University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.17 No.2 June 2022

The use of logistic regression method in data classification with practical application of Covid-19 patients in Nasiriya General Hospital

Fahad Hussein Enad/ University of Thi Qar / Department of Studies and Planning

Email: fahadh@utq.edu.iq

Zainab Nihad Mohammed Alrawi/ University of Baghdad / Computer Center

Email: zainab.alrawi@yahoo.com

https://doi.org/10.32792/utq/utj/vol17/2/4

Abstract:

This research deals with the study of binary logistic regression method for classifying data for Covid-19 patients, Covid disease is one of the largest health problem that swept the world at the present time, causing many economic problems in the world, as the research aims to know the correct classification of the influencing factors and to demonstrate the accuracy of the method in the classification, as data for Covid-19 patients were used in the Nasiriya General Hospital in Dhi Qar, several criteria were relied on, the most important of which is the correct classification ratio of the observations to indicate the performance and efficiency of the method used. The Maximum Likelihood Estimation and Weighted Least Squares methods were also used to estimate the parameters, the obtained results showed that the logistic regression method performed well in the accuracy of the correct classification of observation.

Keywords: Logistic regression, weighted least squares method, correction ratio.

1-Introduction:

Data classification is one of the important topics in various economic, social and medical sciences. With technological progress, it has led to the possibility of benefiting from the statistical classification of data, therefore logistic regression has been used which has recently attracted the attention of many researchers to apply medical research and know the factors that affect patients, so the logistic regression model was used to classify patient data, as the response variable is classified into two categories (infected and uninfected), and the main objective of the research is to display the logistic regression method for Covid-19 patients and demonstrate its accuracy with the correct classification and through the use of the correct classification ratio criterion for observations, where the importance of the research lies in the binary classification of the data. The reliance was on the factors affecting the determination of (Covid-19), as the researcher (H. Yusuff:2012) and others conducted a study on the analysis of cancer disease by using the logistic regression as a tool for x-ray imaging in order to diagnose the factors that affect Breast cancer, and the researcher (Abdul Razzaq:2016) studied the application of the logistic regression model to find out the effect of stress and psychological tensions on hypertension injury. The method of Maximum Likelihood was used to estimate the parameters, and it was concluded through the use of two tests (t and Wald) that there is an effect on the occurrence of hypertension. The researchers (Alrahamneh & Hawamdeh:2017) used the logistic regression method to classify the injured patients and the factors affecting them, the maximum likelihood method was applied to estimate the parameters and they reached through the results that the logistic regression method is effective and gives more accurate results when classifying the data through several tests.

2- **Research Problem**

The research Problem lies in the lack of applied research that use binary logistic regression to classify the affecting factors for Covid-19 patients.

3-Aim of Research

The research aims to present the logistic regression model and its accuracy according to the correct classification criterion.

4-Research Methodology

To achieve the aim of the research, a simplified definition of Covid-19 was adopted, some of its main symptoms were mentioned, the logistic regression and its characteristics were clarified and the model parameters were estimated using estimation methods and some tests for the model using the programs MATLAB (2018) and (SPSS).

5-Corona disease (Covid-19) and its symptoms

Corona viruses are a widespread family that is known to cause diseases ranging from the common cold to most severe diseases such as the Middle East Respiratory Syndrome and Severe Acute Respiratory Syndrome (SARS). Covid-19 is the disease caused by novel coronavirus called SARS-COV-2.

The World Health Organization discovered the virus for the first time on December 31st 2019 after reporting a group of viral pneumonia cases in Wuhan in the People's Republic of China. It has many symptoms such as Fever, Dry Cough and Fatigue, and there are other less common symptoms such as the loss of Taste and Smell, Nasal Congestion, Conjunctivitis, Sore Throat, Headache, Muscle or Joint Pains, various types of Rash, Nausea, Diarrhea, and Tremors, these symptoms are usually mild, and some people become infected but have only very mild symptoms or no symptoms at all.

Signs of Covid-19 Severe include:-

- Shortness of breath.
- Lack of appetite.
- Confusion.
- Persistent pain or feeling of pressure in the chest.
- High temperature (higher than 30 degrees Celsius).

