Turkish Journal of Computer and Mathematics Education Vol.12 No. 4 (2021), 638-646

Research Article

Utilizing the Logistic Regression Model in Analyzing the Categorical Data of Economic Effects

[1*]Mahdi Wahhab Neamah, [2] Enas abid alhafidh Mohamed albasri, [3] Zainb Hassan Rathy

[1*] Department of Statistics, Faculty of Administration and Economics, Kerbala University, Iraq
mehdi.wahab@uokerbala.edo.iq

[2]Department of Statistics, Faculty of Administration and Economics, Kerbala University, Iraq
enas.albasri@uokerbala.edu.iq

[3]College of computer science and information technology - University of Al-Qadisiyah - Iraq
Zainb.hassan@qu.edu.iq

Article History: Received:11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: The categorical data has a significant role in representing statistical binary variables, and they are analyzed by means of grouping the response variable into ordered categories. Thereby, the dependent variable becomes of type binary qualitative variable. The data related to the financial position of world countries is classified within the categorical data. This work is to study the economic effects of an individual's different factors on determining the richness or poorness levels of a selected population of countries. Moreover, a logistic regression model is to be created to estimate these levels. As a sample of research, the categorical data relevant to the financial status of 20 Arabic countries were drawn from the website of the World Bank, WB. In addition, for comparison purpose, another similar set of categorical data was generated by MATLAB too. The paper has been based on two hypotheses, first is the well-known regression models, like the ordinary least squares or maximum likelihood, are not accurate in case of binary qualitative variables. Second, is utilizing the logistic regression model as an alternative model to achieve the paper goal. The paper results, for both WB data and MATLAB data, have successfully proved the ability of the logistic regression model in manipulating the categorical data and predicting the coefficients of the corresponding regression models.

Introduction

Qualitative variables are of binary values (0 or 1) (Yes or No) are almost based on the variable nature (e.g. colour of the eye, black or blue, / gender, male or female, etc.). Regression models of these variables cannot be accurately estimated by applying the conventional regression methods, such as the Ordinary Least Squares method (OLS). This is because the conventional models encounter several problems when used in estimating the coefficients of regression models whose dependent variables are qualitative. These problems can be summarized by; Multicollinearity, Autocorrelation and the non-homogeneous variance. [1-2][3B][4-6].

Alternatively, the logistic regression model, of binary response, is regarded as the most proper model to overcome such obstacles. For logistic regression, the predicted dependent variable is expressed by a function of the probability that a certain event will be in one of the binary categories which commonly specified by (true or false) (zero or one). Practically, it is not possible to create a regression model for binary data. Therefore, Mathematical solution is presented by the logistic regression model, LGM, by utilizing a logarithm function called "logit". This function is regarded as a transfer function to transfer the probability of binary events into non-binary regression values [16-20].

$$Y = logit = \ln\left(\frac{p}{1-p}\right) \quad \dots (1)$$

Where p is the probability that the logistic regression value is at logic "one" which means a certain event is true. In contrast, $1-p$ is the probability of the logistic regression value is "zero"; i.e., the event is false. Accordingly, the range of dependent variable value "Y" will vary from negative infinity, when p=0, to positive infinity, when p=1. Then, it becomes predictable by the conventional regression models like the OLS or the maximum likelihood, ML [2] [8-12]. So:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_n X_n \quad \dots (2)$$

Description of the data

The economic statistical data employed in this work, to achieve the paper issue, were drawn from the website of the World Bank (WB), for the year 2019. The WB publishes, on its website, a per capita Gross Domestic Product (GPD) matrix. This GDP matrix breaks down the domestic economic outputs of the world countries (per person) relative to the country population [7]. Twenty of the Middle East Arab countries were selected for the paper study from the GPD matrix.

