# Predicting the Probability of Infected with Lung Failure in Covid-19 Patients Using a Logistic Regression Model

**Sabah Hasan Jasim**
Mathematic Department, College of Basic Education, Misan University, Iraq.
E-mail: sabah.h.alsaedi@gmail.com

*Received August 22, 2021; Accepted December 02, 2021*
*ISSN: 1735-188X*
*DOI: 10.14704/WEB/V19I1/WEB19074*

## Abstract

Logistic regression has increased in social and medical sciences and in education research. A program SPSS version (22) was used to estimate the logistic regression model after entering the required information. Discussion of the results obtained through the use of the outputs the results to analysis and discussion. The model was applied, about 125 samples for Covid-19 patient was entered from people has Covid-19 disease, and all independent variables were entered to infer the adoption of lung infection and its relationship to the variables together. The aims of study were probability of infected with covid-19 disease regarding to use some variables and forecasting probability infected with lung failure in Covid-19 patients which lead to cytokine syndrome and death. Use the logistic regression in medical fields as a statistic analysis tools.

## Keywords

Logistic regression, Forecasting, Probability, Covid-19.

## Introduction

In 2019, the world health organization (WHO) declared spread a new corona virus or Covid-19 diseases as a cold or flu in china which causes respiratory acute illness and dry coughing, sneezing and high fever and lead to death as epidemic disease in the world. (1). Logistic regression was discovered in the end 1960 and early 1970 and use as a statistical tool routinely available in statistical branches in early 1980(2). Logistic regression has increased in social and medical sciences and in education research (3). It considered as one of the important preferred regression models that can be implemented in modelling binary dependent variables. It is a statically mathematical modelling approach used to define the relationship between such independent variables as $X_1$, $X_2$,…, $X_n$ and Y binary dependent variable which is coded as 0 or 1 for two possible categories. The independent variables may be continuous, discrete, binary or a combination of them. Maximum likelihood methods may be used to estimate the parameters of the logistic model. The interpretations of coefficients are made with odds rate values. In other words, the logistic regression analysis has been reviewed that can define the relationship between the binary result variable and independent variables comprising of both continuous and discrete variables. The aims of study the probability of infected with Covid-19 disease according to variables and forecasting probability infected with lung failure in covid-19 patients which lead to cytokine syndrome and death. Use the logistic regression in medical fields as a statistic analysis tools (4).

## Introduction to Mathematics of Logistic Regression

It is the more important title of statistical is regression analysis. Regrading to Gujarati use regression at predicting to depend variable or other expressive variables. One of types of regression is binary logistic regression analysis. The model is:

$$\widehat{Y_\iota} = b_0 + b_1 x_i + \varepsilon_i$$

y mean represents the dependent variable X mean independent variable.

Some program as SPSS, EVIEWS, SAS and EXCEL used regression model in solve a problem such as mathematics, physics, biological and economy and economy and used for deletion the correlation between two or more variables which have relation (5):

**Model simple regression model:**

$\widehat{Y_\iota} = b_0 + b_1 x_i + \varepsilon_i$

**Multiple Regression Model:**

$\widehat{Y_\iota} = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \cdots\cdots\cdots + b_n x_{in} + \varepsilon_i$
$\widehat{Y_\iota}: \ dependent\ varible$
$x_{i1}, x_{i2}, x_{i3}, \cdots\cdots\cdots, x_{in} : Independent\ variabels$
$b_0, b_1, b_2, b_3 \cdots\cdots\cdots, b_n : Coefficiente$
$\varepsilon_i$=Error term (Average = 0 & Variance = $\sigma^2$)

Logistic regression: It used in estimates the effects of independent variable on risk factors as probability, also it is used for detection risk factors as probability and methods investigation a correlation of dependent variables with independent variables in binary logistic regression. Dependent variables as binary like (0,1) (4).

## Logistic Regression Analysis System

Steps of logistic regression are briefly as

* $\widehat{Y}_1, \widehat{Y}_2, \widehat{Y}_3, \ldots\ldots\ldots \widehat{Y_n}$ , $\widehat{Y}_1$ 1is statistically explanatory.
*Explanatory variables $(X_n)$  are explanatory from each other.

$$\widehat{Y}_\iota \in (0,1) \qquad\qquad i = 1,2,3, \ldots \ldots \ldots \ldots n$$

$$P_{ro.}( \widehat{Y}_\iota = \frac{1}{x_i}) = P_i \qquad\qquad i = 1,2,3, \ldots \ldots \ldots \ldots n$$

But linear probability function which is one of assumption on logistic regression, also present error term with distribution.

$$P_{ro.} = \frac{exp^{(b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + \cdots\cdots\cdots + b_nx_{in})}}{1 + exp^{(b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + \cdots\cdots\cdots + b_nx_{in})}}$$

$$P_{ro.} = \frac{1}{1 + exp^{-(b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + \cdots\cdots\cdots + b_nx_{in})}}$$

$P_{ro.}$ = probability event
$b_0$= statement of dependent variable when independent variables are zero.
$b_0, b_1, b_2, b_3$=mean coefficients of independent variables $x_{i1}, x_{i2}, x_{i3}$=independent variables (p mean independent variables).
$B_o$= mean coefficient of independent variables.
$p$ =mean probability of event analysis in equation of logistic regression.

