Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 1, March 2020, pp.277-283

To determine the most important factors affecting pancreatic cancer in Iraq using the logistic regression model

Wadhah S. Ibrahim

Mustansiriyah University, College of Management and Economics

ABSTRACT

Cancer is one of the most prevalent malignant diseases in the world, and despite the efforts to find a cure for this disease, but the disease takes a lot of human lives, this study is related to pancreatic cancer, which is the fifth most common in the world, and the cause of death in many Sometimes, The data were obtained from the Ministry of Health / Cancer Center / Statistics Division, the number of people infected with the disease reached (509) for the year 2017, excluding a number of patients due to lack of information about them, which affect the results of the study which shows according to available changes the extent of their impact on this The number of variables was (7) variables by using logistic regression model and It is noticed in many statistical studies that variables with influence on disease are confused with variables that help in the healing or treatment process of the injury.

Keywords: Pancreatic cancer, Logistic regression model, Maximum likelihood

Corresponding Author:

Ass. Prof. Dr. Wadhah S. Ibrahim
Mustansiriyah University / College of Management and Economics
Baghdad, Iraq
E-mail: dr_wadhah_stat@uomustansiriyah.edu.iq

1. Introduction

Cancer is one of the most common malignant diseases in the world. The study will address pancreatic cancer, the fifth most common cause of death in the world [1]. It was revealed that fewer than 10% of patients survive for more than a year after diagnosis of the disease. The 5-year survival rate (0.4%) is the lowest of any disease Other cancer.

It consists of irregular genes in pancreatic tissue and abnormal cells resulting from environmental or genetic causes [2, 3].

It is made up of irregular genes in pancreatic tissues that are abnormal cells from environmental or genetic causes. It tends to spread very quickly and is detected only in very early stages. This is the main reason that cancer causes the highest proportion of deaths from cancer. Often, the symptoms of pancreatic cancer do not appear until after the cancer has reached a relatively advanced stage, and the tumor can no longer be removed by surgery.

Pancreas is a large internal organ located horizontally in the posterior lower part of the abdominal cavity [4]. The pancreas secretes enzymes that aid in digestion and hormones to regulate the metabolism of sugars (carbohydrate) in the body.

The pancreas is 15 centimeters in size and resembles, in its shape a pear lying on its side. The pancreas is a central organ of the digestive system. It secretes hormones which include insulin which help in the treatment of sugars. It also produces digestive acids that aid in the digestion of food [5].

Pancreatic cancer begins to form in the tissues of the pancreas. Symptoms of pancreatic cancer often do not appear until pancreatic cancer has reached a relatively advanced stage.

2. Research problem

In recent years the prevalence of cancer has been observed for a number of reasons, including wars, the environment, and other causes. More than 140,000 Iraqis are currently suffering from cancer and malignant tumors. This led to the importance of the subject in prevention, diagnosis, early treatment and improvement of the lifestyle of patients with cancer, which included the importance of research, and the reality of the spread of the disease through the use of logistic regression model.

3. Aim of the research

The aim of this research is to study the statistical factors affecting cancer tumors and provide a brief introduction to what is known about the nature of these tumors in general and pancreatic cancer tumors in particular using the logistic regression model.

4. The theoretical side

4.1. The logistic regression model

Regression Model Logistic is one of the non-linear regression models that can be converted to linear models. Logistic is increasing day by day because it is interested in analyzing binary response data (zero, one) [6].
The logistic regression function (probability of response) can be represented in the case of several independent variables as follows:

$$p_{\mathrm{i}} = \frac{e^{\beta_0+\sum_j^k \beta_j x_{ij}}}{1+e^{\beta_0+\sum_j^k \beta_j x_{ij}}} \qquad (1)$$

$$j = 1,2,\dots k \qquad i = 1,2,\dots,n$$

$$-\infty < X < \infty \quad -\infty < \beta_0 < \infty \quad 0 < \beta_1$$

Where p_i is response probability and specified $0 \le p(x) \le 1$. β_0, β_1: are the parameters of the model to be estimated. *X*: the independent variable. n: the number of observation and *K*: represents the number of independent variables [7].

4.2. The maximum likelihood method

The maximum likelihood method is one of the many methods used in statistics because of its desirable statistical characteristics and the fact that this method depends on finding the values of the parameters that make the function at its great end.
Using the Newton Raphson method, all estimates of the model parameters can be obtained. Newton Raphson formula for the logistic model parameters estimates can be developed as follows:

$$\underline{\hat{B}}_{s+1} = \underline{\hat{B}}_s - \left(X'VX\right)^{-1} X'\left(Y - \hat{Y}_s\right) \qquad (2)$$

Where $\underline{\hat{B}}_{s+1}$ is a vertical wave of the values of the estimates in the cycle (*s + 1*) of the rank *((K + 1) * 1)*. $\underline{\hat{B}}_s$: A vertical wave of the values of the estimates in Cycle (s) of the rank (*(K + 1) * 1*). *X*: matrix of independent variables of rank (*r * (K + 1)*) and *V*: diagonal matrix of rank (*r * r*).

$$V = \begin{bmatrix} n_1\hat{p}_1\hat{q}_1 & 0 & 0 \\ 0 & n_2\hat{p}_2\hat{q}_2 & 0 \\ 0 & 0 & n_k\hat{p}_k\hat{q}_k \end{bmatrix} \qquad (3)$$

The estimation of model parameters is obtained when the difference between the previous and subsequent cycles is very small and approaches zero [8].