The data under study is of a binary response-dependent variable known as "Economic Status", which is equal to 1 if the country citizen has an annual income of more than 15 thousands USD. Otherwise, the dependent variable is of zero value**.** There are five predictors X1, X2, X3, X4 and X5 to specify the person; annual income, life rate, school life, unemployment condition and the continental location(1 for Asia and 0 for Africa) respectively. The predictors X2 through X5 explicitly affect the value of X1 predictor, which was determined, in this work, as a base to define the dependent variable status.

Description of the proposed model

Figure 1 is a descriptive block diagram to illustrate the various stages of the proposed logistic regression model. The predictors X1 through X4, which have continuous values, are fed to the logistic decision block. In addition, the predictor X5 (which labelled by cont. because it is represented by "0" or "1" binary values) is also fed to this block. The output of the logistic decision block is the dependent variable in its binary form "0" or "1". This output form represents the input of the "log function" block which has to widen the range of the independent variable "Y" into (-infinity to +infinity). By this range transformation, values of "Y" become ready to be manipulated by the likelihood estimation block. The output of this block is the required logistic regression model.

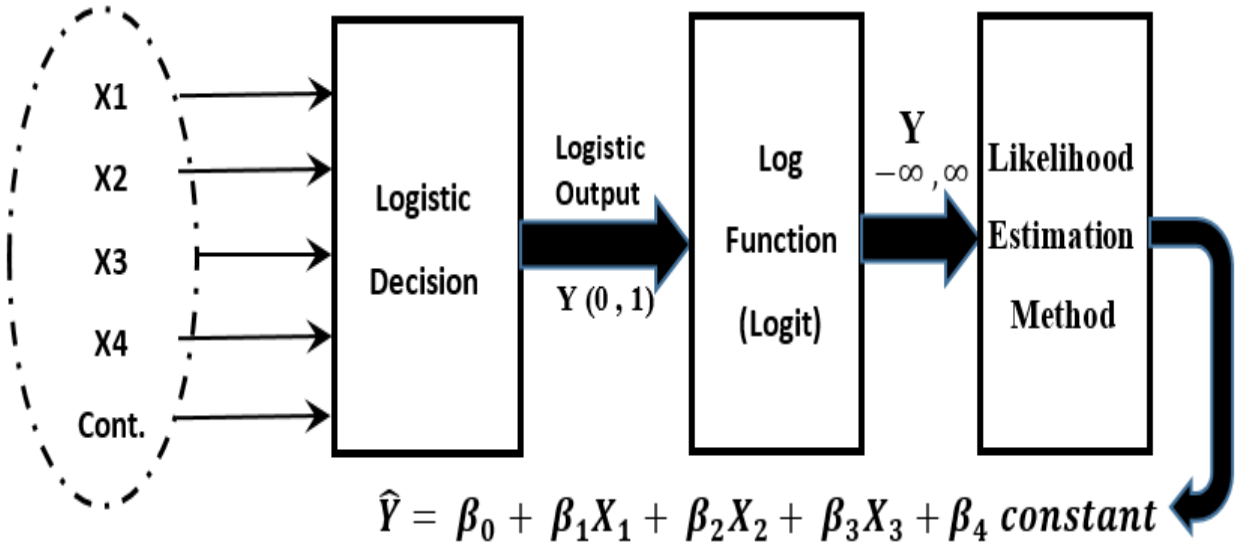

Fig. 1 The proposed logistic regression model

Results and Discussion

The data under this study is shown in appendices.1 and 2. This data was fed and processed by the statistical software SPSS. The output results of this software for the data of appendix 1 are shown and discussed by the tables shown in figures 2 through 6 given below.

