ZANCO Journal of Pure and Applied Sciences
The official scientific journal of Salahaddin University-Erbil
https://zancojournals.su.edu.krd/index.php/JPAS

ISSN (print):2218-0230, **ISSN (online):** 2412-3986, **DOI:** **http://dx.doi.org/10.21271/zjpas**

RESEARCH PAPER

Application of Binary Logistic Regression Model to Cancer Patients: a case study of data from Erbil in Kurdistan region of Iraq.

Khwazbeen Saida Fatah [1], Zhyan Rafaat Ali Alkaki [2]

[1] Department of Mathematics, College of Science, Salahaddin University-Erbil, Kurdistan Region, Iraq.
[2] Department of Mathematics, College of Basic Education, Salahaddin University-Erbil, Kurdistan Region, Iraq.

ABSTRACT:

In this paper, Binary Logistic Regression technique for fitting the best model for analyzing cancer diseases data is introduced using SPSS software based on forward stepwise procedures using different tests. The main objective is to investigate the last status of patients suffer from various types of cancer in Kurdistan Region of Iraq for the period 2010-2019 based on the main factors that contribute significantly to their last situation. A random sample of size 821 cancer patients from two main public hospitals (Rzgari and Nankali) in Erbil, where the patients take permanent and appropriate treatment, is selected of 619 alive and 202 dead. The results of the study showed that binary logistic regression is an appropriate technique to identify statistically significant predictor variables such as gender, age, cancer site and region to predict the probability of the last status (alive or dead) for each cancer patients. Moreover, it is deduced that despite the higher rate of cancer for female patients than the male; the chance for female cancer patients to be alive is more than male patients.

KEY WORDS: Binary Logistic Regression, Odds Ratio, Probability, and Hosmer and Lemeshow Test.
DOI: http://dx.doi.org/10.21271/ZJPAS.33.4.12
ZJPAS (2021), 33(4);117-128.

1. INTRODUCTION:

Logistic Regression Models analyze the relationship between a response (dependent) variable and a set of predictors (independent) variables. The main purpose for designing these models is to analyze the existence of the association between a nominal dependent variable and a group of independent (continuous or categorical) variables. With this model, instead of predicting the value of the response variable directly, the logistic regression equation predicts the odds of the event of interest occurring; this enables the approach to become a widely used statistical technique in medical research studies particularly over the last two decades in which the majority of the medical investigations are published (Yusuff, et al.,2012; Abdalrada, et al., 2019; Yuri, et al., 2016; Nankani, et al., 2019).

Binary Logistic Regression (BLR) , which studies the association between a binary response variable with only two categories and a set of predictors, is the most common used statistical model for the analysis of binary data in various applications such as physical, biomedical and health sciences, in which BLR model can be implemented for many applications in which data analysis comprise predicting the value of the outcome variable (Vupa and Çelikoğlu, 2006). This approach is particularly appropriate for models in which the dependent variable represents disease state (diseased or healthy) or patient status (dead or alive) for patients suffering from a certain disease (Sweet and Martin, 2011; Osborne, 2012; Mabula, 2015; Neupane, et al., 2002; Şirin, and Şahin, 2020; Bozpolat, 2016; Bahadır, 2016; Huang, and Moon, 2013). In situations when the response variable takes more than two categories it is then referred to as multinomial logistic regression (El-Habil, 2012).

* **Corresponding Author:**
Khwazbeen Saida Fatah
E-mail: khwazbeen.fatah@su.edu.krd
Article History:
Received: 15/03/2021
Accepted: 19/05/2021
Published: 18/08 /2021`

However, cancer is one of the diseases that causes major health problems with millions of human deaths in the world with numerous of new cancer cases arising each year and millions of cancer-related deaths with many factors that affect the increase or presence of this disease. The report by WHO (World Health Organization, 2021) on 5th March 2021, stated that the second leading cause of worldwide deaths is cancer; it is responsible for about 10 million deaths per year. Moreover, the report indicated that in low- and middle-income countries approximately 70% of deaths are due to cancer and that the most popular types of cancers are: Lung, Breast, Colorectal, Prostate, Skin cancer (non-melanoma); while Lung; Colorectal; Stomach; Liver and Breast cancer are the major causes of cancer .

Therefore, the aim of this study is to apply BLR model to analyze cancer patient's data for both genders male and female and then identify the predictor variables associated with the last status for all cancer patients in Kurdistan Region of Iraq and hence, based on these factors, determine the chance of being alive of both male and female patients.

