ISSN- 2394-5125 VOL 08, ISSUE 03, 2021

Determining the variables affecting the adequacy of the monthly income of the Iraqi family using logistic regression

Prof. Dr. Obaid Mahmmood Alzawbaee
Cihan University Sulaimanya Camp
Obed.muhsin@sulicihan.edu.krd

Ass. Lec. Ahmed Obaid Mahmood
Al-Maarif University College, Iraq
Ahmedalzawb3ee@gmail.com

Abstract:

Researchers are increasingly interested in identifying variables that have a moral impact on the adequacy of family income, because many variables are related to the level of family income that determines consumption. There are many repercussions that threaten the structure and cohesion of the family and affect the cohesion and structure of society, the most important of which is the insufficiency of income. Our research aims to identify the variables affecting the sufficiency of the monthly income of the Iraqi family by studying a sample size (383) families representing all (4696265) Iraqi families, and the technique of logistic regression analysis was adopted to build an efficient model, both in terms of independent variables with a moral effect, or determining the importance of each independent variable and arranging them according to their importance on the other hand. A large part of the research hypotheses has been fulfilled, where the number of independent variables with a moral effect on the adequacy of family income was (4) variables, while (3) variables were ineffective. It was found through the parametric and non-parametric tests the efficiency of the model that was built, and that the possibility of the model that was built on the correct classification amounted to (80.2%).

Key words: Logistic Regression, Family Income, Multiple Regression.

Introduction:

Researchers are increasingly interested in identifying the variables affecting family income, due to many variables are related to the level of family income that determines consumption and there are many and dangerous repercussions that threaten the structure and cohesion of the family and affect the cohesion and structure of society, the most important of which is the insufficient income. [1]

This research aims to determine the variables that affect the sufficiency of the monthly income of the Iraqi family through a study of a sample that represents the entire Iraqi society. Meanwhile, the technique of logistic regression analysis was adopted to determine the variables with moral impact on the one side, and arrange them according to their importance on the other side, then the research concluded with a set of conclusions and recommendations.

The importance of the research comes through providing an effective tool that can be adopted in making decisions related to the improvement and development of the Iraqi family and taking appropriate decisions in formulating policies. In addition, the research depends on the use of logistic regression technique to build an efficient model, whether in terms of independent variables with a moral effect, or determining the importance of each variable.

The research depends on the hypothesis of significant regression parameters (βi, i= 0,1, 2,…..,7) for all independent variables. The deductive analytical approach was adopted in the research by applying the basics of statistical theory and relying on the technique of logistic regression analysis to build an effective model that can be adopted in forecasting and future planning.

The subject of determining the variables in the adequacy of family income was addressed by a large number of researchers. The studies varied in terms of the quality of the data, the samples used and the nature of the instrument adopted in the analysis.

The study [7] searched the most important social and economic determinants of family income adequacy and adopted the logistic regression technique in the analysis. Besides, the study dealt with

ISSN- 2394-5125 VOL 08, ISSUE 03, 2021

seven independent variables, all are qualitative variables of binary type. The most important results were that the variable of the presence of university students in the family represents the variable with the highest moral impact on the adequacy of the family's income, the monthly income variable came in the second place in terms of the effect on the dependent variable, the housing ownership variable came in the third place, and the family size variable came in the fourth in terms of the influence on the dependent variable. On the other hand, the study shows that the variables of educational level, Place of residence and employment status insignificant influence on the dependent variable.

The [15] study aimed to identify the variables that determine the length of cancer patients staying in the hospital, and the research relied on a sample of (73) cancer patients in Baghdad governorate during the period 2010-2011. The study used the technique of logistic regression analysis and concluded that the two variables (treatment, anemia) are the two variables affecting the dependent variable.

According to [1] his study relied on the technique of logistic regression analysis and multiple discriminant analysis, and it was applied to a sample of (545) heads of household in the state of Khartoum in Sudan. A trade-off was made between logistic regression analysis and multiple discriminant analysis in order to study the most important factors affecting the adequacy of family income. The study reached a number of results, the most important of which is that the logistic model is better than the discriminant function model in analyzing data with qualitative dependent variables.