5. Model quality tests

5.1. Chi square test

It is one of the important tests used to match the quality of the model [9].

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

Since: O_i: represents watch values and E_i: represents expected values.

5.2. The coefficient (R^2)

The statistical coefficient of determination (R^2) does not give an explanation for the quality and relevance of the model but is an indicator of the importance of independent variables in terms of prediction of the dependent variable and its formulas are as follows [9]:

$$R_n^2 = \frac{(l_m^{2/n}) - (l_0^{2/n})}{(l_m^{2/n}) - (l_m . l_0)^{2/n}} \quad (5)$$

6. Results and discussion

The research data included the information available in the statistical form of cancer tumors taken from the Ministry of Health / Cancer Center / Statistics Division. Pancreas Cancer (509) for 2017. Risk factors have been identified and coded according to the following information: Gender, Age, Smoking, Adders Code, Occupation, Disease Detection and The response variable represents the injured condition.

Table 1. Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	508	100.0
	Missing Cases	0	.0
	Total	508	100.0
Unselected Cases		0	.0
Total		508	100.0
a. If weight is in effect, see classification table for the total number of cases.			

Table 1 shows the size of the sample taken for patients with pancreatic cancer (508) without losing any of the values of the variables and the percentage was 100%.

Table 2. Gender

		Frequency	Percent
Valid	male	285	56.1
	female	223	43.9
	Total	508	100.0

Table 2 shows the number of people with gender by pancreatic cancer. Where the of proportion of male (56.1%) is slightly higher than the proportion of females (43.9%).

Table 3. Age

		Frequency	Percent
Valid	20_40	44	8.7
	40_60	174	34.3
	60_80	251	49.4
	80_100	39	7.7
	Total	508	100.0

Table 3 shows that the age groups affected by pancreatic cancer in a large proportion is the category of (60_80) which includes the number (251) by (49.4%) followed by the category of (40_60) which includes the number (174) by (34.3%). These two categories are larger than any other groups.

Table 4. Status

		Rank	Frequency	Percent
Valid	Dead	1	179	35.2
	Alive	2	329	64.8
	Total		508	100.0

Table 4 shows the coding of the approved variable with the sample size for each type.

Block 0: Beginning block

Table 5. Classification Table[a,b]

Observed			Predicted		
			Status		Percentage Correct
			Dead	Alive	
Step 0	Status	Dead	0	179	.0
		Alive	0	329	100.0
	Overall Percentage				64.8
a. Constant is included in the model.					
b. The cut value is .500					

Table 5 shows that the patient with pancreatic cancer, which does not have any information recorded about him, is placed alive (64.8%) It represents a fixed model without variables and is a significant model.

Table 6. Variables in the Equation

Step 0	*B*	*S.E.*	Wald	*df*	*Sig.*	*Exp(B)*
Constant	.609	.093	42.949	1	.000	1.838

Table 6 shows the estimation of the constant limit($\beta = 0.609$). Wald statistics (42.949) and the level of significance of the zero block (0.000) which confirms the efficiency and significance of the model, and the expected value is equal to (1.838).

Table 7. Variables not in the Equation

			Score	*df*	*Sig.*
Step 0	Variables	Age	24.992	3	.000
		Age(40_60)	3.303	1	.029
		Age(60_80)	3.321	1	.028
		Addrcode	96.050	18	.000
		Baghdad	18.139	1	.000
		Nineuha	12.511	1	.000
		Erbil	5.550	1	.018
		Diyala	4.422	1	.035
		Babil	16.050	1	.000
		Karballa	5.550	1	.018
		Kirkuk	6.756	1	.009
		The_Qar	10.153	1	.001

			Score	*df*	*Sig.*
		Al_Sulimanyia	11.832	1	.001
		Duhok	5.550	1	.018
		Basis	327.583	7	.000
		Death Certificate	311.233	1	.000
		Cytology Heamatological	25.531	1	.000
		Histology of a metastasis	10.072	1	.002
		Histology of a primary	75.145	1	.000
	Overall Statistics		375.781	38	.000

Table 7 shows the variables that affect the model and have a significant effect according to statistical significance. Variables that had no significant effect on the model were excluded.

Block 1: Method = enter

Table 8. Omnibus Tests of Model Coefficients

		Chi-square	*df*	*Sig.*
Step 1	Step	482.834	38	.000
	Block	482.834	38	.000
	Model	482.834	38	.000

Table 8 shows the statistical significance of accepting the entry of variables into the model, which shows the semantic value of Chi-square, which is considered significant.

