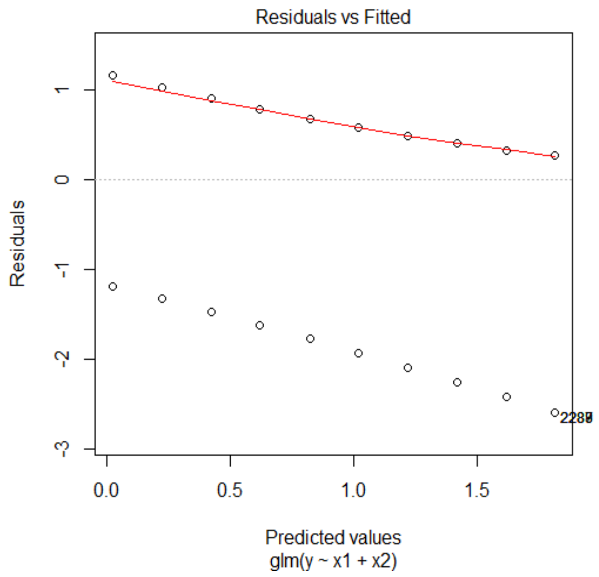


Figure (1) Expected compositions Random Limit

```
> plot(predict(myprobit),residuals(myprobit))
> abline(h=0,lty=2,col="green")
```

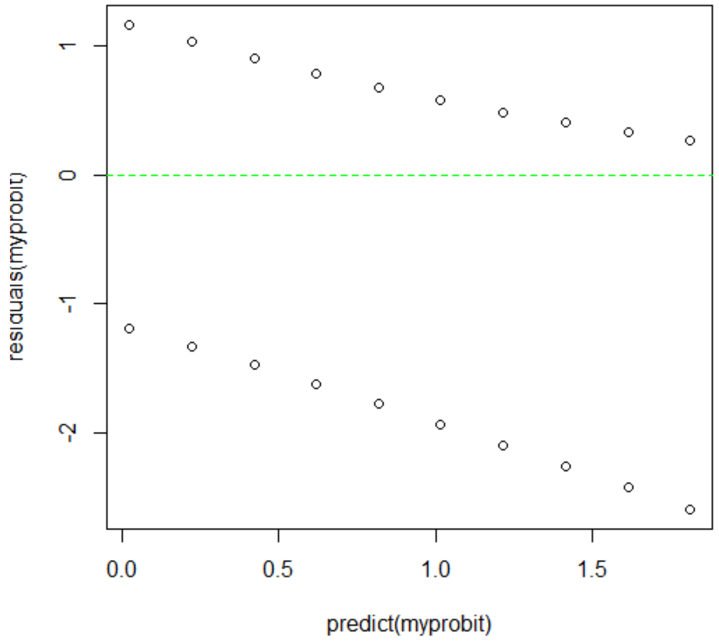


Figure (2) Distribution of the random limit according to the formula probit

```
> plot(predict(myprobit),residuals(myprobit),col=c("blue","red")[1+df$y])
>  abline(h=0,lty=2,col="green")
> lines(lowess(predict(myprobit),residuals(myprobit)),col="black",lwd=2)
```

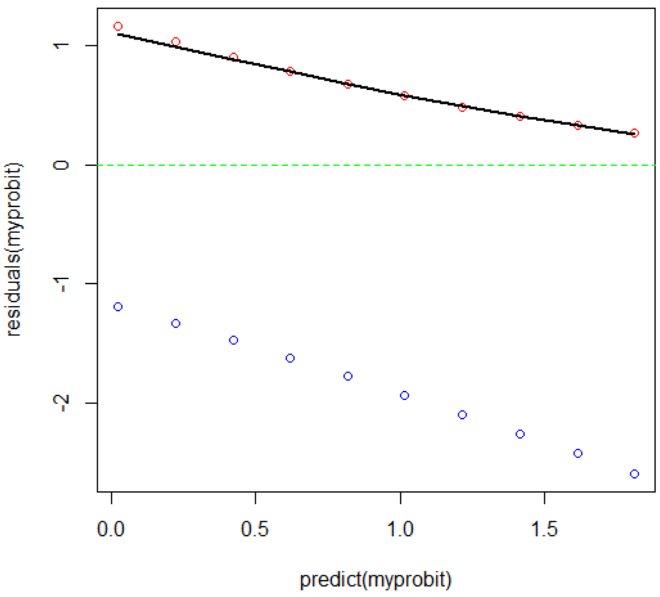


Figure (3) Connect the values in a straight line to distribute the random limit according to the probit formula