In [4] The logistic regression technique has been used to find out the most important economic and social determinants of family income adequacy in Kirkuk governorate \ Iraq, through a sample size of (250) household heads, and nine independent variables were identified. Hence, the study concluded that there are three variables that have a significant effect on the dependent variable, which are the family's monthly income, the family's housing in the event that if it is owned or rented, and finally when the head of the family is working.

1.1- Logistic Regression:

Logistic regression is used to build a relationship between the qualitative dependent variable and the independent variables, whether these variables are qualitative or quantitative. Logistic regression depends on a basic assumption that the dependent variable (y) is a binary variable that takes the value (1) with a probability (p), and the value (0) with a probability (q=1-p) and it is: [9] [6]

$$E(y \mid x) = p(y = 1) = p \text{ ----------(1)}$$

$$\text{Where } 0 \leq p \leq 1$$

In order for the model to be applicable from the regression point of view, it is

$$0 \leq \frac{p}{1-p} \leq \infty$$

And by Taking the natural logarithm, can get

$$-\infty \leq \ln\left(\frac{p}{1-p}\right) \infty \leq \text{---------- (2)}$$

Where $\ln\left(\frac{p}{1-p}\right)$ called Logit Transformation

$$\text{Ln}\left(\frac{p}{1-p}\right) = B_0 + \sum_{i=1}^{n} Bj\ X_{ij} \text{ ---------- (3)}$$

$$j = 1, 2, \ldots, k$$

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

Logistic regression is considered one of the powerful tools because it provides a test of the parameter's morale and gives an idea for the researcher about the influence of the independent variable on the dependent one, in addition to that it arranges the effects of the independent variables according to their

ISSN- 2394-5125 VOL 08, ISSUE 03, 2021

importance. Plus, the logistic regression is less sensitive to deviations from the normal distribution and the linearity relationship of the research variables as in the regression analysis. [1] [14] [2]

The logistic regression analysis model is used to predict the probability of a particular event occurring by reconciling the data in the form of a logistic curve, meaning that the regression depends on the logistic curve, which takes the following form:

$$P = \frac{e^{a+Bx}}{1+ e^{a+Bx}} = \frac{1}{1+ e^{a+Bx}} \text{ ---------- (4)}$$

Instead of the straight line equation ($y = B_0 + B_1x + e$)

[4] [13]

1.2- Practical side:

1.2.1 The research sample:

The research sample was represented by (383) families that were withdrawn from the total number of Iraqi families (4696265), and by applying (Stephen *Thompson's* equation) to determine the sample size:

$$n = \frac{N*pq}{[\left[(N-1)*\frac{d^2}{z^2}\right] +pq]} \text{ ... ---------- (5)}$$

where p=q=0.5

d=0.05 error

z=1.96

$$n = \frac{4696265 * (0.5)(0.5)}{[\left[(4696264) * \frac{0.0025}{3.8416}\right] + 0.25]} \text{ ...}$$

n =383

The data was collected through a questionnaire prepared for this purpose (appendix No. (1)).

The dependent variable (y) and seven independent variables were identified as follows:

Y is the dependent variable, which is a qualitative variable since:

Y = 1 if the family's monthly income is more than or equal to (1,500,000) one million and five hundred thousand dinars.