2.Logistic Regression Model (LRM)

LRM analyzes the relationship between a response (categorical) variable, which is measured on a nominal scale, and a set of predictors (explanatory) variables. This model, which sometimes is called the logistic or logit model, estimates the probability of an event occurring by using a logistic curve. Logistic regression (LR) works very similar to linear regression except that in case of binary response variable the assumption of normality fails. The model is mathematically flexible; the interest in using this method has been increased due to not having any assumption limitations (Pregibon, 1981; Santner and Duffy, 1986; Abaye, 2019).

2.1Binary logistic regression model (BLRM)

The BLRM, which has begun to be a widely used model especially in medical research studies in which the dependent variable is dichotomous, for example: live/die; disease/no disease, can be considered as a statistical tool for predicting the association between the explanatory variables (predictors) and a binary (dichotomous) dependent or response variable with only two categories. With this model, the probability that an event falls into one of the two categories of the response variable is predicted by using one or more independent variable (continuous or categorical).

2.1.1 The Logistic Curve

In the BLRM, the variable representing the response categories takes values of 0 or 1 and the resulted value for the event of interest, the probability of the event occurring, should be in [0,1]. Hence, LRM uses the logistic curve to express the association between the predictors and the response variable. At a very low level of the predictor, the probability approximates to 0, but never becomes 0. While when the predictor variable increases the predicted value increase too; it approaches 1 but never equal to 1 (Nelder, 1961).

2.1.2 Transforming a Probability into Odds and Logit Values

Probability and odds both measures how likely it is that something will occur. The transformation of the logistic curve ensures that the predicted values fall inside the range of 0 and 1; this is obtained when the probability is measured as odds. Odds is defined as the ratio of the probability of the event occurring to the probability of the event does not occurring.

$$Odds\ (OR) = (probability\ of\ an\ event\ occuring)/(probability\ of\ event\ not\ occuring)$$

Then, to ensure that the odds values do not go below 0 which is the lower limit (there is no upper limit) the logit value is computed; it is obtained by taking the logarithm of the odds. Then, the odds value can be converted back into a probability; thus,

$$\text{Probability (event)} = \frac{\text{odds(event)}}{1+odds(event)} \qquad (1)$$

When there is an increase of one unit change in one of the independent variables, while the other predictors remain constant, there will be an increase in the odds of the response variable measured by a factor $exp(\beta_i$); this factor is called the Odds Ratio (OR) and its value is greater than zero. This is a measure of the relative quantity by which the odds of the response variable increase (OR > 1) or decrease (OR < 1) when the value of the corresponding predictor increases by one (1) unit. Therefore, OR is a measure of association which is easy to be calculated and interpreted; it is a ratio of two odds measured by the odds of an event occurring in one category to the odds of that event when occurring in the other category.

2.2 Model building

In BL function, the likelihood or chance of an outcome based on individual characteristics is modeled. Because chance is a ratio, the BLRM applies a transformation called a logit to the probabilities to ensure its boundaries. The logit is defined as the logarithm of the OR; it is given by

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) \qquad (2)$$

In equation (2), π is the probability of an event occurring; the mean of a binary variable is π, given the values of the predictor variables; Or the BLRM relates the log odds [ln(π/ 1 - π)] in a linear way to the variation in the predictor variables with $\pi = \Pr(Y = 1|X = x)$ and the information of $0 \leq \pi \leq 1$ can be transformed into information of $-\infty \leq logit(\pi) \leq \infty$.

Variables are explained as follows:
Variable Y is a binary response with two categories:
Y_i = 1 if the event occurs in observation i
Y_i = 0 if the event does not occur in observation i
and variable $X = (X_1, X_2, \dots, X_k)$ is a set of predictors that can be discrete, continuous, or a combination of both with x_i , the observed value of the predictors for observation *i*.

$$\pi_i = \Pr\left(Y_i = 1 \middle| X_i = x_i\right) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \qquad (3)$$

Hence, $\text{logit}(\pi_i) = \ln(\pi_i /(1 - \pi_i)) = \ln[\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})]$. (4)

Therefore, when the logits is defined as a linear relationship which takes the following form:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} \qquad (5)$$

The Main assumptions for constructing the model:

- The data $Y_1, Y_2, \dots, Y_n$ are identically independently distributed with binomial distribution $Bin(n_i, \pi_i)$, does not need to be normally distributed.
- There is no linear relationship between the outcome variable and the explanatory variables, but the linearity is between the logit of the response and the predictor variables; $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, so that changing an input variable multiplies the probability of the event of interest (output variable) by a fixed amount.
- The homogeneity of variance, which is not even possible in many cases, does not satisfy. The error terms are independent but normally do not satisfied. For parameter estimation the maximum likelihood estimation (MLE) is used instead of ordinary least squares (OLS), and thus relies on large-sample approximations. Goodness-of-fit measures rely on sufficiently large samples.