Odds rate is odd values of two different analyzed events to each other as Exp(B) in logistic analyzed (Gujarati.,1999).

Exp(B) mean number of folds in what percentage y variables has observing probability with effect of yp variables.

- Features of logistic regression dependent variables value 0 or 1 therefore , probability y =1(X)=( $\pi$ X) .

Y is interpreted as odds rate

## Binary groups logistic regression analysis.

If dependent variables has two level like 0,1 with probability of p(y=1) to get value of I (of odd event) expected value will be

$$E(\widehat{Y_\iota}) = one\ x\ pro.(\widehat{Y_\iota} = 1) + zero\ x\ pro.(\widehat{Y_\iota} = 0) = \ (\widehat{Y_\iota} = 1)\,.$$

Shown as a regression equation

$$E(\widehat{Y_\iota}) = \ pro.(\widehat{Y_\iota} = 1) = \ b_0 + \sum b_n x_{in}$$

Binary group logistic regression

1. 1-$\widehat{Y}_i \in (0,1)$ $\qquad i = 1,2, \ldots \ldots.., n$
2. $pro.\left( \widehat{Y_\iota} = \frac{1}{x_i} \right) = p_i$

$$P_i = \frac{exp^{b_0 + \sum\ b_n x_{in}}}{1 +\ exp^{b_0 + \sum\ b_n x_{in}}} \qquad (2,3,2)$$

3. $\widehat{Y_1}, \widehat{Y_2}, \widehat{Y_3}, \ldots\ldots\ldots \widehat{Y_n}$ , $\widehat{Y_1}$1is statistically explanatory.
4. Ex, Expressive are independent from each other according to discrete or continuous or both

   a. if expressive variable is discrete of logistic regression.

   $$\ln(\frac{P_i}{1 -\ P_i}) = \ b_0 + \sum\ b_n x_{in}$$

   b. (b)if all expressive variables are continuous as

   $$\ln = \frac{P_{(x_1\ ,x_2\ ,\ldots\ldots\ldots,x_n)}}{1 -\ P_{(x_1\ ,x_2\ ,\ldots\ldots\ldots,x_n)}} = \ b_0 + \sum\ b_n x_{in}$$

   c. If some expressive variables are discerning and some are continuous, multiple variables distribution is $f_1(x_1, x_2, \ldots \ldots.., x_n\ )$ for success and $f_0(x_1, x_2, \ldots \ldots.., x_n\ )$ for failure and equation of logistic regression model be

   $$\ln = \frac{P\ f_1(x_1,x_2,\ldots\ldots,x_n\ )}{(1-P)\ f_0(x_1,x_2,\ldots\ldots,x_n\ )} = b_0 + \sum\ b_n x_{in}$$

P: pre- probability to get rate 1

## Estimate methods by Maximum Likelihood

Is statistical methods which make maximum the probability of giving a data set observed, (6). The aim of use the maximum likelihood is to detect estimates of p expressive variable as to make maximum the probability of y variable observing. $b_0$ and $b_1$ is a model values in logistic regression for detected value of $\widehat{Y_\iota}$.

probability is not depend $P_i = pro.(\widehat{Y_\iota} = \frac{1}{x_i})$

probability is not depend $1 - P_i .( i = 1,2,3, \dots \dots . , n )$

It can be made as $pro.\left(\frac{\widehat{Y_\iota}}{x_i}\right) = (P_i)^{y_i}\,(1 - P_i)^{1-y_i}$

$$L\left(\frac{\widehat{Y_\iota}}{x_i}\right) = (\prod P_i)^{y_i}\,(1 - P_i)^{1-y_i}$$

$P_i$ : probability of the event happening

$1 - P_i$ : $P_i$ : probability of the event not happening

$$L\left(\frac{\widehat{Y_\iota}}{x_i}\right) = \Pi\,(\frac{exp^{(b_0+b_1x_{i1}+\cdots\cdots+b_nx_{in})}}{1+exp^{(b_0+b_1x_{i1}+\cdots\cdots+b_nx_{in})}})^{y_i}\,(\frac{1}{1+exp^{(b_0+b_1x_{i1}+\cdots\cdots+b_nx_{in})}})^{1-y_i}$$

$L\left(\frac{\widehat{Y_\iota}}{x_i}, b\right)$ is probability function, this method select estimate value of $b_0$.

