

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2026): 4.49
Maths 2026; 11(2): 01-04
© 2026 Stats & Maths
<https://www.mathsjournal.com>
Received: 02-11-2025
Accepted: 06-01-2026

Obaid Mahmmood Mohsin Alzawbaee
Professor, Al-Siraj Private
University, Fallujah, Iraq

Ali Obaid Mahmmood
Lecture, Middle Technical
University - Institute of
Technical Anbar, Iraq

Ahmed Obaid Mahmood
Assistance Lecture, Middle
Technical University - Institute
of Technical Anbar, Iraq

Khalid Obaid Mahmood
Master of Engineering, Middle
Technical University - Institute
of Technical Anbar, Iraq

Corresponding Author:
Obaid Mahmmood Mohsin Alzawbaee
Professor, Al-Siraj Private
University, Fallujah, Iraq

Constructing a statistical model to identify variables that have a significant impact on COVID-19 infection

Obaid Mahmmood Mohsin Alzawbaee, Ali Obaid Mahmmood, Ahmed Obaid Mahmood and Khalid Obaid Mahmood

DOI: <https://www.doi.org/10.22271/math.2026.v11.i2a.2253>

Abstract

The Corona pandemic (COVID-19) is a widespread humanitarian crisis that has led to human suffering and pushed its economic and social well-being to the brink of collapse, a pandemic that the world has never faced.

Our research aims to study the variables that have a moral impact on the infection of the Corona pandemic (COVID-19) (by building a model), and determine the variables which have significant effect, through studying a sample with size of (213) units and a questionnaire prepared for this purpose. To build an efficient model the logistic regression technique analysis was adopted that includes the variables which have significant effect and then arranged according to their importance, and an important part of the research hypotheses was achieved, the possibility of correct classification is (70%), hence, the research reached a set of conclusions and recommendations.

Keywords: (COVID-19), logistic regression, qualitative variable, binary variable

1. Introduction

The world has faced the Corona pandemic, a pandemic that the world has never faced before, as it is fast spreading, claims many lives, and has a devastating impact on economic and social life. The world has experienced its repercussions on all segments of societies. Therefore, Covid-19 is not only described as a health crisis, but rather a humanitarian crisis on a large scale that has caused human suffering and brought economic and social well-being to the brink of collapse. To confront Covid-19, countries have mobilized their resources to confront health and economic repercussions and various aspects of life, as they have taken preventive, educational, and therapeutic measures to limit the spread of the epidemic [4, 11, 12, 13, 1, 14].

Several countries have succeeded in producing several vaccines that have contributed to limiting the spread of the pandemic, reducing infections and reducing deaths.

The importance of the research is evident in finding an effective tool through which the appropriate decision can be made regarding infection with Covid-19.

The paper aims to study the variables affecting the infection with Covid-19, and then identify the variables with a significant effect and arrange them according to their importance, as well as identify the variables with nonsignificant effect and exclude them. This is done by studying a sample and building a model that can be used to predict infection. The paper uses the binary logistic regression analysis technique to suit such studies. The importance of the study comes by providing an effective tool that depends in making decisions related to the variables that have a significant impact on infection with (COVID-19).

The research hypothesis was determined based on the significance of all independent variables as well as their parameters.

As for the research methodology, the analytical approach was adopted by applying the logistic regression analysis technique based on the foundations of statistical theory to achieve the research objective.

Several researchers dealt with the topic of the research, as well as, study ^[2] (2020) the time-series analysis technique was adopted to predict the numbers of people infected with (COVID-19) in Iraq and the countries of the Gulf Cooperation Council, where a daily time-series of (76) individual countries was studied, and a model was built for each country through which Predicting the number of daily cases, the study reached to build models for each country, in addition to the fact that the turning points were not reached, which confirms that the pandemic continues to escalate in the number of cases and that the crisis represented by the pandemic is still relatively far away.

In addition, ^[19] (2016) The study focused on identifying the factors affecting income adequacy in the Darfur region of Sudan. Five independent variables were identified, all of which were qualitative. The sample size was (136) observations. The study adopted the dual response model for data analysis purposes. The most important results of the study are that family spending has a significant impact on income adequacy. The same applies to the family size variable, which had a significant impact on income adequacy. As for other variables, such as ownership of agricultural land, the educational level of the head of the household, or the number of children in schools, they were not all significantly influential.

