ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 10, No. 2, April 2022, pp.260-266

Determine the most important factors affecting breast cancer in Iraq

Wadhah S. Ibrahim[1], Ghiath Hameed Majeed [2], Jinan Abdullah Anber [3]
[1]College of Management and Economics, Dep. of Statistic, Mustansiriyah University, Iraq
[2]College of Basic Education, Dep. of Computer Sciences, Mustansiriyah University, Iraq
[3]Baghdad Technical College of Management, Dep. of Information Techniques, Middle Technical University, Iraq

ABSTRACT

Today cancer is one of the diseases leading to death in the world in general and in the Middle East in particular, according to the reports of the World Health Organization. In spite of this, early detection through the availability of the necessary data on the disease in patients and the variety of variables studied helps in the large percentage of treatment, and that Increasing awareness among members of the Iraqi community, especially women, about breast cancer, proves its ability to recover from this disease through early and periodic examination and self-examination, and it was found that the number of injuries in terms of ages was for the second category (70-40), and this indicates that menstruation or what is known as the stage of hope is the stage that females must do early and periodic examination of this disease, and that the diagnosis of the disease was a clinical test or tissue culture, and therefore it was recommended to spread awareness for early and self-examination to detect the disease early and this greatly helps for treatment, with changing the data registration form for this type of disease to include more important variables, especially the genetic factor variable.

Keywords: Breast cancer, nonlinear logistic regression model, maximum likelihood

Corresponding Author:
Wadhah S. Ibrahim,
College of Management and Economics,
Dep. of Statistic, Mustansiriyah University
Baghdad, Iraq
E-mail: dr_wadhah_stat@uomustansiriyah.edu.iq

1. Introduction

According to previous studies of the World Health Organization, cancer is one of the most important leading causes of death among adults in the world in general and in the Middle East region in particular [1] [2] and that the increasing incidence of this disease over the years poses a risk of the possible spread of the disease through the multiplication and aging of the population. Despite this a large part of cases of this disease can be cured if detected in the early stages of the injury and breast cancer is one of those cases that can be treated with early examination [5].

Breast cancer is the most common type of cancer among females in the world in general and in the Middle East region and Iraq in particular as the disease in Iraq represents about a quarter of the incidence of other types of Iraqi women (23%) [8]. According to the latest update of the log data Iraqi cancers, this species occupies the first place among other types that can be affected by the Iraqi citizen [7].

Obtaining accurate data for variables may help the researcher to study the increase in the incidence of breast cancer in Iraq is one of the difficult things, because health institutions overlook many of the variables that relate to that increase rate, and these variables are genetic factors, increased body mass and environmental pollution due to wars, the deterioration of health status and lack of awareness of the Iraqi family and others, which affect the reality of this type of disease, and the impact of the reality of the disease on Iraqi women, the family and society, and these variables cannot be obtained for the purpose of the study.

The research aims to study the most important factors affecting the incidence of breast cancer in Iraq and to briefly identify the nature of this tumor in particular through the use of the non-linear logistic regression model and according to the data obtained from the Ministry of Health / Cancer Center / Statistics Division.

2. The nonlinear logistic regression model

The non-linear logistic regression model is the most common type of regression model in the process of data analysis which is used in various areas of life including the medical field as this model examines the relationship between the response variable and one or more independent variables [3][9]. The non-linear logistic regression models are divided into two type. The first is the binary logistic regression model which is used in the event that the response variable is composed of two levels and the second type is the multi-response logistic regression model which is used in the event that the response variable is composed of more than two levels [4]. The first type of this model will be used, depending on the condition of the person with the disease (live or death) as the response variable takes the value (1) in the case of live and the value (0) in the case of death this relationship is represented by the following formula [11]:

$$P = \frac{e^{\acute{X}\beta}}{1 + e^{\acute{X}\beta}} \qquad -\infty < X < \infty \qquad -\infty < \beta < \infty \qquad (1)$$

Where P represents the probability of the response and its value is limited. β: represents the parameter of the model to be estimated. X: the array of independent variables.

3. The maximum likelihood method

For the purpose of estimating the parameters of the non-linear logistic regression model that makes those at their maximum ends, the maximum likelihood method was used using the Newton-Raphson method shown as follows [4]:

$$\underline{\hat{\beta}}_{r+1} = \underline{\hat{\beta}}_{r} - (X'WX)^{-1} X'(Y - \hat{Y}_r) \qquad (2)$$

Where: $\underline{\hat{\beta}}_{r+1}$: represents the vector of the values estimated within the range (r + 1). $\underline{\hat{\beta}}_{r}$: the vector of values estimated within the range (r). X: the array of independent variables of rank (r * (D + 1)). W: is a diagonal array of rank (r * r) and is as follows [10]:

$$W = \begin{bmatrix} m_1\hat{p}_1\hat{q}_1 & 0 & \cdots & 0 \\ 0 & m_2\hat{p}_2\hat{q}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_r\hat{p}_r\hat{q}_r \end{bmatrix} \qquad (3)$$

4. Wald test

The Wald test is one of the tests used to test the significance of the parameters of the non-linear logistical model as it is based on testing the independent variables separately with the dependent variable and the test statistic is represented by the following formula [6]:

$$W = \frac{\hat{\beta}}{S.E.(\hat{\beta})} \qquad (4)$$

Where: $\hat{\beta}$: the parameter of the model is estimated. $S.E.(\hat{\beta})$: The standard error of the model parameter. This type of test follows the Chi-Square distribution and (d.f.=1).