Table 9. Model Summary

Step	*-2 Log likelihood*	*Cox & Snell R Square*	*Nagelkerke R Square*
1	176.445[a]	.613	.844

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Table 9 shows the suitability of the logistic model for the data under study. The value of ($R^2_{Cox\ \&\ Snell}$) which equals (0.613) indicates that (61.3%) of the changes taking place in the dependent variable are total differences that can be explained by variance and the rest (39.7%) are changes. Random value and ($R^2_{Nagelkerke}$) which is equal to (0.844) indicates that (84.4%) of the changes in the dependent variable are total differences that can be explained by variance and the rest (15.6%) are random changes. The model is compatible with data analysis.

Table 10. Classification table[a]

Observed			Predicted		
			Status		Percentage Correct
			Dead	Alive	
Step 1	Status	Dead	148	31	82.7
		Alive	14	315	95.7
	Overall Percentage				91.1
a. The cut value is .500					

Table 10 shows that the classification of data increased to (91.1%) from table (5), where the classification of admission for life increased by (31) number of deaths, and that the value is a high indication of acceptance.

Table 11. Variables in the equation

	B	S.E.	Wald	*df*	Sig.	Exp(B)
Gender(1)	.484	.682	.505	1	.047	1.623
Age(2)	-.301-	.902	.111	1	.039	.740
Age(3)	-.663-	1.141	.338	1	.041	.515
Smoke(1)	-.137-	.430	.102	1	.049	2.872
Baghdad (1)	-2.043-	.613	11.112	1	.001	.130
Nineuha (3)	-2.250-	.906	6.170	1	.013	.105
Diyala (7)	.359	1.381	.067	1	.045	1.431
Karballa (10)	-3.675-	.891	17.028	1	.000	.025
The_Qar (13)	.498	1.228	.165	1	.045	1.646
Duhok (15)	-1.702-	1.427	1.422	1	.033	.182
Al_Muthana (16)	-1.751-	.846	4.281	1	.039	.174
Al_Najaf (18)	-1.639-	1.334	1.509	1	.021	.194
AGRICULTURE	2.662	1.670	2.542	1	.011	14.332
OFFICE WORKER	2.694	2.521	1.142	1	.025	14.792
PROFESSIONAL	3.049	1.677	3.306	1	.049	21.101
OTHERS	2.426	1.705	2.023	1	.155	11.310
Constant	-39.44-	1.614	.103	1	.000	1.679

a. Variable(s) entered on step 1: Gender, Age, Smoke, Addrcode, Occup, Basis.

Table 11 shows the estimated parameters using the maximum likelihood method, the Wald statistic with the significance of the estimator and its probabilistic value.

7. Conclusion

That the gender has a slight effect on the incidence of pancreatic cancer, where the increase by one unit of the female is offset by an increase of (1.6) which is a slight increase. The amount of change in the ages of people was close to the two categories of (40-60) and from (60-80) and this is indicated in the number of people with those two groups. In terms of governorates. It was found that the injured in the southern governorates have an impact on the total number for a number of reasons, including registered and unregistered, as shown in table 10. The effect of the professions was clear from the employee, the profession and other related jobs. There is no effect on the variable of diagnosis of the disease in terms of infection. It is noticed in many statistical studies that variables with influence on disease are confused with variables that help in the healing or treatment process of the injury.

8. Recommendations

Approving other variables that were not included in the form of the Ministry of Health, including the variable of drinking alcohol, its quantity, genetics, number of injured in the family, type of injury and the degree of relationship to the injured.

References

[1] A. R. Al-Humrani, et al., " Carcinoma of the pancreas: A six year experience", *Basrah Journal of Surgery*, March, 9, pp. 16, 2003.

[2] A. M. Layla, A. M. Zena ,"Isolation of Trypsin from Serum of pancreatic Cancer Patients and determination some biochemical parameters", *Tikrit Journal of Pure Science* ,vol. 23, no. 6, pp. 31 ,2018.

[3] R.L. Siegel, K.D. Miller and A. Jemal, "cancer statistics". *CA Cancer J.Clin*. vol. 65, no. 1, pp. 5-29, 2015.

[4] F. A. Dalya and J. S. Entsar , "Prevalence of Toxoplasmosis Infection in Iraqi Women with Different Types of Cancer", *Diyala Journal of Medicine*, vol. 13, Issue 2, pp. 56_57, 2017.

[5] G. H.Hamid,"The Significance of a-L-Fucose as Biomarker in Cell Proliferation", *Medical Journal of Babylon*, vol. 2, no. 3, pp. 311, 2005.

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

[7] D. R. Cox and E. J. Snell, "*Analysis of binary data*", Chapman & Hall/CRC Monographs on Statistics and Applied Probability 32, 1989.

[8] 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.

[9] 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.