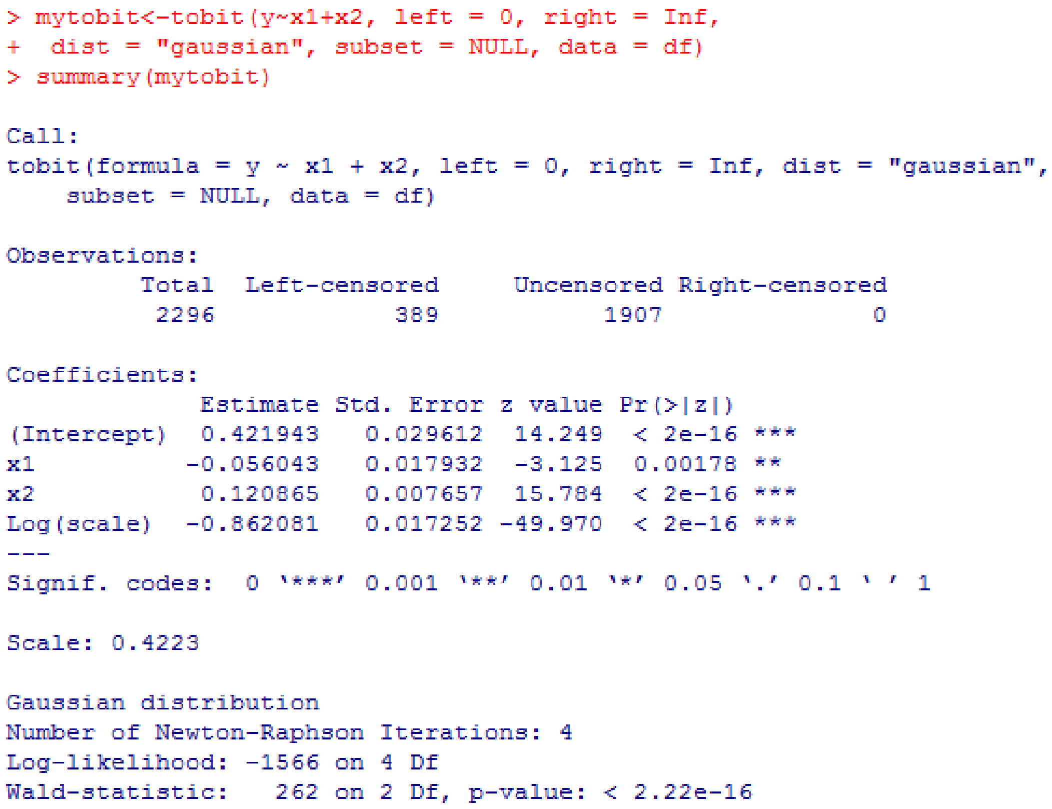
Second : Tobit regression [17]

We can use the tobit function of the package "AER" , so we need to install AER package first and its own library by typing :

> install.packages("AER")

> library(AER)

To display the result ,we use the function for Tobit regression then the function summary() , and to get a summary for the model as follows :



```
> mytobit<-tobit(y~x1+x2, left = 0, right = Inf,
+  dist = "gaussian", subset = NULL, data = df)
> summary(mytobit)

Call:
tobit(formula = y ~ x1 + x2, left = 0, right = Inf, dist = "gaussian",
    subset = NULL, data = df)

Observations:
         Total  Left-censored     Uncensored Right-censored
          2296            389           1907              0

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.421943   0.029612  14.249  < 2e-16 ***
x1          -0.056043   0.017932  -3.125  0.00178 **
x2           0.120865   0.007657  15.784  < 2e-16 ***
Log(scale)  -0.862081   0.017252 -49.970  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale: 0.4223

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -1566 on 4 Df
Wald-statistic:   262 on 2 Df, p-value: < 2.22e-16
```

We note the results of tobit , the high level of the parameters of this model according to the z test since the level of significance corresponding to the values, which is less than (0.05) or (0.01), indicating the importance of these factors (age and sex) in the case of the patient being (inpatient or outpatient) and This is confirmed by the wold test with a value of (262) for a level of significance less than the predefined level of significance