Less common symptoms include:-

- Irritability.
- Confusion.
- Decreased level of consciousness (sometime associated with seizures).
- Anxiety.
- Depression.
- Sleep disorders.
- More severe and rare neurological complications such as strokes, encephalitis, delirium and nerve damage.

People of all ages with fever and/or cough associated with difficulty breathing or shortness of breath, pain or pressure in the chest, or loss of the speech or movement should seek medical attention immediately. First if possible, contact your health care provider, hotline or health facility to direct you to the appropriate clinic (World Health Organization).

2-Methods and Techniques:

1-2- Logistic Regression [7] [5]

Logistic regression is a statistical modeling of categorical data and is a special case of general linear regression models, and it is called (logistic model) or the (logit model). The model analyses the relationship between the explanatory variable and the response variable that depends on the categories. The Logistic regression is used in many fields such as medical, human and social sciences and it can also be used in the fields of engineering sciences.

2-1-1- Assumptions of logistic regression model [2]

1. The logistic regression method does not assume linear relationship between the response variable and the explanatory variables.
2. The error is not normally distributed and the variance is heterogeneous.

3. The response variable or the dependent variable must be binary for example (infected or uninfected).
4. Logistic regression does not assume equal variance within each field and does not require that the explanatory or independent variables be of continuous type and do not follow a normal distribution, and this makes logistic regression more flexible than the rest of the models.
5. The categories are supposed to be comprehensive and specific so that each item belongs to one category.
6. The simple size used in the logistic regression model must be greater than the sample size used in the linear regression model because the estimation of the logistic regression coefficients is done using the maximum likelihood function which is a method that needs a relatively large sample size.

2-1-2– Binary logistic regression model

The logistic regression model is defined as one of the nonlinear regression models in which the relationship between the response variable (Y) and the explanatory variables (x1,x2,x3,….xk) is nonlinear , logistic regression is based on a basic assumption which is that the dependent variable (Y) is a binary response that takes one of the two value (0,1), either the success of any response occurs with the probability of (p_i) or the failure of the response to occur with probability (1- p_i), so the dependent variable (Y) has a Bernoulli distribution [9pp8],[11pp3] .

Where:

$Y_i \sim Ber(p_i)$

$i = 1,2,\dots,n$

And the probability density function is in the following form:-

$P_r(Y=Y_i) = p_i^{Yi}$ (1)

$Y_{i=}0,1$

Where:

Y_i: Binary dependent variable response.

p_i: The probability of the response occurring when $Y_i = 1$.

$1 - p_i$: The probability that the response will not occur when $Y_i = 0$.

So, the expectation of the variable (Y_i) represents the probability of a response occurring (p_i):

$E(Y)$=P_r(Y=1) =p_i (2)

As for the variance of the variable (Y_i) according to the Bernoulli distribution:

$V(Y_i)$ =$p_i(1 - p_i)$ (3)

let (x_1, x_2, x_3, … …x_k) be a set of explanatory variables and (n) represents the number of observations for these variables make up the following matrix:-

$X = (x_{ij})n * k$ (4)

Where:

X: represents the matrix of independent variables.

$i = 1,2,3, \ldots\ldots n$ n: The number of observation (sample size).

$j = 1,2,3, \ldots \ldots k$ k: Represents the number of explanatory variables.

And if y_i=[$y_1, y_2, y_3 \ldots \ldots, y_n$] represents a random sample of binary variable response and $y_i \in \{0,1\}$.

Thus, the logistic regression model can be expressed by the following formula:-

$y_i = p + \varepsilon i$ (5)

And:-

μ_i: represents the logistic function (the logistic response function).

$\mu_i = p(y = 1) = \frac{e^{xi'\beta}}{1+e^{xi'\beta}}$ (6)

β: Is the parameter vector of order ($p \times 1$).

$x_i = \{x_{i1}, x_{i2} \ldots x_{ip}\}$ is a raw vector of order ($1 \times p$).

εi: Represents random error.