The study ^[16] (2012) focused on the increasing importance of the logistics model in human studies at the present time, where study aimed at guiding the reader to how to benefit from this technique and using an illustrative example on how to deal with the outputs that can be obtained from this technique by using the loud statistical programs.

2. Logistics Regression

The logistic regression model is used to predict the probability of an event occurring and uses independent variables, which can be quantitative or qualitative, and a dependent variable, by analyzing the relationship between the independent variables and the dependent variable by fitting a logistic curve (Figure. (1)).

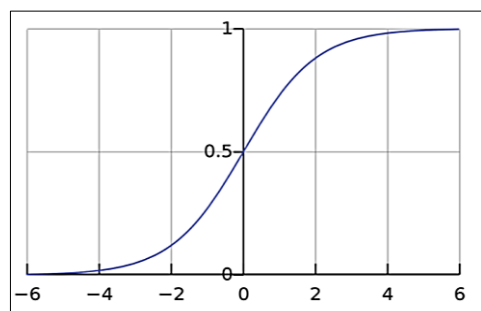


Fig 1: Logistic curve (Function)

The application of the logistic regression technique is common in medicine, social sciences, and marketing.

Logistic regression is of three types, the most important of which is binary logistic regression, as the binary logistic regression model is considered one of the nonlinear regression models, and is characterized by the fact that the dependent variable (y) is binary, with two mutually exclusive states, the first of which is the state of the occurrence of the event (y=1)

and the second state, the state of the non-occurrence of the event (y=0) where: ^[3, 5, 6, 18, 7, 9].

$$P(y=1) = P \quad (1)$$

$$P(y=0) = 1-p$$

$$0 \leq P \leq 1$$

$$\text{Thus: } 0 \leq \frac{P}{1-P} \leq \infty$$

$$\text{And: } -\infty \leq \ln\left(\frac{P}{1-P}\right) \leq \infty \quad (2)$$

$\ln\left(\frac{P}{1-P}\right)$ is called (Logit Transformation. ^[8, 10, 9, 17, 15].

Logistic regression uses the logistic function:

$$f(z) = \frac{1}{1+e^{-z}} \quad (3)$$

Where: $0 \leq f(z) \leq 1$

3. Practical side

We study a sample with a size of (220) units after checking the forms and excluding the missing forms, a sample size of (213) was adopted in the statistical analysis. See a questionnaire in appendix (1).

Search variables are:

Y = 1 if the person is infected with the virus, Y = 0 if the person is not infected with the virus. (Dependent variable)

The independent variables, and their number are (11) variables, and they are:

X₁: age

X₂: The profession where (1) was given if the profession was a lot of people mixing, such as students, teachers, doctors, and those who support them, (0) for others.

X₃: The monthly income of the family, where (0) was given to those whose income is less than one and a half million dinars per month, (1) to those whose income ranges from one and a half to less than two million dinars per month, (2) to those whose income is two million dinars or more per month.

X₄: Number of family members

X₅: Infection with a chronic disease, infection (1), uninfected (0).

X₆: gender, male (1), female (0).

X₇: Is there an increase in weight than normal, there is (1), there is no (0).

X₈: Have you received both doses of the vaccine, yes (1), no (0).

X₉: There are children under ten years of age in the family, yes (1), no (0).

X₁₀: blood type

X₁₁: Chronic disease type.

4. Analysis & Results

By using SPSS package to analyze the data, and using rotation of the -2Log likelihood function, where:

-2 Log likelihood function = 258.392

and the change in the parameters became zero at the fourth rotation, as shown in Table (1).

