Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.2 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

Use of logistic regression to study the most important factors affecting the incidence of tuberculosis

Hayder Raaid Talib
Department of Statistics
College of Administration and Economics
University of Sumer
haider.r.t86@gmail.com

Ahmed Razzaq Abd
Department of Statistics
College of Management and Economics
University of Wasit
ahmedrazzaq@uowasit.edu.iq

Ali Hamdullah Ahmed
Directorate – General for Education of Qadisiyah
Department of Educational Planning
ali_sn2007@yahoo.com

Recived : 11\12\2017 Revised : 16\1\2018 Accepted : 4\2\2018

Available online : 25 /2/2018

DOI: 10.29304/jqcm.2018.10.2.374

Abstract :

What distinguishes this research Is to focus on the concept of logistic the regression and its characteristics and how to build the model analytical descriptive approach was used to describe the logistic regression How to estimate Its Features And its use as an important tool in studying the influencing factors in the injury of tuberculosis the sample included 299 patients.

The first topic: methodology of research

1.1 Introduction: -

The models of logistic regression have been noticed in many years. They are widely used in life experiences, which are among the main concerns of the countries of the world for their relation to human life and development through finding the best way to provide the best services in the fields of pharmaceuticals, vaccines, vitamins, pesticides, hormones and others. The method of analyzing the path of efficient statistical methods in the analysis of data may enable the researcher to identify and clarify the potential negative relationships of a group of factors and to determine the direct and indirect impact of the phenomenon of the total study and thus helps to derive logical explanations of the phenomenon and more efficient in the analysis of data. The importance of logistic regression analysis is highlighted by the ability to study the effect of several factors on a particular phenomenon in a direct or indirect manner and we use logistic regression in the response variables.

2.1 Research Objective

There are two main objectives of research : The first objective is represented by the concept of logistic regression and the methods of estimating the parameters of this model.

The second objective is to apply the logistic regression in the health aspect of the tuberculosis phenomenon to identify the most important causes of this disease

3 .1 Research sample

This study was applied to data collected from Al-Hussein Hospital through tests of chest diseases for each patient. A group of doctors specializing in chest diseases was used to classify the most important factors affecting the disease. A sample of 299 patients was taken from the center of the chest 2017/2/3 and until 2017/5/1

4.1 Research Hypotheses

There are several hypotheses for the logistic regression model, including:

1-Variable period (0.1) variables can be continuous or discreat or binary versions or multiple.

2 - There is a relationship between the dependent variable and explanatory variables take the following pictures:

$$\mathrm{pr}(Y_k = {}^{1}/_{x}) = \frac{\exp(B\acute{X})}{1+\exp(B\acute{X})} \qquad \ldots (1)$$

3-There is no correlation between random errors (independence of errors)

4 - There is no correlation between random error and explanatory variables

5 - There is no correlation between the explanatory variables with each other in a complete way if the variables that have a relationship between them should be correlation completely.

6- The random variable (y_i) is assumed to be distributed by Bernoulli y ~B(θ) in the mean E (y_i)= θ and variance V (y_i)= $\theta(1-\theta)$.

7-The expected value of the random error is zero, since Pr (x) [1-pr (x)] and the variance of the random error U_n expands based on the bar

The second topic: Theoretical side.

1-2 Logistic Regression[2][3] [5]

It is common in human, social and economic studies that the dependent variable is a separate variable to take the fixed value or more. This is a significant criterion for the researchers of their attempts to employ linear regression analysis (simple or multiple), which is somewhat restricted by requiring that The dependent variable is a quantitative variable that is connected rather than descriptive and separate.

Quantitative models are important models in natural, engineering and social studies. The choice of an appropriate model depends on the nature of the data, especially when the variable of response is binary, ie, the occurrence of the phenomena, such as death or life. This response is affected by the existence of a set of independent variables that affect the response variable. The nonlinear models are among the best models representing such phenomena. The best models are the probability model and the logistic model, which is the focus of our research. Logistic model of commonly used nonlinear models.

(Lea 1997), which is seen in such cases, although there are many statistical methods developed to analyze data with descriptive variables (qualitative)

(Function Analysis Disctianant) such as analysis of the functions of excellence

However, logistic regression has many features that make it suitable for use in such situations. Logistic regression is a useful way to illustrate the relationship between independent variables (age, sex, etc.) and the variable of the answer, the probability, which takes two different values.