Figure 2 shows the sample size utilized in this work. It tells that all the twenty input data of the twenty countries, concerned in this study, were processed and there is no any missing in the input data.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	20	100.0
	Missing Cases	0	.0
	Total	20	100.0
Unselected Cases		0	.0
Total		20	100.0

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

Fig. 2, Statistic for the undertaken sample data

Figure 3a illustrates the two states binary coding of the dependent variable, Y, (the work outcome) and its corresponding classification into explanatory categorizes poor and rich. While figure 3b points out the numbers for "poor" and "rich" cases (14 and 6 respectively) and the overall poorness percentage (70%)

(a)

Dependent Variable Encoding

Original Value	Internal Value
poor	0
rich	1

(b)

Classification Table[a,b]

			Predicted Y		
	Observed		poor	rich	Percentage Correct
Step 0	Y	poor	14	0	100.0
		rich	6	0	.0
	Overall Percentage				70.0

a. Constant is included in the model.

b. The cut value is .500

Fig. 3, Coding and classification of the dependent variable

Table for variables in the equation is given in figure 4. By the last column in this table, it can be noticed that the value 0.429 represents the right-hand side of the logit logistic function. This value comes from the following:

By figure 3, probability of true = p = 6/20 = 0.3

So, $logit\left(\frac{p}{1-p}\right) = \ln\left(\frac{0.3}{0.7}\right) = 0.429$ … (2)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.847	.488	3.015	1	.082	.429

Fig. 4, Coding and classification of the dependent variable

Figure 5 shows the table of the iteration history of estimations of the predictor coefficients. The proposed model was constructed by a procedure based on an iterative maximum likelihood, ML. The initial values of the regression coefficients, βs, were arbitrarily chosen. In each iteration, the SPSS predicted new, more accurate values for regression coefficients. Thereby, the likelihood of the observed data would be made greater under the new model coefficients. Iterations procedure continued till model converging was taken place, which means that the differences between the values of previous and current model coefficients can be are neglected. The iteration history table shows that the coefficient estimations processes proceeded for 20 steps. The table also shows the deviance statistic (-2LL). These statistics are obtained from the natural logarithm of likelihood multiplying by (-2). It represents a criterion of how the coefficient estimations are good and, correspondingly, how the logistic regression model exactly fit the data. The smaller value of this Statistic the better estimation of predictor coefficients [1] [13-15].

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients					
			Constant	X1	X2	X3	X4	X5
Step 1	1	8.945	4.684	.000	-.116	.176	-.025	-.776
	2	4.205	9.468	.000	-.211	.266	-.067	-1.136
	3	1.766	16.395	.000	-.309	.236	-.163	-1.285
	4	.684	24.241	.000	-.412	.156	-.282	-1.300
	5	.265	32.261	.001	-.520	.086	-.402	-1.264
	6	.102	40.673	.001	-.637	.027	-.529	-1.180
	7	.039	49.476	.001	-.760	-.026	-.662	-1.048
	8	.015	58.578	.001	-.889	-.075	-.800	-.879
	9	.005	67.883	.001	-1.022	-.120	-.942	-.684
	10	.002	77.318	.001	-1.158	-.160	-1.084	-.476
	11	.001	86.837	.001	-1.296	-.198	-1.227	-.260
	12	.000	96.411	.001	-1.435	-.233	-1.370	-.041
	13	.000	106.024	.002	-1.575	-.267	-1.513	.179
	14	.000	115.663	.002	-1.716	-.299	-1.656	.399
	15	.000	125.324	.002	-1.858	-.330	-1.798	.617
	16	.000	135.002	.002	-2.000	-.360	-1.941	.835
	17	.000	144.694	.002	-2.142	-.389	-2.084	1.051
	18	.000	154.397	.002	-2.285	-.418	-2.226	1.267
	19	.000	164.110	.002	-2.428	-.446	-2.368	1.481
	20	.000	173.831	.002	-2.571	-.474	-2.511	1.695

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 24.435

d. Estimation terminated at iteration number 20 because maximum iterations has been reached.