2.2.1 The Stepwise LRM

In model building, the stepwise LR is a widely used method specifically when the outcome variable being studied is relatively new (AIDS, some cancer types…) or the major predictors may not be identified and their relationship with the response variable may not be well defined. For these cases, stepwise procedures are used as a useful and effective tool to study a large number of such variables and fit a number of logistic regression equations. The technique for these procedures is to consider an initial model and then implement certain rules for selecting variables to arrive at a final model (Cristensen, 1997). There are two techniques in stepwise LRM: the first one is the forward selection, which sequentially adds different predictor variables to the model and then at each stage, the variables that give largest improvement in the fit are chosen until there is no improvement in the fit. The main criterion for this process is the maximum p-value for the final model. The second approach is the backward elimination process which starts by a complicated model and then sequentially eliminates variables; variables with minimum effect on the model are removed at each stage.

2.2.2 Likelihood Ratio Test (LRT)

The LRT, which is sometimes called the chi-square test, is a practical test applied to check the variation between the likelihood ratios L_1, L_2 for two models, the null model, which is with only a constant, and the one with predictor variables respectively. This test checks the contribution of each effect to the model based on the value $-2\log(L_1/\,L_2)$, which is the ratio between the log-likelihood of the null model (L_1) to the log-likelihood (L_2) for the model with more predictors. The model is significance at the level of $\alpha = 0.05$ or less means that the model including the predictors is significantly different from that one with the constant only or null model (all 'β' coefficients are zero). Otherwise, when the probability unable to reach the value 0.05 means that predictors has no influences on response variable then they are to be rejected. (Burns and Burns, 2008).

2.3 Measures of Goodness of Fit Test

Goodness of fit is a test used to see if the sample data fits a distribution or shows how effectively the model describes the variables. In linear regression method, the coefficient of determination R^2 is employed to measure the goodness of fit or the variation ratio explained by the model but with LR there is no such measure to test the goodness of fit, and therefore other statistics similar to R^2 such as Cox and Snell, Nagelkerke R Square and Hosmer and Lemeshow (Hosmer and Lemeshow, 1989) test are applied; the value of each test statistic is used to test the hypothesis of the model. The following hypotheses are considered:

H_0: The modified model fits the data well.

H_1: The modified model does not fit the data well.

When the test is applied, at a standard level of significance α= 0.05, the determined p-value is compared with α and H_0 will be rejected if its value is less or equal to α otherwise H_0 cannot be rejected.

2.4 Statistical Significant Test

Statistical significant test is applied to ensure the real contribution of the predictors in the model building and the prediction of the outcome. Wald test is one of these tests used to verify specifically whether the coefficient of predictor variables is significantly different from zero. Wald test, which is similar to the t-test performed on the coefficients of regression in a linear regression, is used to evaluate the fit of a logistic regression model; the larger value for this measure is an indication for the significance (Kleinbaum, and Klein, 2010).

2.5 Odds Ratios (OR) with 95% Confidence Interval (CI)

To estimate the accuracy of the OR, which is a measure of association, the 95% confidence interval (CI) is used. It is unlike the p-value, the 95% confidence does not refer to the statistical significance for the measure, but it is often used as a representative for the presence of the significance. A large CI indicates a low level of accuracy of the OR, while a small CI indicates a higher accuracy for the OR (Morris, and Gardner, 1988).

3.Methodology

3.1 Data collection and representation

The aim of the study is to analyze cancer patients data in Kurdistan Region of Iraq (with population approximately 6.5 million) where, over the years, health authorities have recorded annually thousands of patients who suffer from different types of cancer. The focus is specifically on Erbil city, the capital of Kurdistan Region, where cancer patients take permanent and appropriate treatment in two main public hospitals (Rzgari and Nankali). The patients involved in this study are those who suffer from different types of cancer and get treatment from Erbil hospitals; they are from two different parts of Iraq, Erbil city and others, who come from different places in Kurdistan Region and all over Iraq. The analysis for cancer patients treated at two the public hospitals (Rzgari) and (Nankali) in Erbil city covered the period 2010-2019. A sample of 821 cancer patients data was provided by the General Directorate for Health in Erbil; they confirmed that it represents 10% of total number of patients for that period; this sample consists of 619 survived patients of which 357 females and 262 males and 202 died including 87 females and 115 males. The main variables for this study are described by Table 1:

The predictor variables are: the time period, Year variable, for the investigation, from 2010-2019 which is divided into three classes; the first class is from 2010 - 2012 which covers 11.5% of all cases, and the second class is 2013-2015 which represents 37.2%, finally the last class is 2016-2019 with 51.3% of all cases; it is the major class of the study.