ε_i= y_i $-p$

The error term has a mean equal to zero and a variance equal to the variance of the dependent variable :

$E(\varepsilon_i) = E(y_i) - E(p) = p_i - p_i = 0$ (7)

$V(\varepsilon_i) = V(y_i) = p_i(1 - p_i)$ (8)

This model is converted to a linear form that is represented by a linear relationship through the raw vector $(\acute{X_\iota})$ of explanatory variables with the probability log $[logit\ p(Xi)]$ according to the following mathematical formula:-

$$Z = logit\ p(\ ) = ln\left[\frac{p(x_i)}{1-p(x_i)}\right] = \beta_O + \beta_1 X_{i1} + 000 + \beta_p X_{ip} \qquad (9)$$

$$= [1\ Xi1\ \dots\ Xip\ ]\begin{bmatrix}\beta_O\\ \beta_1\\ 0\\ 0\\ 0\\ \beta_p\end{bmatrix} = \acute{X_\iota}\beta$$

$\therefore Zi = \acute{X_\iota}\beta + \varepsilon i$

To estimation parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ we use one of the estimation methods (Maximum Likelihood Estimation) [6].

When we have a sample from a set of independent observations with a size (n) for each pair (X_i, Y_i)

$i = 1,2, \dots, n$

Where:

Y_i: Represents the rank of the binary response variable for the observation i.

X_i: Represents the value of the independent variable for the observation i.

Then:

$(Yi = 1\backslash X) = f(X)^{Yi}$ P when $Yi = 1$.

$(Yi = 0\backslash X) = [1 - f(X)]^{1-Yi}$ P when $Yi = 0$.

The maximum likelihood function can also be expressed in the following form:

$L(B) = \prod_{i=1}^{n} f(X)^{Yi}\, [1 - f(X)]^{1-Yi}$

By taking the logarithm of both sides, we get the following equation:

$$L(B) = \ln L(B) = \sum_{i=1}^{n} \{Yi\ln[f(X)] + (1 - Yi)\ \ln[1 - f(X)]\}$$

Where the above equation is derived for the parameters to be estimated (B_i) and equating them to zero, thus we get a number of equations that can only be solved through an iterative algorithm called iterative weighted least squares algorithm [10].

2-1-3 -Weighted Least Squares Method (WLS)

To estimate the parameters of equation (9), weighted least squares are used according to the following equation [10:pp.8]:

$$\hat{\beta} = (X'WX)^{-1}X'WZ \qquad (10)$$

Where:

Z: represents the value of a linear transformation of order n*1.

W: A square matrix with diagonal elements representing the variances of order n*n.

$X'WX$: represents the square matrix resulting from multiplying the variances matrix with the matrix of explanatory variables with rank $[(\mathrm{p}+1)(\mathrm{p}+1)]$.

$X'WZ$: represents the value of the transpose of the explanatory matrix and the variances matrix with the value of the linear transformation of rank $[(p+1) * n]$.

Where:-

$$\underline{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$$

$$W = \begin{bmatrix} W_1 & 0 & \cdots & 0 \\ 0 & W_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_n \end{bmatrix}$$

$$\boldsymbol{X}'WX = \begin{bmatrix} \sum_{i=0}^{n} W_i & \sum_{i=0}^{n} W_i X_{i1} & \cdots & \sum_{i=0}^{n} W_i X_{ip} \\ \sum_{i=0}^{n} W_i X_{i1} & \sum_{i=0}^{n} W_i X_{i1}^2 & & \sum_{i=0}^{n} W_i X_{i1} X_{np} \\ \vdots & \ddots & & \vdots \\ \sum_{i=0}^{n} W_i X_{ip} & \cdots & & \sum_{i=0}^{n} W_i X_{ip}^2 \end{bmatrix}$$

$$X'WZ = \begin{bmatrix} W_i Z_i \\ W_i X_{i1} Z_i \\ \cdot \\ \cdot \\ W_i X_{ip} Z_n \end{bmatrix}$$

2-3- Classification Table Criterion (CTC)

The classification table shows the number of cases that were classified correctly and the number of cases that were classified incorrectly, the table consists of four cells containing the cases that were classified correctly or incorrectly, the table can be illustrated as follows:-

Table (1) represents the classification

SUM	**Expected**			**Classification**
	N Neg.	**P Pos.**		
TP+FN	**False FN Negative**	**True (TP) Positive**	**P**	**Observation**
TN+FN	**True TN Negative**	**False (FP) Positive**	**N**	
+TP+FP FN+TN	**FN+TN**	**TP+FP**		**SUM**

And these criteria are calculated based on the following indicators:-

1- (True Positive) (TP) Diagnosing the sample correctly uninfected.
2- (True Negative) (TN) Diagnosing the sample correctly infected.
3- (False Positive) (FP) Diagnosing the sample incorrectly uninfected.