Table 1: Iteration history

Iteration		-2 Log likelihood	Coefficients									
			Constant	x1	x2	x3	x4	x5	x6	x7	x8	x9
Step 1	1	258.821	-.157	.043	-.237	.228	-.072	.153	-.605	.588	-.655	.174
	2	258.393	-.184	.050	-.240	.257	-.088	.181	-.687	.668	-.717	.205
	3	258.392	-.183	.050	-.239	.259	-.089	.183	-.690	.671	-.720	.206
	4	258.392	-.183	.050	-.239	.259	-.089	.183	-.690	.671	-.720	.206

The estimates of the model parameters' values were found, as well as the standard error (S.E) for each parameter, the Wald

statistic, and the significance of each parameter, in addition to the logit value. Table No. (2) shows this:

Table 2: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	x1	.050	.030	2.692	1	.100	1.051	.990	1.116
	X2	4.028	1.653	5.941	1	.015	56.169	2.201	1433.230
	X3	.259	.414	.390	1	.532	1.295	.575	2.917
	X4	4.900	1.820	7.251	1	.007	134.254	3.794	4750.518
	X5	2.318	1.047	4.897	1	.027	10.150	1.303	79.052
	X6	-.690	.315	4.812	1	.028	.502	.271	.929
	X7	.671	.444	2.286	1	.131	1.956	.820	4.668
	X8	2.582	1.495	2.983	1	.084	13.218	.706	247.403
	X9	.206	.328	.397	1	.529	1.229	.647	2.338
Constant		-.183	1.169	.025	1	.887	.833		

The chi-square test was adopted, where the log likelihood ratio follows the chi-square distribution and according to the relationship:

$$\chi^2 = 2 [\text{Loge } L_0 - \text{Loge } L_1] \quad (4)$$

L_1 : the value of Log likelihood function that contain (i) variable

L_0 : the value of Log likelihood function that contain (i-1) variable

From the results, ($\chi^2 = 22.524$) and is significant, and this demonstrates the effectiveness and efficiency of the model, as shown in Table No. (3).

Table 3: Omnibus tests of model coefficients

		Chi-square	df	Sig.
Step 1	Step	22.524	9	.007
	Block	22.524	9	.007
	Model	22.524	9	.007

Hosmer and Lemeshow test were conducted to determine the goodness of fit of the model that was built. The test depends on the difference between the observed values and the expected values. The no significance of the test, which represents the chi-square (χ^2), means that the model is of good quality and appropriate. It is noted from table (4) that the value of the ($\chi^2 = 15.036$) with a degree of freedom of (df=8) and it is not significant.

Table 4: Hosmer and lemeshow test

Step	Chi-square	df	Sig.
1	15.036	8	.058

To know the strength and ability of the model that was built to reclassify the units correctly, Table (5) shows that, as the percentage of correct classification of the model reached (70%).

Table 5: Classification table

Observed		Predicted		
		y		Percentage Correct
		0	1	
Step 1	y	0	123	91.8
		1	53	32.9
	Overall Percentage			70.0

The model with the significant independent variables that was built after estimating the parameter values (Table (2)) is as follows:

$$\text{Ln} \left(\frac{p}{1-p} \right) = -0.183 + 4.900 X_4 + 4.028 X_2 + 0.05 X_1 + 2.582 X_8 + 2.318 X_5 - 0.69 X_6 \quad (5)$$

From the same table (2), We note that there are five independent variables (X, s) that have a significant effect on the dependent variable (Y) and their order is as follows: The variable (X_4) ranked first in influencing the dependent variable (Y), while (X_2) came in second, (X_8) came in third place, the variable (X_5) came in fourth place, and the variable (X_6) came in fifth place,. As for the other four independent variables, (X_1, X_3, X_7, X_9), do not have a significant effect. And the parameter's values are:

$B_0 = -0.183$ ----- $\text{EXP} = 0.833$

$B_4 = 4.600$ ----- $\text{EXP} = 134.254$

$B_2 = 4.028$ ----- $\text{EXP} = 56.169$

$B_8 = 2.582$ ----- $\text{EXP} = 13.218$

$B_5 = 2.318$ ----- $\text{EXP} = 10.150$

$B_6 = -0.690$ ----- $\text{EXP} = 0.502$

5. Conclusion and Recommendations

The research reached a set of conclusions and recommendations, the most important of which are:

1. A statistical model was created that included variables with a significant effect, to determine whether the infection was caused by the Corona pandemic -COVID-19.
2. Some of the model hypotheses were met, others were not met, and the independent variables with a significant impact were identified.
3. The model that was built has been proven to be efficient, and also the good fitness of it was tested, and this makes it an effective tool of studying the phenomenon.
4. The model's classification ability reached (70%), which is evidence of the model's strength in discrimination between cases of infection with the Corona virus.
5. Based on the research results, researchers recommend using the logistic regression analysis technique more widely due to its effectiveness in studying many phenomena.