The second predictor is the patient's age; it is divided into five groups: 0-20 which represents 2.5%; 21-40 represents 15.5%; 41-60 as 39.8%; 61-80 with 38%; finally 81 and above which covers only 4.1% . Then the Gender is another predictor which is either male or female.

Another important predictor is Cancer Site which is divided into four main groups; they are G1 group the breast cancer with 21.5%; the G2 group (Stomach, brain, lung, pancreas and liver) which represents 31.2%, and the G3 group (Prostate, colon, bladder and ovary) covers 17.5% and finally the G4 group (bone, skin, thyroid gland, hypo pharynx, other parts of tongue, ear, small intestine, parts of mouth, palate, tonsil, ureter) which covers 29.9% of all cases.

Finally, the region of residence is classified in to Erbil state with 73.4% and the others, who are from other parts of Iraq but treated in Erbil, 26.6% of all cases.

While the response variable in this study is the patient's last status (dead, alive) represented by values (0, 1).

In this study, the Statistical Package for Social Sciences (SPSS) software has been used for data analysis for the main variables and the results are shown by the following tables:

Table 2 shows that theme an alive percentage is equal to 75.4 since no independent variable has yet been entered.

Table 3 shows the intercept-only model where no variable contributes the prediction. It is shown that $\mathrm{B} = 1.120$, if the exponential function is taken for both sides of this expression then the predicted odds is $\mathrm{EXP(B)} = 3.064$, which explains that the predicted odds for being alive equals 3.064. Since 619 of all patients are alive while 202 are dead then the observed odds are $619/202 = 3.064$.

In classification table, Table 4, for the model with predictors, it is shown that the predicted number of dead patients is 27 out of 619, and from 202 only 139 are alive, and the percentage can be calculated as follows: $\frac{63}{202} \times 100 = 31.2\%$ dead, $\frac{592}{619} \times 100 = 95.6\%$ alive.

And the overall percentage equals: $\frac{619}{821} \times 95.6 + \frac{202}{821} \times 31.2 = 79.8$

It also shows that introducing independent variables to the model will improve the model in the ratio from 75.4% to 79.8%.

In Table 5, it is shown that all predictor variables are significant depending on P-value; this means that introducing each of the independent variables is significant to improve the model.

It is noted that in fitting a linear regression model, the amount of variance in the output variable associated with the independent variables is determined by the coefficient of determination, R^2. Large value for R^2 indicates that most of the variation is explained by the model, to a maximum of 1. While for LRMs, in which the dependent variable is categorical, it is not possible to determine a value for R^2; instead there are various ways to determine an R^2 for LR and no consensus on which one is the best; the following are three measures of fit:

- Cox and Snell's R^2 : for this measure, which is proposed by Cox and Snell, the log likelihood for the model is compared with the log likelihood for a baseline or the null model, the model with no predictor; it has a theoretical maximum value of less than 1, even for a "perfect" model.

- Nagelkerke's R^2 is an adjusted version of the Cox & Snell R-square that modifies the scale of the measure in order to cover the full range from 0 to 1.
- Hosmer and Lemeshow Chi-square test of goodness of fit for logistic regression tells how well the data fits the model. This test determines if the observed and expected data are matched.

These tests, which show how R^2 are calculated, are explained by table 6.

Table 6, for each model $(-2LL)$ is determined, it is an estimate of a model's suitability for the data which is often used to see if adding additional variables to the model results in a significant reduction.

The results indicate that the hypothesis H_0 is rejected if the p-value (sig.) for the overall model is less than 0.05; this concludes that there is evidence that at least one of the independent variable contributes to the prediction of the outcome. In addition, the results explains that adding any variable to the model, the value of $(-2LL)$ will be decreased.