An example of a cancer diagnosis is that the two values for the response variable are either infected or not.

The importance of logistic regression when compared with other statistical methods (linear regression and differential analysis) is that logistic regression is a more powerful tool to provide a test of the significance of transactions. It also gives the researcher an idea of the effect of the independent variable in the binary dependent variable.

In addition, the logistic regression calculates the effect of independent variables, allowing the researcher to conclude that a variable is stronger than the other variable in understanding the emergence of the desired result, and that regression analysis can include independent qualitative variables as well as the interaction of independent variables in the dependent variable The advantage of using logistic regression is that it is less sensitive to deviations from the normal distribution of the study variables, compared to other statistical methods such as differential analysis and linear regression, and the logistic regression Biddable exceed many restrictive assumptions to use the method of least squares (OLS) in the linear regression which makes the ultimate logistic regression analysis the best method in the case of the binary variable value. (Binary Logistic Regression), which we use in our multinomial logistic research used in the case of the multivalent nominal variable (more than two values).

There is also a third type of logistic regression called ordinal logistic regression, which is used In cases where the dependent variable is a class variable, we use the two-valued dependent variable in values (0,1) without any other coding.

There are also several definitions of the regression model:

The logistic regression model can be defined as a model used to predict the probability of an event by fitting the data on the logistic curve. Logistic regression uses several expected variables that can be numerical or factional. For example, logistic regression in marketing is used to calculate the consumer's tendency to buy a product or refrain from purchasing. Logistic regression is used broadly or broadly in medicine and social sciences.

It is also possible to define the statistical method used to examine the relationship between the dependent variable and one or more independent variables, that is, the binary-value variable and one or more independent variables of any kind called Binary Logistic Regression.
It is also known as the type of regression used to predict the value of dependent binary or class dependent variables depending on the set of independent variables mixed, such as continuous variables or measurements, and the other section in the form of intermittent qualitative or class variables.

2-2 Characteristics of the logistic regression model[1][5]

1.This model does not put any preconditions on the explanatory variables
2.The model does not specify which vectors belong to the new observations, but also determines the probability of this affiliation. It can also be used to analyze the binary and multivariate descriptive variable.
3.The maximum possible method (ML) is used to estimate its parameters and therefore the quality conditions are met in these variables
4.ease of calculations used in the form Model.
With these characteristics, logistic regression becomes one of the most appropriate models for the analysis of binary and multivariate descriptive variable .

3-2 Method of building the regression model[1][3]

1.Achieving the relationship between the binary nominal variable and the nominal independent variable by means of a single analysis using the Chi-square and the correlation test
2.To achieve the correlation between the variable logarithm of the binary nominal and the independent quantitative constant variable by the scheme of dispersion between the two variables where the relationship must be positive
3.Analysis of the linear relationship between independent variables

2-4 Analysis of simple and multiple logistic regression[2][4]

The analysis of logistic regression is used in epidemiological and medical studies, in which the quantitative and qualitative independent variables that affect the probability of the resulting variable are determined when the logistic regression is applied.

2-4-1 Required for simple logistic regression[1][3]

Quantitative or nominal independent variable such as weight, height, marital status and gender. It is a binary variable, such as a disease (yes, no), gender (male, female) and others.

2-4-2 Required in multiple logistic regression[2][6]

Two or more independent quantitative or nominal variables such as weight, height, marital status and gender. One variable my name is a dual-type follower such as an illness (yes, no) or gender (male, female) and others.

2-5 Logistic regression model[2][4][6]

The logistic regression model is defined as one of the regression models where the relationship between the y variable and the explanatory variables (x_1, x_2, ..., x_k) is nonlinear where the y variable of the binary response takes the values 0,1 and the success is the probability of __i Or failure Failure to respond to the probability of 1-__i
so the dependent variable y follows the Bernoulli distribution and the probability density function will be as follow:

$$P(Y = Y_i) = \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \quad . . . \quad (2)$$