4- (False Negative) (FN) Diagnosing the sample incorrectly infected.

Depending on these indictors, a set of criteria is calculated which is:-

1- Accuracy:- Represents the ratio of the correctly diagnosed results (Positive, Negative) to the overall results.

$$\mathbf{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} * \mathbf{100\%} \qquad \mathbf{(11)}$$

2- Specificity:- Represents the ratio between samples that were diagnosed correctly infected (True Negative) to the total samples that were diagnosed correctly infected and the samples that were diagnosed incorrectly uninfected.

$$\mathbf{SPE} = \frac{\mathbf{TN}}{\mathbf{TN+FP}} * \mathbf{100\%} \qquad \mathbf{(12)}$$

3- Sensitivity:- Represents the ratio between samples that were diagnosed correctly uninfected to all samples that were diagnosed uninfected (correctly and incorrectly).

$$\mathbf{SEN} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} * \mathbf{100\%} \qquad \mathbf{(13)}$$

4- False Positive rate:- Represents the ratio between samples that were diagnosed uninfected incorrectly to all samples that were diagnosed uninfected correctly (TN) or incorrectly (FP).

$$\mathbf{FPR} = \frac{\mathbf{FP}}{\mathbf{TN+FP}} * \mathbf{100\%} \qquad \mathbf{(14)}$$

2-4 correct classification ratio law [3:pp38]

It can be calculated according to the following formula:-

$$\boldsymbol{PC} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FN+FP}} * \mathbf{100} \qquad \mathbf{(15)}$$

2-5- The criterion for comparing the methods

The mean squared error (MSE) criterion was relied upon, and it is calculated according to the following formula:-

$$\mathbf{MSE} = \frac{\mathbf{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{i=1}}^{\mathbf{N}} (\boldsymbol{Y_i} - \widehat{\boldsymbol{Y}})^{\mathbf{2}} \qquad \mathbf{(16)}$$

2-6- Evaluation of the explanatory power of the logistic regression

The use of the statistic $R^2{}_{cox\&snel}$ or the statistic $R^2{}_{Nagelkerke}$ in logistic regression is to know the quality of the estimated regression equation in explaining the relationship between the dependent variable and explanatory variables, as the two statistic have the same objective as the coefficient of determination R^2 that is used in multiple linear regression , but the value of $R^2{}_{cox\&snel}$ can not reach one unlike $R^2{}_{Nagelkerke}$ that can, because the limits of $R^2{}_{Nagelkerke}$ goes from zero to one, and that makes it more reliable than $R^2{}_{cox\&snel}$ and its value is naturally higher than $R^2{}_{cox\&snel}$. The statistic $R^2{}_{cox\&snel}$ can be calculated through the following formula :- [11:2014:pp51]

$$R^2{}_{cox\&snel} = 1 - [\frac{L_0}{L_1}]^{(\frac{2}{n})} \qquad (17)$$

Where:

L_0:- The maximum likelihood function in the case of the model that includes the fixed term.

L1:- The maximum likelihood function in the case of the model that includes all the variables.

While the statistic $R^2{}_{Nagelkerke}$ is calculated through the following relationship:-

$$\mathbf{R^2{}_{Nagelkerke} = \frac{R^2{}_{cox\&snel}}{1-[L_0]^{(\frac{2}{n})}}} \qquad \mathbf{(18)}$$

2-6- Tests for the logistic regression model

2-6-1- Wald’s test

The Wald statistic that follows the chi-square χ^2distribution and has a degree of freedom of df=1 for each parameter of the logistic regression model is tested in order to test the null hypothesis, which states that the effect of the logit coefficient is equal to zero and is calculated according to the following formula:-

$$\text{Wald} = [\frac{B_j}{S.E(B_j)}]^2 \qquad (19)$$

Where:

S.E: Represents the standard Error of the estimated parameter of rank j.