Y = 0 if the family's monthly income is less than (1,500,000) one million and five hundred thousand dinars.

independent variables:

X1 = the amount spent per month on foodstuffs

X2 = the amount spent per month on housing and lighting

X3 = the amount spent per month on transportation

ISSN- 2394-5125 VOL 08, ISSUE 03, 2021

X4 = the amount spent per month on clothes and shoes

X5 = the amount spent per month on furniture and household appliances

X6 = the amount spent per month on other miscellaneous things

X7 = the number of family members

1.2.2 The Results:

The SPSS program was used to analyze the search data and for the purpose of estimating the model parameters, iteration was adopted to derive the maximum possibility function for the least negative value twice the logarithm of the greatest possibility function (-2log likelihood), where the optimal estimate of the model parameters was obtained, and that:

(- 2log likelihood = 357.042)

In the fifth cycle, which we stopped at because the change in parameters (B_0, B_1, ..., B_7) became almost zero, and therefore it is considered the best result that can be obtained for the parameters and that (-2log likelihood) is at their minimum ends, as shown in table no. (1)

Table No 1. Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients Constant	x1	x2	x3	x4	x5	x6	x7
Step 1	1	380.741	-1.367	.002	.000	.002	.003	-.001	.001	-.092
	2	359.672	-1.991	.003	.001	.002	.003	-.002	.002	-.131
	3	357.086	-2.315	.004	.002	.002	.003	-.002	.002	-.145
	4	357.042	-2.365	.004	.002	.002	.004	-.002	.002	-.147
	5	357.042	-2.366	.004	.002	.002	.004	-.002	.002	-.148

a. Method: Enter
b. Constant is included in the model.
c. Initial -2 Log Likelihood: 493.358
d. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

the values of the parameters were found as table no. (2) shows the parameters of the model (column B), the standard error (S. E) for each parameter, and Wald Statistics, in addition to the significance of the parameters (sig) and the exponential value of the parameter (EXP(B)). which will be explained later.

Table No 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	.004	.001	31.807	1	.000	1.004	1.002	1.005
	x2	.002	.001	8.692	1	.003	1.002	1.001	1.004
	x3	.002	.001	1.876	1	.171	1.002	.999	1.005
	x4	.004	.001	6.784	1	.009	1.004	1.001	1.006
	x5	-.002	.001	2.519	1	.112	.998	.996	1.000
	x6	.002	.001	2.433	1	.119	1.002	1.000	1.004
	x7	-.148	.050	8.745	1	.003	.863	.782	.951
	Constant	-2.366	.465	25.926	1	.000	.094		

a. Variable(s) entered on step 1: x1, x2, x3, x4, x5, x6, x7.

ISSN- 2394-5125 VOL 08, ISSUE 03, 2021

In order to test the efficiency of the model and its quality (goodness of fit), where the test depends on the Chi-Square $\chi 2$ distribution, because in the case of logistic regression the log like hood ratio is used, where:

$$\chi^2 = 2\{\log_e l_0 - \log_e l_1\}$$

where:

L_1: the value of the maximum possibility function that contains a variable (i)

L_0: the value of the maximum possibility function that contains a variable (i - 1)

The value of ($\chi 2 = 136.315$) is significant. As shown in the following table no. (3): [10]

Table No 3. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
	Step	136.315	7	.000
	Block	136.315	7	.000
	Model	136.315	7	.000

For the purpose of knowing the quality of the model fit, a non-parametric test was adopted for the quality of the model fit and depends on calculating the ($\chi 2$) statistic for the difference between the observed values and the expected values, to detect deviations of the logistic model. The statistic for this test consists of an observed part that is not based on a theoretical model and an expected part computed from the logistic model estimates. A statistic ($\chi 2$) is calculated for the quality of fit. Where the test works on the basis that the built model is a model that has a high fit quality when the result of the ($\chi 2$) test is not significant, and table no. (4) represents the test: [11]

Table No 4. Contingency Table for Hosmer and Lemeshow Test

		y = .00		y = 1.00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	35	36.004	3	1.996	38
	2	34	34.376	4	3.624	38
	3	34	33.080	4	4.920	38
	4	30	31.464	8	6.536	38
	5	30	29.599	8	8.401	38
	6	27	27.230	11	10.770	38
	7	30	23.954	8	14.046	38
	8	18	19.381	20	18.619	38
	9	11	12.949	27	25.051	38
	10	2	2.963	39	38.037	41

It is the sums of the binary dependent variable (y) with the sums of the estimated probability. And from the following table no. (5), can notice that the value of (H- Statistics = 5.622), which represents the value of ($\chi 2$) with a degree of freedom (8), which confirms the acceptance of the null hypothesis (H_0) as (Sig = .612) by df = 8), and this confirms the fit-quality of the whole model.