$y_{i=0,1}$ Since
Y_i : binary variable dependent response
θ_i : The probability of a response when $y_{i=}1$
Therefore, signing the dependent variable represents the probability of a response

$$E(y_i) = p(y = 1) = \theta_i \quad . . . \quad (3)$$

The variance of the dependent variable by Bernoulli distribution is

$$V(y_i) = \theta_i (1 - \theta_i) \quad . . . \quad (4)$$

Let $X_\circ, X_1, X_2, \ldots\ldots X_p$ be a set of explanatory variables and let n have many observations of these variables that are matrix X

$$X = (X_{ij})nxp \quad . . . \quad (5)$$

(n) i = 1,2 .... n represents the sample size . (P) J = 0.1,2 .... P represents the number of parameters
If $y_{i=}[y_1, y_2, \ldots . y_n]$ is a random sample of the binary response variable and $y_i \in \{0,1\}$ This leads to a regression model given as follows

$$y_i = \theta_i + \varepsilon_i \quad . . . \quad (6)$$

represents the regression function (probability of response)

$$\theta_i = p(y = 1) = \frac{e^{X_i \underline{\beta}}}{1 + e^{X_i \underline{\beta}}} \quad . . . \quad (7)$$

β: vector of information dimensions (p * 1)
Xi = {X_{i0}, X_{i1}, ⋯ X_{ip}} A class vector of explanatory variables
(1 * p):
ε_i: The error limit that will have an average of zero as in the following formula:

$$\varepsilon_i = y_{i-} \theta_i$$

$$E(\varepsilon_i) = E(y_i) - E(\theta_i) = \theta_i - \theta_i = 0 \quad . . . (8)$$

Either the error limit variance is equal to the variance of the adopted binary response variable.

$$V(\varepsilon_i) = V(y_i) = \theta_i (1 - \theta_i) \quad . . . (9)$$

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.2 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

The error threshold follows the Bernoulli distribution with an average of 0 and the variance (θ_i-1) θ_i
It is noted that the variance of the error limit depends on the values of the response probability θ_i i.e. on vector values X_i and there are Variance of error non homogenoce .

2-6 Methods for estimating the parameters of the logistic regression model

To estimate the parameters of the logistic regression model, it was based on the weighted low squares method (WLS), which is used in regression analysis to address some analysis problems. The maximum possible method (MLE) is also used, and the maximum possible method requires repetitive methods for its calculation. Therefore, initial values of β_1 and β_0 are required.
Before analyzing these methods, it is important to know that binary data such as success and failure appear in most areas of study. The analysis of log regression is most often
used to examine the relationship between intermittent responses and total explanatory variables. There are those who discussed logistic regression (Agresti 1990, 1989) Cox & Snell) Since in binary data the distribution of random error (U_i) is sporadic, it is an abnormal distribution, but a binomial distribution is distributed. If the error is distributed naturally, the variance is not equal in all aggregates. In the case of discrete response models, the model is as follows:
$Y_i = \beta_\circ + \beta_1 X_{i1} + \ldots\ldots + \beta_k X_{ik} + U_i \quad . \ . \ . \quad (10)$
X_i : independent Variable .
$\beta_\circ , \beta_1 , \beta_k$Regression equation parameters. U_i: is a random error
Y_i : dependent variable
Since the variable (Y_i)distribution of any Bernoulli there are only two values, so the random distribution of error (U_i) distribution is sporadic and does not continue any
$Y_i = 1 \ ; \ Y_i = 0 \ ; \ E(Y_i) = P_i \ ; Q_i = 1 - P_i$
While
P_i: Response Ratio When $Y_i = 1$; Q_iResponse Ratio When $Y_i = 0$
$\mathrm{Var}(U_i) = E(U_i)^2 - [E(U_i)]^2 \quad . \ . \ . \quad (11)$
The compensation shall be $\mathrm{Var}\,(U_i) = P_i\, Q_i \quad . \ . \ . \quad (12)$
Therefore, the Variance error in this case is nonhomogeneous and the ordeal least square method cannot be used to estimate the parameters of the linear model in the case of dichotomous binary data distributions.

2-6-1 Weighted least squares method (WLS)[1][3][6]

In WLS, the relationship between the dependent variable (Z) and the independent variables is as follows:
$Z = X\beta + U \ldots (13)$ Where as
U: A vertical direction Rank (r * 1) represents random errors
X: matrix of independent variables of order [r * (k + 1)]
β: Parameters of vector [(k + 1) * 1] ; Z: vector axis of rank (r * 1)
Since the expectation and variance of variable Z is

$$E\,(Z_i) = X\beta \quad . \ . \ . \quad (14)$$

$$\mathrm{var}\,(Z_i) = \frac{1}{n_i p_i q_i} \quad . \ . \ . \quad (15)$$

$$Z \sim N\left(X\beta, \frac{1}{n_i p_i q_i}\right)$$

The estimated formula of the model parameters can be derived by multiplying the two ends of the model by the inverse of the symmetrical country matrix (p), which is at the Ranke of (r * r). We obtain the following model

$$P^{-1}Z = P^{-1}X\beta + P^{-1}U \quad . \ . \ . \quad (16)$$

$$P = \begin{bmatrix} \frac{1}{\sqrt{n_1 p_1 q_1}} & . & . & 0 \\ . & . & 0 & . \\ . & 0 & . & . \\ 0 & . & . & \frac{1}{\sqrt{n_r p_r q_r}} \end{bmatrix}$$