References

1. Achia TN, Wangombe A, Khadioli N. A logistic regression model to identify key determinants of poverty using demographic and health survey data. 2010.

2. Al Zawbaee OM, Tabra HM. Identifying factors influencing decision making using logistic regression. *General Letters in Mathematics (GLM)*. 2020;8(2).
3. Alzawbaee OM, Hamad APDAA, Mahmood ALAO. Using time series analysis to forecast the prevalence of the COVID-19 pandemic (Iraq and Gulf Cooperation Council countries (GCC)). *Journal of Critical Reviews*. 2020;7(15).
4. Andrews DW. Chi-square diagnostic tests for econometric models: theory. *Econometrica: Journal of the Econometric Society*. 1988;1419-1453.
5. Brown CE. Applied multivariate statistics in geohydrology and related sciences. [Electronic resource]. 1998.
6. Cokluk O. Logistic regression: concept and application. *Educational Sciences: Theory and Practice*. 2010;10(3):1397-1407.
7. Dutta A, Bandopadhyay G, Sengupta S. Prediction of stock performance in the Indian stock market using logistic regression. *International Journal of Business and Information*. 2012;7(1):105.
8. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.
9. Ingersoll RM. Teacher turnover and teacher shortages: an organizational analysis. *American Educational Research Journal*. 2001;38(3):499-534.
10. King JE. Logistic regression: going beyond point-and-click. 2002.
11. Lee S. Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS. *Environmental Management*. 2004;34:223-232.
12. Li H, Sun J, Wu J. Predicting business failure using classification and regression tree: an empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications*. 2010;37(8):5895-5904.
13. Litwin H, Sapir EV. Perceived income adequacy among older adults in 12 countries: findings from the survey of health, ageing, and retirement in Europe. *The Gerontologist*. 2009;49(3):397-406.
14. Mincer J. Schooling, experience, and earnings. *Human Behavior & Social Institutions No. 2*. 1974.
15. Nummela OP, Sulander TT, Heinonen HS, Uutela AK. Self-rate health and indicators of SES among the ageing in three types of communities. *Scandinavian Journal of Public Health*. 2007;35(1):39-47.
16. Osborne JW. Logits and tigers and bears, oh my! A brief look at the simple math of logistic regression and how it can improve dissemination of results. *Practical Assessment, Research & Evaluation*. 2012;17(11).
17. Petrovska T, Sova V, Voronova V, Khmel'nitska I, Borysova O, Kurdybakha O. Features of self-esteem and level of ambition in athletes of different qualifications. *Journal of Physical Education and Sport*. 2022;22(3):593-599.
18. Poston DL. Sociological research: quantitative methods. Lecture 7. 2004.
19. Rehka S. Perception of manufacturers in Perak State towards goods and services tax (GST) implementation. [Doctoral dissertation]. UTAR; 2015.

Appendix No. (1)

Questionnaire

Dears

You have all our appreciation and love, and we would like to inform you that the questionnaire that is in your hands is for the purpose of scientific research to find out the most important variables that have a scientific impact on infection with Corona pandemic (COVID-19).

Therefore, we are grateful for the answer to all the questions that we have made sure that they are simple and clear and require a little time.

Note: No need to mention the name.

Questions					
1	Have you infected with the virus?	Yes			No
2	Age	15-	25-	35-	45-
3	Occupation				
4	Monthly income	less than 1.5 M		1.5 - 2 M	More than 2 M
5	Number of family members				
6	Do you have a chronic disease?	Yes			No
7	Gender	Male			Female
8	Are you overweight?	Yes			No
9	Have you received both doses of the vaccine	Yes			No
10	Are there children under 10 years old in the family	Yes			No
11	Blood type				
12	Type of chronic disease				