> ### The graph ###

> plot(predict(mytobit),residuals(mytobit))

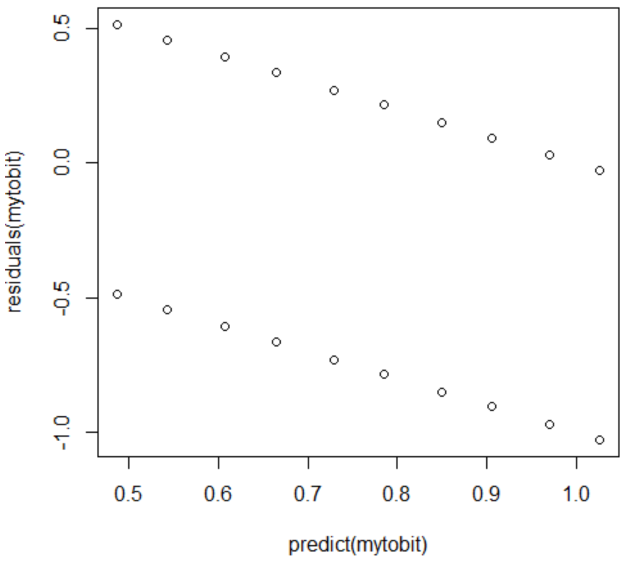
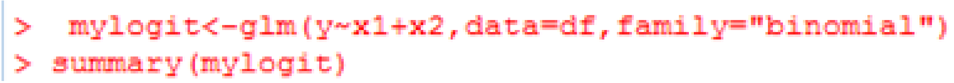
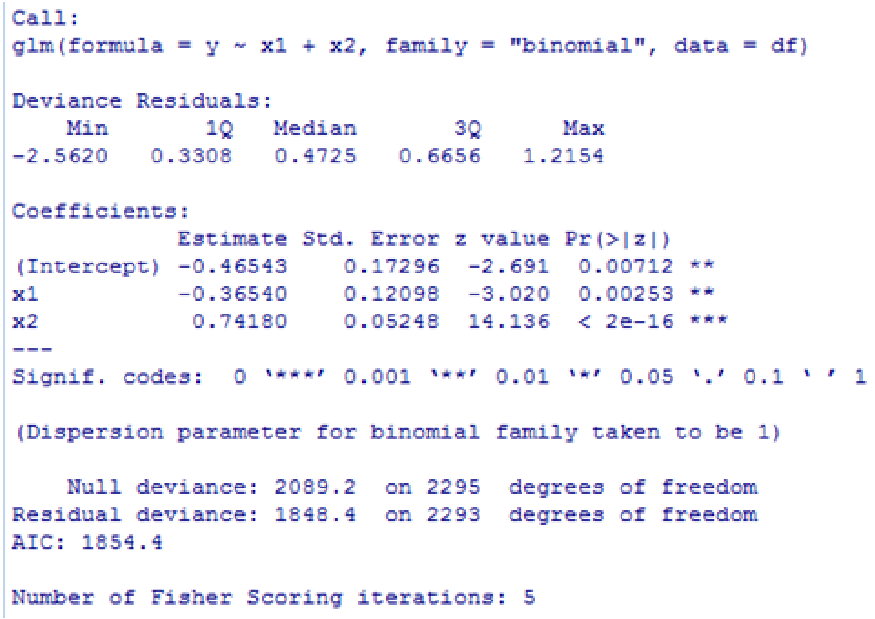