5. The Hosmer-Lemeshow test

It is a good conformance test but it is used for non-linear logistic regression models, especially risk prediction models, and it shows the suitability of data with the model used and that these types of tests are used in the event that the dependent variable is of the binary type. We rearrange the data according to the expected probabilities and form a number of groups G. Therefore the test statistics are represented as follows [4] [9]:

$$\chi^2 = \sum_{i}^{1}\sum_{j}^{G}\frac{(F_{ij}-e_{ij})^2}{e_{ij}} \qquad (5)$$

It is clear that the test formula is the Chi-Square test formula and that F_{ij}: represents the observed frequency and e_{ij}: represents the expected frequency and that the degree of freedom of the test is (G-2).

6. The practical side

The research data was obtained from the Ministry of Health / Cancer Center / Statistics Division. The number of people with breast cancer (4116) for the year 2017 and the variables taken were the type Gender, Age, Adders Code, Occupation, detection of disease and that the dependent variable represents the patient's condition (death, live).

Table 1. Case Processing Summary

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	4116	100.0
	Missing Cases	0	.0
	Total	4116	100.0
Unselected Cases		0	.0
Total		4116	100.0

Table 1 shows the number of patients with breast cancer (n = 4116) which shows that there is no data missing.

Table 2. Gender

		Frequency	Percent
Valid	Male	91	2.2
	Female	4025	97.8
	Total	4116	100.0

Table 2 shows the percentages of the number of patients by type (male and female)), the percentage of gender with a female class is very large (97.8).

Table 3. Status

		Frequency	Percent
Valid	Death	328	8.0
	Live	3788	92.0
	Total	4116	100.0

Table 3 shows the encoding of the dependent variable with the percentage for each type.

Table 4. Classification Table [a]

Observed			Predicted		
			Status		Percentage Correct
			Death	live	
Step 0	Status	Death	0	328	.0
		Live	0	3788	100.0
	Overall Percentage				92.0
a. Constant is included in the model.					

Table 4 shows that the person with breast cancer can be classified as alive (92.0%) which is a fixed model without explanation of other variables.

Table 5. Variables in the equation

		B	S.E.	Wald	df	Sig.
Step 0	Constant	2.447	.058	1806.871	1	.000

Table 5 shows the estimation of the fixed term and Wald statistics for the zero step as it shows the efficiency and significance of the model.

Table 6. Variables not in the equation

Variable, step 0	Score	df	Sig.
Age(10-40)	33.003	1	.000
Age(40-70)	37.020	1	.000
BASRAH	13.456	1	.000
NINEVHA	13.303	1	.000
MAYSAN	30.452	1	.000
ERBIL	7.425	1	.006
DIYALA	4.162	1	.041
ALANBAR	11.901	1	.001
KIRKUK	8.040	1	.005
SALAHELDIN	71.400	1	.000
ALNAJAF	22.655	1	.000
House wife	3.947	1	.047
Basis(Others)	16.735	1	.000
Grade(1)	92.558	1	.000
Grade(2)	43.258	1	.000
Grade(3)	6.670	1	.010
Grade(4)	200.533	1	.000
Cytology Heamatological	29.352	1	.000
Histology of a metastasis	12.921	1	.000
Histology of a primary	416.413	1	.000
Basis(Unknown)	19.681	1	.000

Table 6 shows the type of variables that have an effect on the model, and the internal classifications of those variables, and according to statistical significance, excluding the variables that have no effect.

Table 7. Omnibus tests of model coefficients

		Chi-square	df	Sig.
Step 1	Step	1523.361	42	.000
	Block	1523.361	42	.000
	Model	1523.361	42	.000

Table 7 shows the value of the Chi-square test which represents the acceptance of entering variables into the model.

Table 8. Model summary

	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
Step 1	765.211	.309	.725

Table 8 shows that the changes in the dependent variable for the differences that can be explained by (72.5%) and the rest are random changes and this indicates that the model is compatible with the data analysis.

Table 9. Hosmer and Lemeshow test

	Chi-square	df	Sig.
Step 1	8.284	8	.046

Table 9 shows the value of Hosmer and Lemeshow Test which is significant according to statistical significance.

Table 10. Classification table

Observed			Predicted		
			Status		Percentage Correct
			Death	Live	
Step 1	Status	Death	235	93	71.6
		live	14	3774	99.6
	Overall Percentage				97.4

Table 10 shows that the classification of data increased to (97.4%) when entering the variables affecting the model from what is found in Table 4.