We use the OLS method to find β estimates as it comes

$$U = Z - X\beta$$

$$U'U = (Z - X\beta)'(Z - X\beta)$$

And compensation

$$(P^{-1}U)'(P^{-1}U) = (P^{-1}Z - P^{-1}X\beta)'(P^{-1}Z - P^{-1}X\beta)$$

$$(P^{-1}U)'(P^{-1}U) = \left(Z'p'^{-1} - \beta'X'P'^{-1}\right)(P^{-1}Z - P'X\beta)$$

And unzip the brackets we get

$$(U'P'^{-1}P^{-1}U) = (Z'P'^{-1}P^{-1}Z - Z'P'^{-1}X\beta - \beta'X'P'^{-1}Z + \beta'X'P'^{-1}X\beta)$$

$$(U'W^{-1}U) = Z'W^{-1}Z'W^{-1}X\beta - \beta'X'W^{-1}X\beta$$

$$(U'W^{-1}U) = Z'W^{-1}Z - 2\beta'X'^{W^{-1}}Z + \beta'X'W^{-1}X\beta$$

Taking the first derivative for β we get:

$$\frac{\partial U'W^{-1}U}{\partial \beta} = -2X'W^{-1}Z + 2X'W^{-1}X\beta$$

On the basis of the first derivative of zero, we obtain:

$$X'W^{-1}Z = X'W^{-1}X\hat{\beta}$$

By multiplying the equation by $(X'W^{-1}X)^{-1}$ we obtain an estimate of the values of $\hat{\beta}'S$)

$$\hat{\beta} = (X'W^{-1}X)^{-1}X'W^{-1}Z \quad . . . \quad (17)$$

W=P'P

$$W^{-1} = \begin{bmatrix} n_1p_1q_1 & . & . & 0 \\ . & . & 0 & . \\ . & 0 & . & . \\ 0 & . & . & n_rp_rq_r \end{bmatrix}$$

) r*1 (Where : Z is a vertical direction bar

$$Z = \begin{bmatrix} \ln\frac{p_1}{1-p_1} \\ . \\ . \\ . \\ \ln\frac{p_r}{1-p_r} \end{bmatrix}$$

The weighted least squares method achieves estimates when the Pi (probability of response) value is zero or one, so the method is processed so that the value of the Zi variable is as follows:

$$Z_i^* = \ln\left[P_i + \frac{1}{2n_i} \Big/ q_i + \frac{1}{2n_i} \right] \quad . . . \quad (18)$$

$$\mathrm{Var}(Z_i^*) = \left[(n_i+1)(n_i+2) \Big/ n_i^3 \left(P_i + \frac{1}{n_i}\right)\left(q_i + \frac{1}{n_i}\right) \right]$$

$$\mathrm{Var}(Z_i^*) = \frac{1}{W_i^*}$$

W_i^* : New weighted weight

The estimates can be calculated as follows:

$$\hat{B} = \left(X'^{W^{*-1}}X\right)X'W^{*-1}Z^* \quad . . . \quad (19)$$

2-6-2 Maximum likelihood method MLM[2][5] [6]

This method is based on the finding of β values, which are estimates of vector β, which makes the function at its extreme end, and assuming that we have r of independent random variables $(Y_1, Y_2, \ldots, Y_r)$; (n_i, p_i) and (Y_i)that represents the sum of successes in each end of the function at (n_i) the end of and that (K) of the independent variables in each set of totals, the probability density function of (Y_i) :

$$P_i(Y_i = y_i) = C_{y_i}^{n_i} P_i^{y_i} (1-P_i)^{(n_i - y_i)} \quad . . . \quad (20)$$