Furthermore, Cox and Snell's R^2 is determined. This value indicates that the probability ratio of the variance in the response variable is explained by the predictor variable which is assumed to be good enough. However, Nagelkerke's R^2 is an adjusted version of the Cox & Snell R^2 statistic.

In Table 7, the Forward Stepwise (Wald) method, which is a method for finding the best model among others, is applied; it selects the largest significant value. In Step 4, the model is with highest significance; hence it will be the preferred one among the others.

The table 8 shows how the BLR, including all variables, is implemented and then different tests are applied.

Table 8 displays data analysis for different models with their accuracy within some intervals; it shows that all predictors are significant and that the fourth model is the best model based on the value of the tests. Therefore, the model identified as the fourth model is supposed to be the best one with exclusion of gender variable; the logit function it is defined as:

$$\log\left[\frac{\pi}{1-\pi}\right] = -1.747 - 0.529\text{ age} + 1.034\text{ year} + 1.421\text{ region} + 0.225\text{ cancer site}$$

However, the gender variable has been considered as a major predictor for identifying the last status, the gender variable is taken into consideration and the last model, with inclusion of gender is proposed. It is noted that data analysis shows that females have more chance to get cancer than males, but males have chance to die more than females. The following table is with Gender included.

In the table 9, the predictor variable Gender has been added to the overall model and the logit function it is defined as:

$$\mathbf{\log\left[\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right] = -3.124 - 0.464\,\boldsymbol{age} + 0.651\,\boldsymbol{gender} + 1.01\,\boldsymbol{year} + 1.447\,\boldsymbol{region} + 0.305\,\boldsymbol{cancer\ site}}$$

In this model, for example, if a male patient is selected who is from the second age group with breast cancer (G1) and from first category period in Erbil Region then the chance of being alive will be given by:

OR = $e^{-3.124-0.464\,(2)+0.651(\,1)+1.01\,(1)+1.447(1)+0.305(1)} = 0.52782$ with probability calculated from $\frac{\text{OR}}{1+\text{OR}} = \frac{0.52782}{1+0.52782} = 0.35.$

But if another a female patient from the same groups is selected then the chance of being alive is OR= $e^{-3.124-0.464\,(2)+0.651(\,2)+1.01\,(1)+1.447(1)+0.305(1)} = 1.012072$ with probability $= \frac{\text{OR}}{1+\text{OR}} = \frac{1.012072}{1+1.012072} = 0.5.$

While, if a male patient is selected from the second age group and from first category period in Erbil Region but with cancer of G2 group, then the chance of being alive is

OR=$e^{-3.124-0.464\,(2)+0.651(\,1)+1.01\,(1)+1.447(1)+0.305(2)} = 0.716054$ with probability= $\frac{\text{OR}}{1+\text{OR}} = \frac{0.716054}{1+0.716054} =$ 0.42.

But if the patient is a female and from the same groups then the chance of being alive is OR = $e^{-3.124-0.464\,(2)+0.651(\,2)+1.01\,(1)+1.447(1)+0.305(2)} = 1.373003$ with probability =$\frac{\text{OR}}{1+\text{OR}} = \frac{1.373003}{1+1.373003} = 0.58.$

Table 1: categories and levels of variables included in the study

Category	Level
Years	2010-2012, 2013-2015, 2016-2019
Age	0-20, 21-40, 41-60, 61-80, 81-
Gender	Male, Female
Cancer Site	G1, G2, G3, and G4
Region	Erbil, others
Patients last status	Alive, dead (binary variable)

Table 2: Classification Table

			Predicted		
			Situation		Percentage Correct
	Observed		Dead	Alive	
Step 0	Situation	Dead	0	202	.0
		Alive	0	619	100.0
	Overall Percentage				75.4

Table 3: Variables in the Equation

		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 0	Constant	1.120	.081	190.989	1	.000	3.064

Table 4: Classification Table

			Predicted		
			Situation		Percentage Correct
	Observed		dead	Alive	
Step 1	Situation	Dead	63	139	31.2
		Alive	27	592	95.6
	Overall Percentage				79.8

Table 5: Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Age	25.119	1	.000
		Gender	13.081	1	.000
		Year	75.597	1	.000
		Region	34.125	1	.000
		Cancer site	9.496	1	.002
	Overall Statistics		147.558	5	.000