The test hypothesis is as follows: $H_0: B_j = 0$ vs $H_1: B_j \neq 0$ j=1, 2… k

Test the significance of each parameter of the model and the test is two – sided, if the probability value of Wald's statistic is less than (0.05), we reject the null hypothesis.

2-6-2 Hosmer & Lemshow test (H&L)

This test that follows the chi –square distribution(χ^2) is used to find out whether the model represents the data well or not by evaluating the difference between the observed and expected values [3]. And the hypothesis is as follows:-

H0: There are no significant differences between the observed values and expected values.

H1: There are significant differences between the observed values and the expected values.

If the (H&L) statistic is greater than (0.05), than the model is well represented and the null hypothesis is accepted, which states that there are no differences between the observed and expected values [12].

3-Application side

The application side includes classifying the data using the binary logistic regression model method, the correct classification ratio criterion was used to show the efficiency and accuracy of the methods used. This research was based on data related to the Covid-19 epidemic, the data were collected from Nasiriya General Hospital in Dhi Qar because it is one of the important aspects in the health aspect, as the research included a sample size of (396) patients and the patients were divided into two types, infected was (215) and the uninfected (181), six variables were used (Gender, Academic achievement, Occupation, Age, Weight and Smoking) and the dependent variable was (infected and uninfected) and the variables adopted for the purpose of the study were:

Table (3) represents explanatory variable

Represents the value	The value that the variable takes	Represents	Variable

Represents the male	1	Gender	X1
Represents the female	2		
primary	1	Academic achievement	X2
medium	2		
Middle school	3		
diploma	4		
Medical	1	Occupation	X3
Health	2		
Engineering	3		
Educational	4		
administrative	5		
freelance	6		
retired	7		
student	8		
child	9		
disabled	10		
farmer	11		
worker	12		
housewife	13		
other	14		
Represents Age			X4
Represents Weight			X5
Smoker	1	Smoking	X6
Non smoker	2		
Infected	0	Response variable	Y
Uninfected	1		

These methods were applied writing the program in MATLAB language (MATLAB: 2018b) and using the software package (SPSS), the experiment was conducted on Covid-19 samples and it was of two types: the first type represents the infected samples and the second type represents the uninfected samples, Tables (1) and (2) show the information about the sample and the encoding of the dependent variable.

Table (1) Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	396	100.0
	Missing Cases	0	.0
	Total	396	100.0
Unselected Cases		0	.0
Total		396	100.0

a. If weight is in effect, see classification table for the total number of cases.

Table (1) shows the number of observations included in the analysis, which is the sample size (396), the number of missing observation equal to zero, and the total size of the data is (396). As for Table (2) it represents the encoding of the dependent variable.

Table (2) Dependent Variable Encoding

Original Value	Internal Value
P	0
N	1

Table (2) show the values of the dependent variable (Y) where the affected person was expressed as (0) and the uninfected person (1), and the uninfected person (1), and Table (3) shows the number of iterations.

Table (3) Iteration History[a,b,c,d]

Iteration	-2 Log likelihood	Coefficients						
		Constant	x1	x2	x3	x4	x5	x6
1	381.120	-4.508	.121	.078	-.035	-.010	.061	.298
2	364.689	-6.556	.214	.107	-.055	-.013	.088	.425

	3	363.570	-7.274	.262	.118	-.063	-.013	.098	.465
	4	363.562	-7.342	.268	.119	-.064	-.013	.099	.468
	5	363.562	-7.343	.268	.119	-.064	-.013	.099	.468

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 546.050

d. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Table (3) shows the number of iterative cycles in order to obtain the highest value of the term (-2Log Likelihood), which we obtained in the fifth cycle, which is equal (363.562) after its initial value in the case of the model that includes only a fixed term was equal to (546.050) which is referred to in paragraph (c), and paragraph (d) shows the state of stopping at the fifth cycle because the change in the estimated values of model parameters is less than (0.001), therefore, at the fifth cycle the best estimated values for the model parameters were obtained.