i=1,2,….,r

y_i=0,1,2,….,n_i

$$E(y_i) = n_i p_i \quad . . . \quad (21)$$

$$\mathrm{var}(y_i) = n_i p_i (1 - p_i) \quad . . . \quad (22)$$

The p_i response rate is estimated as follows:

$$p_{i=} {}^{y_i}/_{n_i} \; ;$$

$$Q_i = 1 - p_i = 1 - y_i/n_i = (n_i - y_i)/n_i$$

The Maximum Likelihood function for the co-distribution of data (Yi) is by formula

$$L(P) = \prod_{i=1} C_{y_i}^{n_i} P_i^{y_i} (1-P_i)^{n_i - y_i} \quad . . . \quad (23)$$

$$L(P) = \prod_{i=1} C_{y_i}^{n_i} \left[\frac{p_i}{1-p_i}\right]^{y_i} (1-p_i)^{n_i} \quad . . . \quad (24)$$

$$\ln L(P) = \sum_{i-1}^{r} \left[\ln C_{y_i}^{n_i} + y_i \ln\left[\frac{p_i}{1-p_i}\right] + n_i \ln(1-p_i) \right] \quad . . . (25)$$

Since (pi) the probability of success is affected by the independent explanatory variables according to the logistic model (in the form of matrices) :

$$P_i = \frac{\exp(X'_i\beta)}{1+\exp(X'_i\beta)} \quad . . . \quad (26)$$

$$1 - P_i = \frac{1}{1+\exp(x'_i\beta)} \quad . . . \quad (27)$$

When compensating the formulas (26) and (27) in (25) we get the following formula:

$$L(P) = \sum_{i-1}^{r} \left\{ \ln C_{y_i}^{n_i} + y_i x'_i\beta + n_i \ln\left[\frac{1}{1+\exp(x'_i\beta)}\right] \right\} \quad . . . \quad (28)$$

We use the Newton Raphson method to find the *Maximum Likelihood* estimates according to the formula

$$t_{s+1} = t_s - G^{-1}g_{(s)} \quad . . . \quad (29)$$

t_{s+1} :represents the vector of β parameters to be estimated

t_s:The vector represents the initial values of the parameters

$g_{(s)}$:The first derivative vector of the logarithm represents the function Likelihood

$$g_{(s)} = \frac{\partial \ln L}{\partial \hat{\beta}} = \begin{bmatrix} \frac{\partial \ln L}{\partial \beta_\circ} \\ . \\ . \\ . \\ \frac{\partial \ln L}{\partial \beta_k} \end{bmatrix}$$

G: Matrix The expected value of the negative value of the second derivative of the logarithm is the Maximum Likelihood function

$$G=\left[\frac{\partial^2 \ln l}{\partial \widehat{\beta_\iota}\, \partial \widehat{\beta_\iota}}\right] = \begin{bmatrix} -E\frac{\partial^2 \ln l}{\partial \beta_\circ^2} & -E\frac{\partial^2 \ln l}{\partial \beta_\circ \partial \beta_1} & -E\frac{\partial^2 \ln l}{\partial \beta_\circ \partial \beta_k} \\ . & -E\frac{\partial^2 \ln l}{\partial \beta_1^2} & -E\frac{\partial^2 \ln l}{\partial \beta_1 \partial \beta_k} \\ . & . & -E\frac{\partial^2 \ln l}{\partial \beta_k^2} \end{bmatrix}$$

Thus, the formula for finding estimates is as follows:

$$\hat{\beta}_{s+1} = \hat{\beta}_s + \left[\frac{\partial^2 \ln l}{\partial \hat{\beta}_j\, \partial \hat{\beta}_j}\right]^{-1} \frac{\partial \ln L}{\partial \hat{\beta}_s} \quad . \ . \ . \quad (30)$$

The original Newton-Raphson formula in the case of the logistic model to find the Maximum Likelihood estimates is

$$\hat{\beta}_{s+1} = \hat{\beta}_s + (x'vx)^{-1}x'^{(y-\hat{y}_s)} \quad . \ . \ . \quad (31)$$

$\hat{\beta}_{s+1}$ *Vertical direction of the estimate values in the cycle (s+1) at grade ((k+1)*1)*

$\hat{\beta}_{s+1}$ Vertical direction of the estimate values in the cycle (s) at grade ((k+1)*1)

X: matrix of independent variables in the cycle (r((k + 1)*

*V: a diagonal matrix of the rank (r*r)*

$$V=\begin{bmatrix} n_1\hat{p}_1\hat{q}_1 & 0 & 0 & 0 & 0 \\ 0 & n_2\hat{p}_2\hat{q}_2 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & n_r\hat{p}_r\hat{q}_r \end{bmatrix}$$

Estimation of the vector of parameters (β) is discontinued when the difference from the previous cycle and the subsequent cycle is very small and approaches zero.