Figure (4) Distribution of the random limit according to the formula tobit

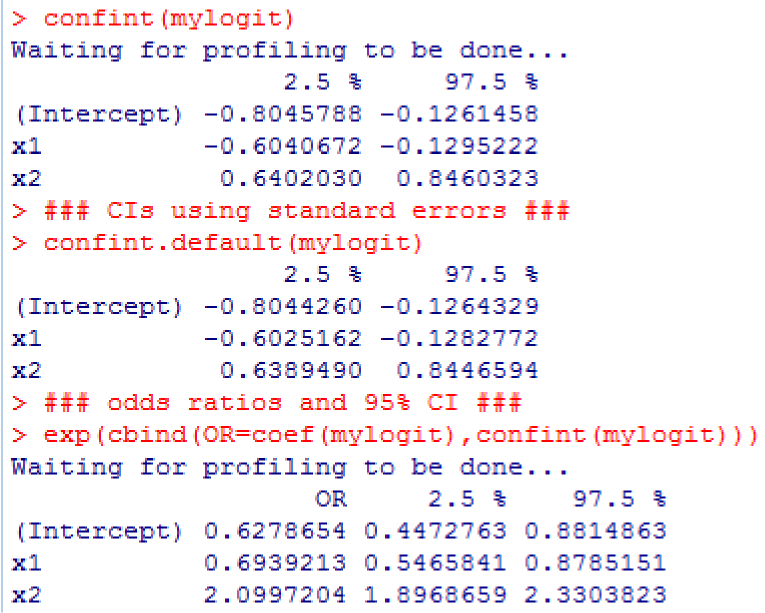
Third : Logit regression [18]



```
>  mylogit<-glm(y~x1+x2,data=df,family="binomial")
> summary(mylogit)
```



```
Call:
glm(formula = y ~ x1 + x2, family = "binomial", data = df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.5620   0.3308   0.4725   0.6656   1.2154  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.46543    0.17296  -2.691  0.00712 ** 
x1          -0.36540    0.12098  -3.020  0.00253 ** 
x2           0.74180    0.05248  14.136  < 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: 2089.2  on 2295  degrees of freedom
Residual deviance: 1848.4  on 2293  degrees of freedom
AIC: 1854.4

Number of Fisher Scoring iterations: 5
```



```
> confint(mylogit)
Waiting for profiling to be done...
                 2.5 %      97.5 %
(Intercept) -0.8045788 -0.1261458
x1          -0.6040672 -0.1295222
x2           0.6402030  0.8460323
> ### CIs using standard errors ###
> confint.default(mylogit)
                 2.5 %      97.5 %
(Intercept) -0.8044260 -0.1264329
x1          -0.6025162 -0.1282772
x2           0.6389490  0.8446594
> ### odds ratios and 95% CI ###
> exp(cbind(OR=coef(mylogit),confint(mylogit)))
Waiting for profiling to be done...
                   OR     2.5 %    97.5 %
(Intercept) 0.6278654 0.4472763 0.8814863
x1          0.6939213 0.5465841 0.8785151
x2          2.0997204 1.8968659 2.3303823
```

The logit results also show the high morale of the model parameters based on the z test and according to the moral level of the values, the value is less than (0.05) or (0.01), indicating clearly the importance of these factors (age and sex) in the case of the patient (inpatient **or** outpatient) The maximum age parameter is (-0.128) and the minimum is (-0.603), which is limited between (-0.603, -0.128), and the sex parameter is limited between (0.639, 0.845)