Table No 5. Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.312	8	.612

ISSN- 2394-5125 VOL 08, ISSUE 03, 2021

In order to find out the extent of the possibility of the model that was built on the classification of vocabulary, the following table no. (6) shows this, and can note that the percentage of the correct classification amounted to (over all percentage = 80.2%), meaning that the number of correctly classified observations is (307), while There are (76) items classified incorrectly.

Table No 6. Classification Tablea

			Predicted		
			y		Percentage
	Observed		.00	1.00	Correct
Step 1	y	.00	232	19	92.4
		1.00	57	75	56.8
	Overall Percentage				80.2

a. The cut value is .500

Referring to table no. (2), can note that the column (β) contains the model parameters in (Log-odds) units, and the model that was built:

$$\log\left(\frac{p}{1-p}\right) = -2.366 + 0.004X_1 + 0.002\ X_2 + 0.002X_3 + 0.004\ X_4\ -0.002X_5+0.002X_6-0.148X_7 \text{-- (6)}$$

And the results shown in the same table indicate that the independent variables that have a significant effect on the adequacy of the monthly income of the Iraqi family are: (X_1, X_2, X_7, X_4), plus the constant.

Thus, the independent variables (X_1, X_2, X_7, X_4) are the variables affecting the dependent variable (y), and the value of the parameters are:

B_0 = -2.366 ------ odd = 0.094

B_1 = 0.004 -----odd = 1.004

B_2 = 0.002 ----- odd = 1.002

B_7 = -0.148 ----odd = 0.863

B_4 = 0.004 ----- odd = 1.004

- The variable (X_1), which represents (the amount spent monthly on foodstuffs) ranked first in the effect on the variable (y) and that this variable has a significant parameter (sig = 0.00), and that (wald = 31.807), and (SE = 0.001), and the value (EXP(B)=1.004).

- The variable (X_4) (the amount spent per month on clothes and shoes) came in second place by influencing the dependent variable (y) where (EXP(B) = 1.004), and that this variable has a parameter with a significant effect (sig = 0.009), and that (Wald = 6.784), and (SE = 0.001).

- The (X_2) variable which is (the monthly amount spent on housing and lighting), came in third place by influencing the dependent variable (y), where (EXP(B) = 1.002), and that this variable has a parameter with a significant effect (sig = 0.003), and that (wald=8.692) and (S.E.=0.001).

- The variable (X_7), which represents the number of family members, ranked fourth in the influence on the dependent variable (y), where (EXP(B) = .863), and this variable has an opposite effect, and it has a parameter with a significant effect (sig = 0.003), and surely (wald = 8.745) and (S.E.=0.050).

- As for the rest of the independent variables (X_3, X_5, X_6), has no significant effect on the dependent variable (y).

1.3- Conclusions and Recommendations:

The most important conclusions reached by the research are:

ISSN- 2394-5125 VOL 08, ISSUE 03, 2021

1- Some of the research hypotheses were achieved as the parameters (B_0, B_1, B_2, B_4, B_7) had significant values, however, the other research hypotheses were not achieved, as the parameters (B_3, B_5, B_6) have non-significant values.
2- It is possible to adopt the logistic regression technique to predict the adequacy and insufficiency of the monthly income of the Iraqi family through the model that was built.
3- The number of independent variables with a significant effect is four, and according to their order of importance, they are:
   (X_1) The amount spent per month on foodstuffs
   (X_4) The amount spent per month on clothes and shoes.
   (X_2) The amount spent per month on housing and lighting
   (X_7) The number of family members - which has the opposite effect and confirms the logicality of the model.