As for Table (4), it shows the efficiency of the model

Table (4) Omnibus Tests of Model Coefficient

		Chi-square	df	Sig.
Step 1	Step	182.488	6	.000
	Block	182.488	6	.000
	Model	182.488	6	.000

Table (4) shows the efficiency and accuracy of the model using the maximum likelihood ratio that follows the ($Ꭓ^2$) distribution and with a degree of freedom equal to the number of independent variables (6), which can be calculated through the following formula**:**

$$Ꭓ^2 = -2LogL_0 - (-2LogL_1)$$

Where:

$\chi^2 = 546.050 - 363.562 = 182.488$

The significance level that was used is (α=0.001) because it was (sig= 0.00) this confirms the efficiency and significance of the model. And Table (5) shows the statistics ($R^2_{cox\&snel}$) and ($R^2_{Nagelkerke}$).

Table (5) Model Summary

Step	-2 Log likelihood	$R^2_{cox\&snel}$	$R^2_{Nagelkerke}$
1	363.562[a]	.369	.494

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Table (5) shows the ($R^2_{cox\&snel}$) and $(R^2_{Nagelkerke})$ statistics, if the value of ($R^2_{cox\&snel}$) equal (0.369) this shows that (36.9%) of variances in the dependent variable are explained by the model, as for ($R^2_{Nagelkerke}$) its value is equal to (0.494) meaning that (49.4%) of the variances in the dependent variable are explained by the logistic regression model and table (6) shows the (H&L) test.

Table (6) Hosmer and Lemeshow

Step	Chi-square	Df	Sig.
1	17.020	8	.030

Table (6) shows the (H&L) test, as it shows the (H&L) statistic that follows the (χ^2) distribution equals (17.020), which is significant because (sig=0.030) which is less than (0.05), therefore, we reject the null hypothesis which states that there are no significant differences between the observed values and the expected values, and this is clarified through the result of Table (7), which shows the observed and expected values.

Table (7) Contingency Table for Hosmer and Lemeshow Test

		y = P		y = N		Total
		Observed	Expected	Observed	Expected	
Step 1	1	35	38.769	5	1.231	40

2	36	36.196	4	3.804	40
3	35	32.862	5	7.138	40
4	34	29.645	6	10.355	40
5	26	25.398	14	14.602	40
6	23	21.339	18	19.661	41
7	13	14.851	27	25.149	40
8	7	9.921	33	30.079	40
9	5	4.824	35	35.176	40
10	1	1.195	34	33.805	35

As for table (8), it shows the classification percentage.

Table (8) Classification Table[a]

Observed			Predicted		
			Y		Percentage Correct
			P	N	
Step 1	y	P	184	31	85.6
		N	43	138	76.2
	Overall Percentage				81.3

a. The cut value is .500

Table (8) shows the percentage of correct classification of the observations where (85.6) of the people who were correctly classified among the infected people and (76.2) of the people were correctly classified among the uninfected people, and the percentage of people who were correctly classified is (81.3) of the people who were classified incorrectly, which is (4.7) and it is a small percentage and this indicates the efficiency of the model in representing the data where the sensitivity of the model is equal to (85.5%) and the specificity is equal (76.2).

Table (9) shows the estimated values of the model parameters

Table (9) Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	x1	.268	.497	.289	1	.591	1.307
	x2	.119	.168	.500	1	.479	1.126
	x3	-.064	.067	.930	1	.335	.938
	x4	-.013	.009	2.333	1	.127	.987
	x5	.099	.010	94.121	1	.000	1.104
	x6	.468	.333	1.979	1	.160	1.597
	Constant	-7.343	1.060	48.008	1	.000	.001

a. Variable(s) entered on step 1: x1, x2, x3, x4, x5, x6.

Table (9) show the estimated values of the parameters of the logistic regression model using the maximum likelihood method, and the model is as follows:

$$\hat{Y} = -7.343 + 0.268X1 + 0.119X2 - 0.064X3 - 0.013X4 + 0.099X5 + 0.468X6$$

The table above shows the standard error values for each parameter of the model and shows the Wald test values for each parameter respectively and also shows EXP (B) that represents the likelihood ratio, where EXP (B) indicates the amount of changes in the likelihood ratio in the case of the infected person when the occurrence of changes in the values of the independent variables associated with parameter (B), and in the case of EXP (B) greater than (1) the probability ratio increases in the case of an infected person, but if it was less than (1), increase in the independent variable leads to a decrease in that percentage.

Table (10) shows the use of the weighted least squares method to estimate the parameter of the model and its standard error.