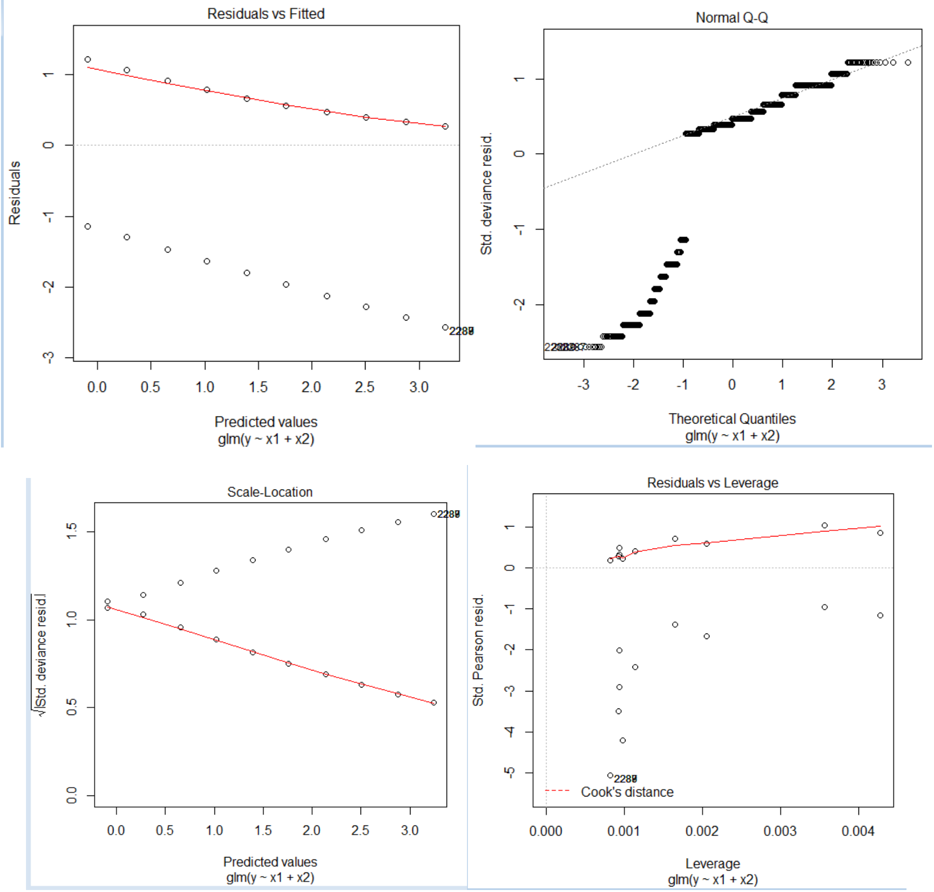


Figure (5) forms to distribute the random limit according to the formula of logit

5. Discussion

The study of the problem of epilepsy has a clear effect on many aspects of health, social and economic, through the data of (2296) patients, and applied probabilistic regression models of probit, Tobit and Logit models for their ability to analyze data and the effect of age and gender variables of the disease, The results show that the age factor and the sexual factor have a significant impact on the disease, according to the Wold test. , Which was confirmed by the selection coefficient as it reached 0.96**.**

6.. Conclusions

1. The wide scope of the R program in biomedical statistics, allows researchers to analyze and process their data with much more flexibly than many other statistical programs, and contains packages to run many statistical analyzes, and show the other packages data for the same disease to other countries.

2 - Although mental illness are important and sensitive, but there is a lack of data for each disease of mental disorders and scarcity in scientific studies (medical and statistical) in Iraq as a whole and Basra in particular.

3. The R program is a famous free program in the field of statistics and display and analysis, data , allowing scientists to classify infectious diseases and conduct some simple analyzes common in biomedical statistics.

4. The probit regression versus logit has similar results and largely depends on individual details

5 . The results of the regression of probabilistic function showed that both the factor of age and sex factor have a very significant effect in cases of disease in terms of stagnation or review.This is confirmed by the wold test. The results showed that these functions represented the best representation, and this was confirmed by the coefficient of selection at 0.96.

6. Although simulation using the R program allows the analysis of statistical data to obtain the results of data generated randomly and that the data entry is only for a few numbers in addition to easy access to online support, but we find in this study for a medical condition and the use of real data is best to give realistic results for the diagnosis of epilepsy and the factors affecting the patients in Basra

7. Recommendations

1- There are dozens of versions of the programming language R on the web pages, where you can download the required version, which includes the functions used in the study as well as libraries as needed, which can be known from the packages available and freely online

2. In advanced study, R program can be used to study the possibility of the disease for people who have not yet been affected by this disease, or to apply the possibility of epilepsy or any other disease when taking a treatment and the extent of response to a certain type of treatment and in dose and response analysis.

3. The most important R program today is after the SAS and SPSS programs for ease of use, being free, and the graphics are the best. As well as its size is small and can be loaded on any operating system, and the programming language R is one of the best languages in the statistics and informatics of the programs that are ready to implement most of the problems and statistical analysis.

4 - conducting a broad study includes most factors that affect the cases of disease

References

[1] Y. C. Yves Croissant and A.Z. Achim Zeileis ,Package 'truncreg, 2016 http://R-Forge.R-project.org/projects/truncreg/

[2] World Health Organization ,NEUROLOGICAL DISORDERS public health challenges, 2006 ,Geneva, Switzerland

[3] Karen L. Parko, M.D., Seizures and Epilepsy Diagnosis and Treatment,2011, San Francisco VA Medical Center.

[4] Brian S. Everitt and Torsten Hothorn, Title A Handbook of Statistical Analyses Using R (1st Edition),2017, Version 1.3-9
.

[5] K.S. Kim Seefeld, MS, M.Ed. and E.L. Ernst Linder, Ph.D. , University of New Hampshire, Durham, NH , Department of Mathematics & Statistics, Statistics Using R with Biological Examples ,2007.