To find the maximum like hood as $L\left(\frac{\widehat{Y_\iota}}{x_i}\right)$ lead to

$\ln\ L\left(\frac{\widehat{Y_\iota}}{x_i}, b\right) = \sum(\widehat{Y_\iota} \ln P_i + \left(1 - \widehat{Y_\iota}\right) \ln(1 - \widehat{Y_\iota}))$ to find probability equation.

$\sum(\widehat{Y_\iota} - P_i)\, x_{ij} = 0 \qquad j = 1,2,3 \dots \dots \dots . , m$

## Results of Assessment and Testing of the Logistic Model for Samples and Discussion

A program SPSS version (22) was used to estimate the logistic regression model after entering the required information. A program was used to estimate the logistic regression model after entering the required information. Presentation and discussion of the results obtained through the use of the outputs in terms of form and in order to facilitate their analysis and discussion. The logistic regression model was applied, where 125 samples for Covid -19 patient were entered from people with Covid-19 disease, where all independent variables were entered to infer the adoption of lung infection and its relationship to the variables together. According to output results from SPSS program in the following tables:

The stage of reading a logistic regression results is called the model testing stage, where the predictive or explanatory regression strength is known by comparing the regression model that includes constant (B) only without the predictor variables with a model that contains predictive variables and comparing the model before entering the values to be predicted, this called Bock (0). In block (0) classification table **(1).** It is evident the results of the predictive values of lung infection in Covid-19 patients, meaning that the probability of not related with a lung infection with a percentage (54%) while the probability of infection is (71%) either the predictive accuracy of the model was (56.8%).

**Table 1 Classification table for predictive value**

| Case | Predicted | | Percentage Correct |
|---|---|---|---|
| | Infected | Non infected | |
| Infected | 0 | 54 | 0 |
| Non infected | 0 | 71 | 100 |
| Overall percentage | | | 56.8% |

Regrading to table (2) based on the -2log likelihood for estimating the model and allows the possibility of using chi-square without predictive variables.

**Table 2 Model Summery**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 137.579 | .234 | .315 |

In Table (3), the value of the constant (0.274), and Wald test (2.298) are shown, with no significant value (0.130), and this means that the model is unable to predict independent variables before entering the predictive variables.

**Table 3 Variable in the Equation**

| | B | S.E | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| Step(0) Constant | 0.274 | 0.181 | 2.298 | 1 | 0.130 | 1.315 |

Table (4) show all independent value (9) variables with significant value for all individual.

**Table 4 Variables not in the Equation**

| Variable | Wald | df | Sig |
|---|---|---|---|
| Age (years) | 10.210 | 12 | 0.598 |
| LDH | 1.030 | 1 | .310 |
| CRP | 1.367 | 1 | .242 |
| Ferritin | .743 | 1 | .389 |
| Neutrophile | 5.161 | 1 | .023 |
| Lymph | 6.035 | 1 | .014 |
| WBcs | 3.503 | 1 | .061 |
| HGB | .248 | 1 | .619 |
| Plts | 17.373 | 1 | .000 |
| Total | 40.559 | 20 | .004 |

Regarding to block (1), In table **(5)** enter all independent value and show effect of predictive value on lung infection in covid-19 patients. Chi-square It contains all independent predictive variables and are statistically significant, this means that the model will improve its conformity and its ability to predict variable disease in Covid-19 patients.

**Table 5 Omnibus Tests of Model Coefficients**

| | Chi-square | df | Sig. |
|---|---|---|---|
| Steps(1) | 48.828 | 20 | 0.00 |
| Block | 48.828 | 20 | 0.00 |
| Model | 48.828 | 20 | 0.00 |

Table (6) In -2log likelihood table, it was noted that the value (122.140) with predictive variables, where its value was smaller compared to block 0, where -2log likelihood value was (137.579) greater in the case of not entering predictive variables, and this indicates the existence of a relationship between the predicted variables and independent variables and indicates that the predicted variables contributed to Improve the model's fit and its ability to predict the independent variable and then check how well the model matches the data.

**Table 6 Model summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 122.140 | 0.323 | 0.434 |

In the Hosmer and Lemeshow test table (7), which is called the conformational quality test, which contains chi-square to judge the quality of the model's matching to the data, where the chi-square (12.172) with df (8), as a not statistically significant (0.144), which indicates that the model is identical to the data and the model's ability to distinguish between a group of patients and healthy people, as well as the predictive ability of each group and its relationship with predicted factors.