And the values of the parameters of the above variables are:

$$B_1 = 0.004 - B_2 = 0.002 - B_7 = -0.148 - B_4 = 0.004$$

4- The tests proved the significance of the built model, as the value of ($\chi 2$) calculated from the log likelihood ratio reached (136.315) with a degree of freedom (7), which is significant.
5- The correct classification rate for the proposed model was (80.2%).
6- The quality of the fit of the model with the non-parametric test (Hosmer and Lemeshow), where the results showed non-significance of differences, which confirms the quality of the fit of the built model.
7- Expand the use of logistic regression technique as an effective method for determining the adequacy or insufficiency of the Iraqi family's income.

References:

[1] Abbas, A. K. (2012). "Use of logistic regression model in the prediction of functions with economic variables of specific nature". Journal of Kirkuk University for Administrative Sciences and Economics. 2(1).

[2] Ahan, A.E; & Okafor, R. (2010): Application of Logistic Regression Model to Graduating (CGPA of University Graduate-University of Lagos). Journal of Modern Mathematics and Statistics, 2(2), pp. 58 – 62.

[3] Alzawbaee &others , Obaid Mahmmood & Hatem Hatef (2019)-(Using logistic regression to study the most important determinants of household income adequacy: Field study on a sample of households

in Kirkuk / Iraq)- The Scientific Journal of Cihan University – Sulaimaniya (SJCUS) –vol(3) No.(1)-Jon-2019-[ISSN 2520-7377 (online),ISSN 2520-5102 (print)]

[4] Dutta A., and Bandopadhyay G.,(2012). Performance in the Indian Stock Market Using Logistic Regression", International Journal of Business and Information, Vol . 7, No. 1, June, 105-136.

[5] Dutta A., and Bandopadhyay G.,(2012). Performance in the Indian Stock Market Using Logistic Regression", *International Journal of Business and Information*, Vol . 7, No. 1, June, 105-136.

[6] Fagoyinbo, I.S, Ajibode, I.A., Olaniran, Y.O.A , (2014). The Application of Logistic Regression Analysis to the Cummulative Grade Point Average of Graduating Students: A Case Study of Students' of Applied Science, Federal Polytechnic, *Ilaro-Developing Country Studies*, Vol.4, No.23, 26-30.

[7] Ghanem, A. & Al-Ga'ouni, F. K. (2011). "Use of binary regression technique in the study of the most important economic determinants of family income efficiency". 27(1), 113-132.

[8] Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*, Wiley, N. Y, 1989.

[9] Lee, S. (2004). "Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS", Environmental Management, Vol. 34, No. 2, 223-232

[10] Li, H., Sun, J. and Wu, J. (2010). "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, Vol. 37, No. 8, August, 5895- 5904.

[11] Li, H., Sun, J. and Wu, J. (2010). "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*, Vol. 37, No. 8, August, 5895- 5904.

[12] Litwin, H., & Sapir, E. V. (2009). Perceived income adequacy among countries: Findings from the survey of health, ageing, and retirement in Europe. The Gerontologist, 49, 397–406. doi:10.1093/geront/gnp036.

[13] Nummela, O. P., Sulander, T. T., Heinonen, H. S., & Uutela, A. K. (2007). Self-rated health and indicators of SES among the ageing in three types of communities. Scandinavian Journal of Public Health, 35, 39–47. doi:10.1080/14034940600813206.

[14] Osbome W., Jason, (2012). Logits and tigers and bears, oh my! A brief look at the simple math logistic regression and how it can improve dissemination of results, Practical Assesment, Research & Evaluation,17(11), 1-10.

[15] Saleh, A. H. (2013). "Analysis of logistic regression to study the survival time of patients with leukemia". Management and Economics Journal/ University of Mustansiriya. 3(8).

[16] Stoller, M. A., & Stoller, E. P. (2003). Perceived income adequacy among elderly retirees. Journal of Applied Gerontology, 22, 230–251.