Fig. 5, History of predictor coefficient estimations

The most important table is the one given in figure 6. It demonstrates the results of the estimation of coefficients of the logistic regression model. According to these results, the estimated output (Y) of the undertaken logistic regression model is given by:

$$Y = 173.831 + 0.002X1 - 2.571X2 - 0.474X3 - 2.511X4 + 1.695X5 \quad \dots (3)$$

Fig. 6, Variables of the logistic regression equation by WB data

It is clear that the estimated logistic model is well consistent with the economic standards in financial

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X1	.002	1.171	.000	1	.998	1.002
	X2	-2.571	3369.584	.000	1	.999	.076
	X3	-.474	6871.214	.000	1	1.000	.622
	X4	-2.511	1747.026	.000	1	.999	.081
	X5	1.695	16080.307	.000	1	1.000	5.448
	Constant	173.831	164157.100	.000	1	.999	3.118E+75

developing of countries. Discussion of equation (3) can be summarised by the following points:

- The negative coefficients of the predictors X2, X3 and X4 mean that they have a negative effect on tending of output to be a state "1" "richness". Increasing the value of X2, X3 or X4 by one leads to reduce the logit of logistic regression by 2.571, 0.474 and 2.511, respectively.
- In contrast, the X1 and X5 predictors have a positive effect on bringing the output up to "1" state. Increasing each of these predictors by one will improve the opportunity of the logit regression by 0.002 and 1.695, respectively.
- According to the above two points, it is obviously clear that the person life period and its unemployment condition highly affect on reducing the output of the logistic model. In a reverse manner, the output rises with the person annual income and the Asian geographic position of the country.
- The exact output binary was categorizing of countries into poor "0", and rich "1" without any per cent of error affect the value of Wald test making it equals to zero for all estimated coefficients.

Whereas the table given in figure 7 shows the results of the estimation of coefficients of the logistic regression model according to the data generated by MATLAB simulation.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	X1	.003	.692	.000	1	.997	1.003	.258	3.894
	X2	-.130	725.789	.000	1	1.000	.878	.000	.
	X3	3.500	2766.404	.000	1	.999	33.106	.000	.
	X4	.110	302.180	.000	1	1.000	1.117	.000	1.838E+257
	X5	-1.015	9678.531	.000	1	1.000	.362	.000	.
	Constant	-67.904	68570.788	.000	1	.999	.000		

Fig. 7, Variables of the logistic regression equation by MATLAB data

So, the corresponding logistic equation for the MATLAB data is given by:

$$Y = 0.003 - 1302X1 + 3500X2 + 0.11X3 - 1.015X4 - 67.904X5 \quad \ldots (4)$$

Verification

To verify the validity of the obtained logistic regression equation given in (3), the data of the first country, for instance, are substituted in the regression equation. By this substitution yields:

$$Y = -38.24 \quad \ldots (5)$$

Taking the inverse of (logit) function yields:

$$\frac{p}{1-p} = value\ very\ close\ to\ zero \quad \ldots (6)$$

Result of equation (6) is correct if and only if the value of the probability of the country to be rich (p) is very close to zero. This means that the country whose date was substituted in the logistic regression equation (equation 3), is more likely to be a poor country. Thereby, the regression result well fit the data of the first country given in appendix 1. Similarly, if the data of the second country is substituted in the logistic regression equation, a value of (p) very close to one will be obtained. So, this country is more likely to be rich, which is consistent with this country data.

Conclusions

The paper projects a spot of light on the difficulties that may encounter the researchers when they try to apply the traditional regression tools on data of binary form. In addition, the results of this paper have confirmed the ability of the logistic regression model in dealing with the binary qualitative variables and accurately estimating the coefficients of the predicted regression model. It can be concluded that the logistic regression model is convenient in modelling the binary data because of its simplicity and its high explanatory meaning. Comparing the equations of regression models given by equations 3 and 4 has shown that the value estimated coefficients and their effects differ according to the data to be manipulated.