Table 6: Forward Stepwise (Wald) method

Model Summary				Hosmer and Lemeshow Test		
Step	-2 Log likelihood (-2LL)	Cox & Snell R Square	Nagelkerke R^2	Chi-square	df	Sig.
1	843.039[a]	.085	.127	27.875	1	.000
2	802.552[b]	.129	.192	32.326	4	.000
3	778.462[b]	.154	.230	22.408	8	.004
4	771.092[b]	.162	.241	11.080	8	.197
5	759.260[b]	.174	.259	17.596	8	.024

Table 7: Forward Stepwise (Wald) method, Likelihood ratio test

Omnibus Tests of Model Coefficients

		Chi-square	Df	Sig.
Step 1	Step	73.105	1	.000
	Block	73.105	1	.000
	Model	73.105	1	.000
Step 2	Step	40.487	1	.000
	Block	113.592	2	.000
	Model	113.592	2	.000
Step 3	Step	24.091	1	.000
	Block	137.683	3	.000
	Model	137.683	3	.000
Step 4	Step	7.370	1	.007
	Block	145.053	4	.000
	Model	145.053	4	.000
Step 5	Step	11.832	1	.001
	Block	156.884	5	.000

Model	**156.884**	**5**	**.000**

Table 8: Binary Logistic Regression Models using (Forward Stepwise (Wald) method)
Variables in the Equation

								95% C.I.for EXP(B)	
		B	**S.E.**	**Wald**	**Df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Step 1[a]	**Year**	**1.003**	**.121**	**68.587**	**1**	**.000**	**2.726**	**2.150**	**3.456**
	Constant	**-1.176**	**.279**	**17.763**	**1**	**.000**	**.309**		
Step 2[b]	**Year**	**1.036**	**.125**	**69.225**	**1**	**.000**	**2.818**	**2.208**	**3.597**
	Region	**1.423**	**.250**	**32.322**	**1**	**.000**	**4.150**	**2.541**	**6.779**
	Constant	**-2.964**	**.425**	**48.742**	**1**	**.000**	**.052**		
Step 3[c]	**Age**	**-.527**	**.111**	**22.439**	**1**	**.000**	**.590**	**.475**	**.734**
	Year	**1.068**	**.127**	**70.190**	**1**	**.000**	**2.909**	**2.266**	**3.735**
	Region	**1.376**	**.254**	**29.405**	**1**	**.000**	**3.960**	**2.408**	**6.513**
	Constant	**-1.218**	**.557**	**4.784**	**1**	**.029**	**.296**		
Step 4[d]	**Age**	**-.529**	**.113**	**21.998**	**1**	**.000**	**.589**	**.472**	**.735**
	Year	**1.034**	**.128**	**65.116**	**1**	**.000**	**2.811**	**2.187**	**3.613**
	Region	**1.421**	**.255**	**31.006**	**1**	**.000**	**4.140**	**2.511**	**6.827**
	Cancer site	**.225**	**.084**	**7.270**	**1**	**.007**	**1.253**	**1.063**	**1.475**
	Constant	**-1.747**	**.595**	**8.615**	**1**	**.003**	**.174**		
Step 5[e]	**Age**	**-.464**	**.114**	**16.469**	**1**	**.000**	**.629**	**.502**	**.787**
	Gender	**.651**	**.191**	**11.626**	**1**	**.001**	**1.918**	**1.319**	**2.789**
	Year	**1.010**	**.129**	**61.111**	**1**	**.000**	**2.745**	**2.131**	**3.535**
	Region	**1.447**	**.256**	**31.942**	**1**	**.000**	**4.252**	**2.574**	**7.023**
	Cancer site	**.305**	**.090**	**11.625**	**1**	**.001**	**1.357**	**1.139**	**1.617**
	Constant	**-3.124**	**.731**	**18.251**	**1**	**.000**	**.044**		

Table 9: Variables in the Equation

								95% C.I.for EXP(B)	
		B	**S.E.**	**Wald**	**Df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
Step 5[e]	**Age**	**-.464**	**.114**	**16.469**	**1**	**.000**	**.629**	**.502**	**.787**
	Gender	**.651**	**.191**	**11.626**	**1**	**.001**	**1.918**	**1.319**	**2.789**
	Year	**1.010**	**.129**	**61.111**	**1**	**.000**	**2.745**	**2.131**	**3.535**
	Region	**1.447**	**.256**	**31.942**	**1**	**.000**	**4.252**	**2.574**	**7.023**
	Cancer site	**.305**	**.090**	**11.625**	**1**	**.001**	**1.357**	**1.139**	**1.617**
	Constant	**-3.124**	**.731**	**18.251**	**1**	**.000**	**.044**		

3.2 Final Model

When the model is fitted, the following table explains how the data is analyzed and then how the most significant predictors are selected.