Table 11. Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	Gender(1)	1.057	.617	2.932	1	.087	2.877
	Age(40-10)	.478	.457	1.092	1	.296	1.613
	Age(70-40)	-.064-	.481	.017	1	.895	.938
	NINEVHA	-2.419-	.359	45.413	1	.000	.089
	KARBALLA	-2.064-	.348	35.114	1	.000	.127
	WASIT	-1.592-	.554	8.270	1	.004	.203
	SALAHEDIN	-3.801-	.378	100.848	1	.000	.022
	Occup(Others)	-.972-	.262	13.711	1	.000	.378
	Grade(3)	-3.438-	1.490	5.328	1	.021	.032
	Grade(4)	-3.126-	1.465	4.552	1	.033	.044
	Clinical investigation	6.414	.746	73.921	1	.000	610.100
	Surgery/autopsy	6.800	.647	110.504	1	.000	897.971

		B	S.E.	Wald	df	Sig.	Exp(B)
	Cytology Heamatological	8.812	.628	196.621	1	.000	6714.688
	Histology of a primary	7.702	.479	258.531	1	.000	2212.815
	Autopsy With Histology	4.962	.967	26.301	1	.000	142.837
	Basis(Unknown)	6.011	.502	143.437	1	.000	408.021
	Constant	-1.654-	1.688	.960	1	.032	.191

Table 11 shows an estimate of the parameters of the non-linear logistic model and what variables affect the variable adopted in the model.

7. Conclusion

From observing the results, it is clear that the variety has an effect of developing breast cancer as the incidence of one male is offset by an increase in the number of female infections by a large amount (287%) which is a large percentage and that the number of injuries in terms of ages ranged between the first category (10-40) and the second (40-70) and this indicates that females are more affected than males of this age and for several reasons and that the residential areas were distributed between the south (KARBALLA, WASIT), and the north (NINEVHA, SALAHEDIN) as shown in table (11). Occupations have no significance in terms of injury but in terms of the size of the tumor discovered were of the third and fourth degree and this indicates that the diagnosis of cases is not early and that the diagnosis process relied on clinical investigation as well as surgical procedures, anatomy and tissue culture.

From observing the results, it can be said that the small number of variables recorded by the health institution places a study of the factors affecting breast cancer in Iraq in a difficult way, as it leaves variables of great importance, including genetics, environmental pollution and others.

Spreading health awareness among the community members in terms of early examination and increasing awareness of early self-awareness for both sexes may help in not exacerbating infection and help in recovery, registration of data by health institutions for variables of great importance to know the factors affecting such type of diseases, and change the registration forms to collect the largest number of these variables.

Declaration of competing interest

The authors declare that they have no any known financial or non-financial competing interests in any material discussed in this paper.

References

[1] K. M. N. Al Asadi , "Women's Knowledge and Concern about Breast Cancer" *kufa Journal for Nursing sciences*, vol.4, no.1, pp. 121-130, 2014.

[2] I. K. Al-shibly, M. A. Muhsin, and M. S. A. Razak, "Immunological Study on Breast Cancer In Hilla Province", *Karbala Journal of Medicine*, vol.4, no.10, pp.1140-1145, 2011.

[3] D. Bertsimas and A. King,“Logistic regression: From art to science”. Statistical Science, vol. 32, no. 3, pp. 367-384, 2017.

[4] H. T. Falah, " Estimation of the Logistic Regression model and the Severity function of Cox Regression models – Case study", the Higher Diploma in Biostatistics, College of Administration and Economics / University Al-Mustansiriyha, 2018,.

[5] S. S. Hassan, W. H. Yousif, and A. N. AL-Thaweni, "Detection of BRCA1and BRCA2 mutation for Breast Cancer in Sample of Iraqi Women above 40 Years", *Baghdad Science Journal*, vol. 7, pp.400-394, 2010.

[6] S. A. Hussein, "The Strongest Possibilities of the Strongest Weighted and Comparable with Other Methods of Logistic Model with Practical Application", Master Thesis in Statistics, College of Administration and Economics / University of Baghdad, 2009.

[7] M. N. Mezher, A. S. Dakhil, and D. H. Abdul-Jawad, "Role of epstein-barr virus (EBV) in human females with breast cancer", *Journal of Pharmaceutical Sciences and Research*, *vol.9*, no.7, p.1173, 2017.

[8] F. H. Mualla, and N. A. Al-Alwan, "Promoting clinical breast examination as a screening tool for breast cancer in Iraq", *Nursing national Iraqi specility*, *vol.27*, no.1, pp.76-82, 2014..

[9] A. T. Raheem, H. A. Rasheed, and G. H. Majeed, "Constructing a model to determine the most important factors affecting diabetes disease", *Periodicals of Engineering and Natural Sciences*, vol.*9*, no.4, pp.481-490, 2021.

[10] W. S. Ibrahim, "To determine the most important factors affecting pancreatic cancer in Iraq using the logistic regression model", *Periodicals of Engineering and Natural Sciences.* Vol.7 No.2 pp.10-18, 2019.

[11] A. J. Mohammed, W. S. Ibrahim and, "Study of the probability of change in share prices based on the value of the Babylon Hotel using logistic regression", *Journal of College of management and Economics*, Vol. 43, No.123 434-446, 2020.