Table (10) shows the estimation of the parameter of the method

Parameters	Estimated parameter	Standard error

$\widehat{B_0}$	-0.627	8.217
$\widehat{B_1}$	0.030	0.043
$\widehat{B_2}$	0.019	0.026
$\widehat{B_3}$	-0.008	0.006
$\widehat{B_4}$	-0.002	0.001
$\widehat{B_5}$	0.015	0.001
$\widehat{B_6}$	0.074	0.042

So the logistic regression model is as follows:-

$$\hat{Y} = -0.627 + 0.030X1 + 0.019X2 - 0.008X3 - 0.002X4 + 0.015X5 + 0.074X6$$

This shows that the weighted least squares method gives more efficiency in estimating the parameters of the logistic regression model by calculating the mean squares error whose value was in the above method (4.995), but in the maximum likelihood method, its value was (7.542).

5-Conclusions and Recommendations

1-5- Conclusions

Through the results, it was found that the logistic regression model whose parameters were estimated was an efficient and good representative of the data as it passed the tests related to logistic regression, like (H&L) test and the classification ratio of the data was good through the correct classification ratio and sensitivity, the least squares method is more efficient and accurate than the maximum likelihood method.

2-5- Recommendations

We recommend conducting advanced studies to estimate and classify new infections with the corona virus at the local and global levels, knowledge of housing with this disease, and developing appropriate strategic plans to confront this epidemic, and we recommend using the logistic regression model in different medical fields especially when data for contagious diseases are available in large sample sizes, also the use of advanced methods especially artificial intelligence methods (such as KNN) and other methods.

Refrences

1-Abd-allrazak, M.& allimam Zalan, R.,(2016)," Using an Approach of Logistic Regression to Analyze Effect of Psychological Stress on Blood Pressure an Applied Study on a Sample from the Patients in Zubair",Basrah, Gulf Economic Journal,vol.27,pp.48-66

2-Alrahamneh, A. & Hawamdeh, O. ,(2017) , "The Factors Affecting Eye Patients (Cataract) In Jordan by Using the Logistic Regression Model", *MAS*, pp. 38-42.

3-Abdulqader, Q. M., (2015),"Comparison Of Discriminant Analysis and Logistic Regression Analysis: An Application or Caesarean Births and Natural Births Data" ,Zakho Tecnical Institute Duhok, Polytechnic University Duhok, Duhok , Iraq, INAS, pp.34-46.

4-Burns , Robert & Burns , Richard , (2008) ," Business Research Methods and Statistics using SPSS", Five extra advanced chapters , chapter 24 Logistic Regression : pp 568 -575.

5-Cramer, J. S. (2002). The origins of logistic regression.pp.3-7.

6-Gayou, O., & et al., (2008)," A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes", American Association of Physicists in Medicine, pp. 5426 – 5433.

7-Gelman, A. & et al., (2014)," Bayesian Data Analysis", Texts in Statistical Science , *CRC* , *LC* , Chapman and Hall Book.

8-H. Yusuff , N. Mohamad , U.K. Ngah & A.S. Yahaya , (2012) , "Breast Cancer Analysis Using Logistic Regression" , IJRRAS ,Vol 10 , No.1 : pp 14 – 22.

9- Hosmer, D., Lemeshow, S. & Sturdivant , R. ,(2013)," Applied Logistic Regression", 3rd edition ,New York: willey,WSIPS, http : // ihmsi.org.

10-Müller , Marlene , (2004) , " Generalized Linear Models " , Fraunhofer Institute for Industrial Mathematics (ITWM) , (Germany) . www.marlenemueller.ed/publications/HandbookCS.pdf

11- Rodriguez ,G.,(2007), "Logit Models for Binary Data" ,Chapter(3) ,Retrieved from,http://data.princeton.edu/wws509/notes/c3.pdf

12-Sukono, Sholahuddin, A. & et al., (2014)," Credit Scoring for Coopera of Financial Services Using Logistic Regression Estimated by Genetic Algorithm", AMS, pp. 45-57.

12-Wuensch, K. L. (2014). Binary logistic regression with SPSS. Retrieved March, 18, 2015.pp 2 -4.

13- World Health Organization.