[6] G.V Grant V. Farnsworth ,Econometrics in R, 2008.

[7] Greene W (2004). "Fixed Effects and Bias Due to the Incidental Parameters Problem in the Tobit Model.", Econometric Reviews, 23(2), 125-147.

[8] Okoh-Esene, R.U, National Institute for Pharmaceutical Research and Development Idu, F.C.T. O.J. Okogun J.I., Chemistry Department, University of Abuja, F.C.T , O.S. Okwute S.K. And T.S. Thomas S.A. Sheda Science and Technology Complex (SHESTCO), P.M.B.186,Garki Abuja F.C.T., An Overview of the Facts, Myths and Treatment of the Disease Condition Known as “Epilepsy”,2013.

[9] Dekker, P.A. , epilepsy , Amanual for medical and clinical officers In Africa ,2002, world health organization ,Geneva.

[10] W.R. William Revelle , How To: Install R and the psych package,2017 Department of Psychology , Northwestern University .

[11] Maronna, Martin and Yohai, Wiley ,Robust Statistics, Theory and Methods,2006, https://CRAN.R-project.org/package=robustbase

[12] - Brian S. Everitt and Torsten Hothorn , A Handbook of Statistical Analyses Using R (1st Edition),2006 , https://CRAN.R-project.org/package=HSAUR

[13] F.L. Friedrich Leisch , Creating R Packages: A Tutorial, 2009, Department of Statistics, Ludwig-Maximilians-Universit□at M□unchen, and R Development Core Team, Friedrich.Leisch@R-project.org

[14] J.N. Jackie Nicholas ,(2006) ,"Introduction to Descriptive Statistics" ,Mathematics Learning Centre , University of Sydney ,NSW

[15] Oscar Torres-Reyna, Logit, Probit and Multinomial Logit models in R,2014 ,(v.3.3), PRINCETON UNIVERSITY, http://dss.princeton.edu/training/

[16] Natalie Austin ,The Probit Model Alexander Spermann University of Freiburg SoSe , 2009, University of Freiburg.

[17] McDonald, J. F. and Moffitt, R. A. , The Uses of Tobit Analysis. *The Review of* Economics and Statistics,1980 , Vol 62(2): 318-321.

[18] S.P. Stephen Pettigrew , Logit Regression and Quantities of Interest ,2014.

.

.

حل نماذج الانحدار الاحتمالية باستخدام بيئة برمجة R

دراسة حالة للمرضى الصرع

أ.م. ندى بدر جراح

جامعة البصرة / كلية الادارة والاقتصاد/ قسم الاحصاء

الاختصاص العام : علوم حاسبات

الخلاصة

إن تكنولوجيا التعامل مع الحسابات العلمية تتقدم بسرعة كبيرة، وتشكل زخما كبيرا للتقدم العلمي. وقد أصبح بعض مستوى إتقان R، بالنسبة لكثير من التطبيقات، ضرورية للاستفادة من هذه التطورات.

توفر بيئة البرمجة الاحصائية R منصة مثالية لاجراء الدراسة لما تتمتع به من قدرة برمجية قوية ، والرسومات البيانية، ومجموعة شاملة من الوظائف الإحصائية المفيدة ويحتوي على أكثر من 11164 حزمة حيث تم في هذا البحث حل

نماذج الانحدار الاحتمالية والمتمثلة بنماذج Probit ، Tobit و Logit إلى جانب تحديد متغيرات الدراسة (العمر والجنس) التي تؤثر او تزيد من الاصابة بداء الصرع من خلال بيانات لـ (2296) مريض في البصرة .

وبعد تقدير نماذج الانحدار الاحتمالية الثلاثة لمعرفة تأثير عاملي العمر والجنس في احتمال الاصابة بالمرض باستخدام لغة البرمجة R اظهرت نتائج انحدار دوال الاحتمالية ان كل من عامل العمر وعامل الجنس لهما تأثيراً معنوياً عالياً في حالة المرض من حيث كونه راقد او مراجع. وهذا ما اكده اختبار wold . كما ان هذه الدوال مثلت البيانات خير تمثيل وهذا ما أكده معامل التحديد اذ بلغ 0.96.

الكلمات المفتاحية : R language , Statistical Programming , probability regression , Epilepsy