The third topic: The applied side

1-3 Description of data Description of data

The data of the research were collected from Al-Hussein Hospital in Dui Qasr Governorate through tests of chest diseases for each patient. A group of specialized doctors were used to classify the most important factors affecting the disease. A sample of 299 patients was taken from Al- 2017/2/3 and until 2017/5/1 are:

y: - The condition of the patient where he takes two values (infected 1) and (not infected 0) which represents the dependent variable.

X_1 : Age of patient

X_2 : patient sex (male = 1) and (female = 2)

X_3: Classification of the environment of the casualty (rural=1) and (city= 2)

: X_4 : Smoking (Smoker 3) and (Non-Smoker 4)

To follow the test data Of dependent variable .

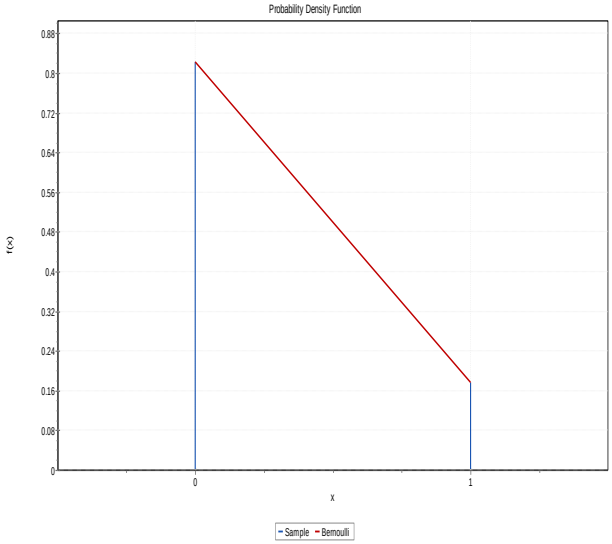
To determine the distribution of variable y variable data (the binary response variable) we will adopt the ready-made program

(EASY FIT) test (Anderson-Darling), which shows that these data are distributed according to Bernoulli's distribution with their estimation parameter (p = 0.17726) as shown in Table (1).

#	Distribution	Parameters
1	Bernoulli	p=0.17726
2	Binomial	n=1 p=0.17726
3	D. Uniform	a=0 b=1
4	Geometric	p=0.84943
5	Poisson	□=0.17726
6	Hypergeometric	No fit
7	Logarithmic	No fit (data min < 1)
8	Neg. Binomial	No fit

Table (1-4) Results of the test of good ness of fit to the variable data adopted in the application

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.2 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504



When the dependent variable is plotted, it is determined by the Bernoulli distribution function as follows

2-3. Statistical analysis

These data are analyzed through the statistical program SPSS to deter mine the importance of variables and their impact on chest diseases through the following tables:

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	299	100.0
	Missing Cases	0	.0
	Total	299	100.0
Unselected Cases		0	.0
Total		299	100.0

Table 2 shows sample size

From the above table, we note that the number of persons (sample items) who took the information for the sample of the research 299 and the missing data (observation) was equal to 0.

Observed			Predicted		
			The patient's condition		
			Uninfected	Injured	Percentage Correct
Step 0	The patient's condition	Uninfected	246	0	100.0
		Injured	53	0	.0
Overall Percentage					82.3

Table (3) shows the dependent variable

The above table shows that the number of people who did not have chest diseases from the study sample 246 and people with chest diseases was 53

.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.453	.105	231.569	1	.0781	.224

Table (4) shows the child's test

By comparing the value of a child test with the value of the kay square test at a significant level of 0.001 and the k-1 = 3 freedom level of 187.326, we find that the value of the child generated is greater than the square value of the square test. Therefore, Dependent on the dependent variable.