Table 10 explains that with the best model, which is the fifth one, all variables are significant except the first and third age categories with third cancer site category (significance is greater than 0.05).

In Table 11, the Wald Chi-Square test statistic is calculated; it shows the unique contribution of each predictor, holding the other predictors constant, wherefore each variable group the comparison is with final category. For example, for cancer site, where each group is compared with $G4$, predictors meets the conventional .05 standard for statistical significance, except for variable $G1(1)$. It is explained that for a patient whose in the second cancer site category, his chance to be alive decreases by the OR which is $EXP(\beta) = 0.159$ with probability 0.14 while if that patient is in third category then the chance will be reduced by $EXP(\beta) = 0.348$ with probability 0.26.

Table 10: Variables not in the Equation

			Score	Df	Sig.
Step 0	**Variables**	**Age**	**26.698**	**4**	**.000**
		0-20(1)	**1.237**	**1**	**.266**
		21-40(2)	**13.255**	**1**	**.000**
		41-60(3)	**1.523**	**1**	**.217**
		61-80(4)	**10.311**	**1**	**.001**
		Gender male(1)	**13.081**	**1**	**.000**
		Years	**120.043**	**2**	**.000**
		2010-2012(1)	**119.035**	**1**	**.000**
		2013-2015(2)	**4.241**	**1**	**.039**
		Region Erbil(1)	**34.125**	**1**	**.000**
		Cancer site	**112.332**	**3**	**.000**
		G1(1)	**23.027**	**1**	**.000**
		G2(2)	**92.605**	**1**	**.000**
		G3(3)	**1.544**	**1**	**.214**
	Overall Statistics		**225.037**	**11**	**.000**

Table 11: Binary Logistic Regression model fitted to Classes and Levels of the Variables Included (final model)

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Age			10.362	4	.035			
	0-20(1)	.843	.856	.969	1	.325	2.322	.434	12.427
	21-40(2)	1.504	.511	8.672	1	.003	4.501	1.654	12.249
	41-60(3)	.748	.431	3.006	1	.083	2.113	.907	4.922
	61-80(4)	.564	.425	1.762	1	.184	1.758	.764	4.042
	Gender male(1)	.049	.214	.052	1	.820	1.050	.690	1.598
	Years			65.495	2	.000			
	2010-2012(1)	-2.376	.296	64.372	1	.000	.093	.052	.166
	2013-2015(2)	-.350	.211	2.749	1	.097	.705	.466	1.066
	Region Erbil(1)	-1.431	.275	27.157	1	.000	.239	.140	.410
	Cancer site			59.401	3	.000			
	G1(1)	-.023	.368	.004	1	.951	.977	.475	2.009
	G2(2)	-1.840	.274	45.210	1	.000	.159	.093	.271
	G3(3)	-1.055	.306	11.845	1	.001	.348	.191	.635
	Constant	2.947	.539	29.851	1	.000	19.057		

a. Variable(s) entered on step 1: age, gender, year, region, cancer site.

4.Conclusion

BLR provides an effective tool for modeling the dependence of a binary output variable on one or more predictors that can be either categorical or continuous variables. In this study, a BLRM is fitted to data obtained from a case study concerns an investigation on the last status for patients suffer from different types of cancer treated in Erbil in Kurdistan Region of Iraq for the period of 2010-2019.

According to the results obtained from the analysis of data, using SPSS software, by implementing various procedures based on forward stepwise and applying different tests, the best model for all cancer patients, male and female patients with the response variable as the patient's last status, can be obtained. The model can easily identify the factors or predictors that will affect the response variable, which is the patient's last status (alive or dead). It is concluded that the main predictors contribute the chance of the cancer patients to be alive or dead are each of variables, age, gender, cancer site and the region. When predictor variables are included in the model, the results show improvement for the model by the ratio from 75.4% to 79.8%. In addition, when Forward Stepwise (Wald) methods are applied for each model, it is shown that the value for $-2LL$ is decreased. All other tests showed that the model is improved by adding a new influence predictor. Moreover, it is concluded that despite the rate of female patients who suffer from cancer is higher than the male patients; the rate of death for male patients is higher than for female. Finally, by applying the best model, if one patient (male or female) is randomly sampled and all information for the factors in the model is identified then the status of that patient can easily be predicted through the estimated logit value.

References

Abaye, D. 2019. A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of data analysis and information processing*, 7(4), 190-207.