**Table 7 Hosmer and Lemeshow Test**

| Observed | | Predicted | | |
|---|---|---|---|---|
| | | Non infected | Infected | Percentage correct |
| Step1 | Non infected | 40 | 14 | 74.1 |
| | Infected | 10 | 61 | 85.9 |
| Overall percentage | | | | 80.8% |

In Classification table (8) shows the ability of the model to correctly classify the group of patients and healthy people through its ability to predict each group, as it classified them to non-infected (40) individuals, with a rate of (74.1%) and infected individuals (10), was (85.9%), and its overall predictive capacity is (80.8%).

**Table 8 Classification Table**

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 12.172 | 8 | .144 |

According to Wald test was found that lymphocytes recorded (5.882) at a statistical significance (0.015) and this indicates an importance of the lymphocyte variable in the prediction of lung infection in Covid-19 patients with Exp (B) was (0.938). This means a decrease of one degree in the lymphocyte variant increases the possibility that the person has lung infection. As for the Wald test for the platelet variable (12.900) where it was observed that an increase of one degree by a percentage of (1.010) with a significant value (0.00) and increase of one degree in the variable blood platelet increases the likelihood that the individual has lung infection and may recover. The final equation of logistic regression as the following:

**Table 8 Variable in the Equation**

| Factor | B | S.E | Wald | df | Sig | Exp(B) |
|---|---|---|---|---|---|---|
| Age | | | 9.218 | 12 | 0.684 | |
| LDH | 0.000 | 0.002 | 0.15 | 1 | 0.901 | 1.000 |
| CRP | - 0.12 | 0.14 | 0.769 | 1 | 0.380 | 0.988 |
| Ferritin | - 0.002 | 0.001 | 2.089 | 1 | 0.148 | 0.998 |
| Neutrophil | 0.008 | 0.009 | 0.767 | 1 | 0.381 | 1.008 |
| Lymph | - 0.064 | 0.027 | 5.882 | 1 | 0.015 | 0.938 |
| WBcs | 0.036 | 0.027 | 1.801 | 1 | 0.180 | 1.037 |
| HGB | 0.147 | 0.144 | 1.043 | 1 | 0.307 | 1.158 |
| Plts | 0.010 | 0.003 | 12.900 | 1 | 0.000 | 1.010 |
| Constant | - 3.339 | 2.235 | 2.232 | 1 | .135 | .035 |

$$\boldsymbol{ln} = (\boldsymbol{p_i}/\mathbf{1} - \boldsymbol{p_i}) = (\mathbf{-3.339}) - (\mathbf{0.064})\,\boldsymbol{x_1} + (\mathbf{0.010})\,\boldsymbol{x_2}$$

It is done by converting Exp(B) to the odds ratio to facilitate reading the results by converting the odds ratio to a percentage that is comparable between two groups of healthy and sick from the following equation:

Percentage (%) = $(\boldsymbol{Exp}(\boldsymbol{B}) - \mathbf{1}) \times \mathbf{100}$<br>
Plts % = $(\mathbf{1.010} - \mathbf{1}) \times \mathbf{100} = \mathbf{10}$<br>
Lymphocyte% = $(\mathbf{0.938} - \mathbf{1}) \times \mathbf{100} = \mathbf{-6.2}$

**Table 9 Percentage of variable predicted**

| Number | Case | Predicted ratio | Variables |
|---|---|---|---|
| 1 | Significant | 10% | Plts |
| 2 | Significant | 6.2% | Lymphocytes |

## References

Unhale, S.S., Ansar, Q.B., Sanap, S., Thakhre, S., Wadatkar, S., Bairagi, R., & Biyani, K.R. (2020). A review on corona virus (COVID-19). *World Journal of Pharmaceutical and life sciences*, *6*(4), 109-115.

Chuang, H.L. (1997). High school youths' dropout and re-enrollment behavior. *Economics of Education review*, *16*(2), 171-186.

Cabrera, A.F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher education: Handbook of theory and research, 10*(10), 225-256.

Korkmaz, M., Güney, S., & Yiğiter, Ş. (2012). The importance of logistic regression implementations in the Turkish livestock sector and logistic regression implementations/fields. *Harran Tarım ve Gıda Bilimleri Dergisi, 16*(2), 25-36.

Gujarati, D.N. (1995). *Basic econometrics.* 3rd ed. McGraw-Hill, New York.

Bircan, H. (2004). Logistic Regression Analysis: Practice in Medical Data. *Kocaeli University Social Sciences Institute Journal, 2,* 185-208.