Iteration		-2 Log likelihood	Coefficients				
			Constant	x1	x2	x3	x4
Step 1	1	279.546	.512	.0014	.231	.123	.116
	2	276.342	.531	.0011	.713	.265	.164
	3	276.089	.891	.0013	.521	.275	.218
	4	276.127	.871	.0017	.338	.267	.281

Table (5) between the number of iterations maximum repeatability capabilities

From the table above, we note that the Maximum Likelihood are stabilized at the fourth frequency, which is used to interpret the results statistically

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	276.34	.67	.18

Table (6) Cox test

The table shows the Cox test, which is a corresponding test for the Kay box test, so that the independent variables explain 69% of the logistic regression model and the remaining 31% are included within the error limit, ie, there are other variables with high impact that are not included in the logistic regression model.

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1[a]	x1	.0120	.120	.217	1	.083	1.834
	x2	.327	.231	1.745	1	.095	.993
	x3	.126	.628	.884	1	.074	1.003
	x4	.542	.231	.777	1	.089	.781
	Constant	.628	1.008	.329	1	.054	.451

Table (7) shows the parameters of Maximum Likelihood potential

From the above table, we note that the second variable, which represents the sex of the patient, has a greater effect on the injury process through the logistic regression model. Then the third variable, which represents the environment classification of the injured (rural = 1) and (city = 2) The first age is the patient who represents the least average error

The fourth subject : CONCLUSIONS & RECOMMENDATIONS

The main conclusions and recommendations reached will be discussed in this section

4-1.Conclusions:

1.Data on the associated variant of chest disease and data distribution was followed by Bernoulli distribution.
2.by the value of a father test that was greater than the value of the scale of the Kay Square test this indicates the existence of the impact of each of the four variables and the age of the patient and the sex of the patient and the classification of the patient's environment and smoking on chest diseases.
3.through the Cox test that the four variables have interpreted 69% of the impact of chest disease.
4.Classification of the patient's environment and gender of the patient had a greater impact than the rest of the variables on the incidence of chest diseases.

4-2.Recommendations:

In Based on the conclusions reached by the researcher in this research, he recommends the following:
1.In the applied side can be used other variables not mentioned in our applied study had a significant role in the process of chest disease.
2.To develop the statistical data collection base in the Ministry of Health to obtain real, realistic and accurate data so that the results are good and satisfactory, which helps us to develop this field to reach the desired goal.
3.Other models such as the Cox model, the model of the analysis can be used, and the comparison with the logistic model.
4.The use of other methods for estimating the logistic regression function such as the bizi methods in statistical analysis.
5. Using nonlinear regression models of various types to analyze health and biological phenomena

References:

[1]Al- raule Narrator, Dr. Khasha Mahmoud (1987) "**Introduction to regression analysis**" University of Mosul - Iraq
[2] Baitin, Adel Ahmed Hassan (2009), "**Logistic regression and how to use it to construct predictive models of data with qualitative and dual-value variables**" .
[3] Brown, C.E. (1988) "**Applied Multivariate, Statistics in Geohydrology and related sciences**" Springer - Verilog. Berlin Heidelberg. Chapter 6, Multiple regression, p.p. 62-66.
[4] Hosmer, D.W. , Lemeshow, S.and Klor, J. (1988) "**Goodness of fit testing for the logistic model when the estimated probabilities are small**".
[5] Kemp, G.C.R. (2000): "**Semi-Parametric Estimation of a Logic Model**", University of Essex.
[6] Jamil, Yassin Abd-el-Kader (1988), "**Using the Greatest Way of Estimating the Parameters of a Logistics Model with Medical Application**", Master Thesis submitted to the College of Management and Economics, University of Baghdad

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.2 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

استخدام الانحدار اللوجستي لدراسة الاصابة بمرض التدرن

حيدر رائد طالب
قسم الاحصاء
كلية الادارة والاقتصاد
جامعة سومر
haider.r.t86@gmail.com

أحمد رزاق عبد
قسم الاحصاء
كلية الادارة والاقتصاد
جامعة واسط
ahmedrazzaq@uowasit.edu.iq

علي حمدالله احمد
المديرية العامة لتربية القادسية
قسم التخطيط التربوي
ali_sn2007@yahoo.com

المستخلص :

ان ما يميز هذا البحث هو التركيز على مفهوم الانحدار اللوجستي وخصائصه وكيفية بناء النموذج له اذ تم استخدام المنهج الوصفي التحليلي في توصيف الانحدار اللوجستي وكيفية تقدير معالمه واستعماله كـــــأداة مهمة في دراسة العوامل المؤثرة في إصابة بمرض التدرن حيث تضمنت عينة البحث ٢٩٩ مـــــريـــض.