Abdalrada, A., Yahya, O., Alaidi, H., Hussein, N., Alrikabi, H., and Al-Quraishi, T. 2019. A Predictive Model For Liver Disease Progression Based On Logistic Regression Algorithm. *Periodicals of Engineering and Natural Sciences*, 7(3), 1255-1264.

Abadi, A., hajizadeh, E., Pourhoseingholi, M. and Mojarad, E. 2019. Comparison of Random Forest and Logistic Regression Methods in Predicting Mortality in Colorectal Cancer Patients and Its Related Factors. *Iranian Journal of Epidemiology*, 14(4), 375-383.

Bahadır, E. 2016. Using Neural Network and Logistic Regression Analysis to Predict Prospective Mathematics Teachers's Academic Success upon Entering Graduate Education. *Educational Sciences: Theory & Practice,* 16(3), 943-964.

Bozpolat, E. 2016. Investigation of the Self-Regulated Learning Strategies of Students from the Faculty of Education Using Ordinal Logistic Regression Analysis. *Educational Sciences: Theory & Practice,* 16(1), 301-318.

Burns, R. and Burns, R. 2008. *Business Research Methods and Statistics Using SPSS*.

Cristensen, R. 1997. *Log-Linear Models and Logistic Regression*. 2nd edition, Springer-Verlag, New York.

El-Habil, A. M. 2012. An Application on Multinomial Logistic Regression Model. *Journal of Statistics and Operation Research*, 8(2), 271-291.

Hosmer, W. and Lemeshow, S. 1989. *Applied Logistic Regression.* Second Edition, Canada.

Huang, L. and Moon, R. 2013. What Are the Odds of That? A Primer on Understanding Logistic Regression. *Gifted Child Quarterly*, 57(3), 197–204.

Kleinbaum, D. and Klein, M. 2010. *Introduction to Logistic Regression*. Springer Science+Business Media. LLC

Mabula, S. 2015. Modeling Student Performance in Mathematics Using Binary Logistic Regression at Selected Secondary Schools A Case Study of Mtwara Municipality and Ilemela District. *Journal of Education and Practice,* 6(36), 96-103.

Morris, A. and Gardner, J. 1988. Calculating confidence intervals for relative risks (odds ratios) and standardized ratios and rates. *British Medical Journal (Clinical Research Ed.),* 296(6632), 1313-1316

Nankani, H,; Gupta, S., Singh, S., Ramesh, S. S. S. 2019. Detection Analysis of Various Types of Cancer by Logistic Regression using Machine Learning. *International Journal of Engineering and Advanced Technology (IJEAT),* 9(1), 99-104.

Nelder, J. A. 1961. The Fitting of A Generalization of the Logistic Curve. *International Biometric Society*, 17(1), 89-110.

Neupane, R. P., Sharma, K. R., and Thapa, G. B. 2002. Adoption of agroforestry in the hills of Nepal: a logistic regression analysis. *Agricultural Systems,* 72(3), 177–196.

Osborne, J. W. 2012. 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, and Evaluation,* 17(11), 1-10.

Pregibon, D. 1981. Logistic Regression Diagnostics. Annals of Statistics. *The Annals of Statistics,* 9(4), 705-724.

Santner, T. J., and Duffy, E. D. 1986. A Note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Oxford University Press on behalf of Biometrika Trust,* 73(3), 755-758.

Şirin, Y. E., and Şahin, M. 2020. Investigation of Factors Affecting the Achievement of University Students with Logistic Regression Analysis: School of Physical Education and Sport Example. *SAGE Open*, 10, 1-9.

Sweet, S. A., and Martin, K. G. 2011. *Data Analysis with SPSS + Mysearchlab with Etext: A First Course in Applied Statistics*. Pearson College Division.

Vupa, Ö. and Çelikoğlu, C. 2006. Model Building in Logistic Regression Models About LUNG CANCER DATA. *Anadolu University Journal of Science and Technology,* 7(1), 127-141.

World Health Organization. Available from https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 5th, March, 2021]

Yuri, P., Rochadi, S., and Danarto, R. 2016. A Device for Predicting Prostate Cancer Risk: A Logistic Regression. *Journal of Prostate Cancer,* 1(2), 1-5.

Yusuff, H., Mohamad, N., Ngah, U.K., and Yahaya, A.S. 2012. Breast Cancer Analysis Using Logistic Regression. *IJRRAS*, 10(1), 14-22.