References

1- A. A. Tôrres FernandesI and et al., Read this paper if you want to learn logistic regression,
2- A. K. Abaas, *Using the Logistic Regression Model to Estimate the Functions of Qualitative Economic Dependent Variables*, Journal of Kirkuk University for Administrative and Economic Sciences, 2 (2012), 234-253.
3- A. AGRESTI, An Introduction to Data Categorical Analysis, Wiley Series in Probability and Statistics, United States, 2019.
4- D. L. HOFFMAN and G. FRANKE, Correspondence Analysis: Graphical Representation of n Categorical Data in Marketing Research, Journal of Marketing Research, Vol. XXIII (August 1986), 213-27
5- E. Brentari, S. Golia and M. Manisera, Models for Categorical Data: A Comparison between the Rasch Model and Nonlinear Principal Component Analysis, Statistica & Applicazioni, V (2007) 53-77
6- https://www.investopedia.com/terms/p/per-capita-gdp.asp#:~:text=Per%20capita%20gross%20domestic%20product,a%20country%20by%20its%20population., available on 18 Dec. 2020
7- https://unstats.un.org/unsd/demographic/products/socind/default.htm, available on 18 Dec. 2020.

8- J. Malar and T. Bhuvaneswari, Data Quality Measurement on Categorical Data Using Genetic Algorithm, International Journal of Data Mining & Knowledge Management Process (IJDKP), 2 (2012) 33-42.
9- Hole, Y., & Snehal, P. & Bhaskar, M. (2019). Porter's five forces model: gives you a competitive advantage. Journal of Advanced Research in Dynamical and Control System, 11 (4), 1436-1448.
10- L. D. Ambraa, O. Köksoyb and B. Simonettic, Cumulative correspondence analysis of ordered categorical data from industrial experiments, Journal of Applied Statistics 36 (2009) 1315–1328.
11- M. B. Pietrzak and et al., The Application Of Local Indicators For Categorical Data (LICD) In The Spatial Analysis Of Economic Development, Comparative Economic Research, 17 (2014) 203-220.
12- M. E. Aguilar and et al., Logistic Regression Model for the Academic Performance of First-Year Medical Students in the Biomedical Area, Creative Education, 7 (2016) 2202-2211
13- M. Mustapha, F. W. Usman and S. Yusuf, A Logistic Regression Model on Academic Performance of Students in Mathematics, Continental J. Applied Sciences 11 (2016) 1 – 15.
14- O. A. Maydeu and J. Harry, Assessing Approximate Fit in Categorical Data Analysis, Multivariate Behavioral Research, 49 (2014) 305–328.
15- Q. H. Vuong, N. K. Napier and T. D. Tran, A categorical data analysis on relationships between culture, creativity and business stage: the case of Vietnam, Int. J. Transitions and Innovation Systems, 3, (2013) 4-24.
16- R. Serban, A. Kupraszewicz and G. Hu, "Predicting the characteristics of people living in the South USA using logistic regression and decision tree," 9th IEEE International Conference on Industrial Informatics, Caparica, Lisbon, 2011, pp. 688-693.
17- S, Byron, K. Rachel and R. Chris, Practical Applications of Correspondence Analysis to Categorical Data in Market Research, 5 (1996) 56-70
18- S. A. Mingoti and R. A. Matos, Clustering Algorithms for Categorical Data: A Monte Carlo Study, International Journal of Statistics and Applications 2 (2012) 24-32.
19- S. Alija, H. Snopce and A. Aliu, Logistic Regression for Determining Factors Influencing Students Perception of Course Experience, The Eurasia Proceedings of Educational & Social Sciences (EPESS), 5 (2016) 99-106.
20- S. Mabula, Modeling Student Performance in Mathematics Using Binary Logistic Regression at Selected Secondary Schools, Journal of Education and Practice, 6 (2015) 96-103.'
21- Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121
22- X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," IEEE 7th International Conference on Computer Science and Network Technology, Dalian, China, 2019, pp. 135-139.

Appendix1: Data from the WB [7]

No.	Country	Annual Person Income ($)	Life expectancy (Years)	School life expectancy (Years).	Unemployment rate	Continental location (Asia=1 , Africa=0)	Economic status (poor=0 , rich=1)
		X1	X2	X3	X4	Cont.	Y
1	Algeria	3,974.0	72	13	8.1	1	0
2	Bahrain	23,504.0	75	13	5.6	0	1
3	Djibouti	3,414.9	57	6	54.6	1	0
4	Egypt	3,019.2	72	12	4.9	1	0
5	Iraq	5,955.1	68	12	16.2	0	0
6	Jordan	4,405.5	72	12	11	0	0

7	**Kuwait**	**32000.5**	**74**	**13**	**2**	**0**	**1**
8	**Lebanon**	**7,583.7**	**71**	**13**	**8.6**	**0**	**0**
9	**Libya**	**7,685.9**	**73**	**16**	**7.6**	**1**	**0**
10	**Mauritania**	**1,679.4**	**57**	**8**	**23.9**	**1**	**0**
11	**Morocco**	**3,204.1**	**70**	**11**	**8.4**	**1**	**0**
12	**Oman**	**15,343.1**	**71**	**13**	**1.9**	**0**	**1**
13	**Qatar**	**62,088.2**	**79**	**12**	**0.2**	**0**	**1**
14	**Saudi Arabia**	**23,139.8**	**73**	**14**	**3.5**	**0**	**1**
15	**Somalia**	**1,26.9**	**50**	**3**	**26.1**	**1**	**0**
16	**Sudan**	**4,41.5**	**60**	**10**	**18.7**	**1**	**0**
17	**Syria**	**2,032.6**	**74**	**12**	**5.7**	**0**	**0**
18	**Tunisia**	**3,317.5**	**73**	**14**	**11.9**	**1**	**0**
19	**United Emirates**	**43,103.3**	**76**	**11**	**2**	**0**	**1**
20	**Yemen**	**774.3**	**65**	**11**	**12.4**	**0**	**`0**

Appendix2: Data by the MATLAB

No.	**Country**	**Annual Person Income ($)**	**Life expectancy (Years)**	**School life expectancy (Years).**	**Unemployment rate**	**Continental location (Asia=0 , Africa=1)**	**Economic status (poor=0 , rich=1)**
		X1	**X2**	**X3**	**X4**	**Cont.**	**Y**
1	**C1**	**20,228**	**64**	**12**	**28.342**	**1**	**1**
2	**C2**	**22,418**	**56**	**8**	**48.647**	**1**	**1**
3	**C3**	**18,290**	**73**	**12**	**32.383**	**1**	**1**
4	**C4**	**38,175**	**60**	**8**	**8.596**	**1**	**1**
5	**C5**	**31,394**	**52**	**11**	**11.072**	**1**	**1**
6	**C6**	**38,599**	**80**	**8**	**22.442**	**1**	**1**
7	**C7**	**46,642**	**60**	**12**	**41.204**	**1**	**1**
8	**C8**	**48,637**	**68**	**11**	**45.425**	**1**	**1**
9	**C9**	**9,601**	**78**	**8**	**43.469**	**0**	**0**

10	**C10**	**6,943**	**64**	**6**	**17.587**	**1**	**0**
11	**C11**	**34,813**	**79**	**12**	**29.420**	**1**	**1**
12	**C12**	**4,691**	**66**	**13**	**5.038**	**1**	**0**
13	**C13**	**26,270**	**77**	**7**	**6.233**	**1**	**1**
14	**C14**	**26,517**	**78**	**11**	**7.582**	**0**	**1**
15	**C15**	**43,056**	**68**	**8**	**37.358**	**1**	**1**
16	**C16**	**24,242**	**51**	**12**	**27.285**	**0**	**1**
17	**C17**	**19,672**	**63**	**11**	**10.515**	**0**	**1**
18	**C18**	**33,571**	**67**	**6**	**27.276**	**1**	**1**
19	**C19**	**37,062**	**61**	**12**	**8.204**	**1**	**1**
20	**C20**	**26,002**	**58**	**13**	**3.118**	**1**	**1**