Application of Logistic Regression for Incidence of Breast Cancer among Sudanese Females (2005-2008)

Eman Obeid Mohammed Alhadi[1]
Adel Ali Ahmed[2]

Abstract

This paper aimed at application of logistic regression model to the data of Sudanese females living with breast cancer with reference to Gezira State, the logistic regression method has been used to study breast cancer situation through a sample of 110 patients. The study basically depended on secondary data collected from the National Cancer Institute (NCI) , the data covered the period (2005–2008). Research methodology included chi-square test to know characteristics of the variables of the study population. The results that the tribe, the stage of illness and marital status are the most statistically significant variables. Multiple logistic regression models are applied in two parts, first part contains all the variables in this research, and the second contains significant variables. The study showed that single women who have breast cancer are more likely to die of breast cancer than women who are married, and women in the late stage are more likely to die of breast cancer than those who are in early stage. The research recommended that, there must be assessment for the current situation of the breast cancer so as to construct strategic plan to stop or eradicate the spread among people. Also case-control data used in this research, together with the cancer registry data, can be used to construct such models of absolute risk.

[1] Lecturer, Dept. of Developmental Planning, Faculty of Developmental Studies, University of Gezira, Email: mam26ecno@ yahoo.com.
[2] Assistant Professor, Dept. of Applied Statistics and Demography, Faculty of Economics and Rural Development, University of Gezira, Email: adelaliahmed@yahoo.com

تطبيق نموذج الانحدار اللوجستي على المصابات بسرطان الثدي بين النساء السودانيات (2005-2008)

ملخص الدراسة

هدفت هذه الورقة إلى تطبيق نموذج الانحدار اللوجستي في بيانات النساء السودانيات المصابات بسرطان الثدي بالإشارة إلى ولاية الجزيرة ، وتم استخدام طريقة الانحدار اللوجستي من خلال عينة ضمت 110 مريضة ، اعتمدت الدراسة بشكل أساسي على بيانات ثانوية جمعت من المعهد القومي للسرطان ، غطت البيانات الفترة (2005- 2008م). تضمنت منهجية البحث تطبيق اختبار كاي (chi-square) لمعرفة خصائص المجتمع محل الدراسة، وقد وجد أن كلاً من القبيلة وطور المرض والحالة الاجتماعية هي أكثر المتغيرات ذات الدلالة الإحصائية. تم تطبيق نموذج الانحدار اللوجستي المتعدد على مرحلتين؛ الأولى تحتوي على كل المتغيرات محل الدراسة؛ والثانية تحتوي على المتغيرات ذات الدلالة الإحصائية. أوضحت الدراسة أن النساء غير المتزوجات المصابات بسرطان الثدي هن أكثر عرضة للوفاة من النساء المتزوجات، وأن النساء في الطور المتأخر من المرض هن أكثر عرضة للوفاة من النساء في الطور المبكر من المرض. أوصت الدراسة بأنه لابد وأن يكون هنالك تقييماً للوضع الراهن لمرض سرطان الثدي وذلك لكي تبنى خطة إستراتيجية لوقف أو استئصال أو تقليل انتشار المرض بين الناس، كما أن بيانات مراقبة الحالة (case-control) التي استخدمت في هذه الدراسة إلى جانب بيانات مكتب تسجيل السرطان، يمكن أن تستخدم لبناء مثل هذه النماذج عن الخطر المطلق.

INTRODUCTION

Regression is one of the most important statistical methods and regression analysis is the use of statistical methodology for predicting values of one or
more response (dependent) variable from a collection of predictor (independent) variable values. A model is a statistical description of a physical, chemical or biological state or process. Using a model can help us think about such processes and their mechanisms, so we can design better experiments and comprehend the result. A model forces us to think through (and state explicit) the assumptions behind analysis. Logistic regression is a part of a category of statistical models called generalized linear models. This broad class of models includes ordinary regression and analysis of variance as well as covariance and log linear regression. Logistic regression allows one to predict outcomes, from a set of variables that may be continuous, discrete and dichotomous or a mix of any of these. Generally, the dependent or response variable is dichotomous such as presence/absence or success/failure. Discriminate analysis can only be used with continuous independent variable. Thus, in instances where the independent variables are categorical, or a mix of continuous and categorical logistic regression is preferred, (David W. Hosmer, 2002). The use of logistic regression modeling has exploded during the past decade. From its original acceptance in epidemiologic research, the method is now commonly employed in many fields but not nearly limited to biomedical research, business and finance, criminology, ecology engineering, health policy, linguistics and wildlife biology. The study problem explain that women in under developing countries exposure to the risk of death by breast cancer without any control or plan to stop the virus from dissemination among women. Most of the data was collected from the medical field in these countries containing some of the problems that leads to be data was defected, on the other side; most mathematical models involve unknown parameters that must be estimated from observed data. Then, the study will attempt to find specific statistical models related to breast cancer.

The objectives of the paper:

1. The paper applies the logistic regression so as to discover the relation between variable denotes breast cancer.
2. The paper also aims to reveal the concept of breast cancer cases through statistical model.
3. Furthermore, the paper aims to test the validity of the models.
4. Determining the factor that affects breast cancer in Sudan particularly Gezira State.

The importance of the paper stems from the fact that finding out an appropriate model for breast cancer in Sudan will help policy makers to conduct a comprehensive strategy to combat this disease thus help improve community health. It is very clear that statistical models and techniques become they become useful in many types of life, because they used data to gain insight in to real problems.

2. Logistic Regression Model:

Regression method have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variable, the goal of an analysis using Logistic
Regression is the same as the usual linear regression models where the dependent variable is assumed to be continuous or discrete.

2.1. General Logistic Regression Model:

The simple linear logistic model $\log\left(\frac{\Pi_i}{1-\Pi_i}\right)=\beta_1+\beta_2 x_i$ used
in the previous example is a special case of the general logistic regression model .

$$\log it\ \Pi_i = \log\left(\frac{\Pi_i}{1-\Pi_i}\right) = x_i^T \beta \quad \text{..........}(1)$$

Where x_i is a vector continuous measurement corresponding to covariates and dummy variables corresponding to factor level s and β is the parameter vector. This model is very widely used for analyzing data involving binary or binomial responses and several explanatory variables it provides a powerful technique analogous to multiple regression and analysis of variance (ANOVA) for continuous responses.

2.2. Multiple Logistic Regression Model:

The legit of the multiple logistic regression models is given by the equation

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_P x_P \quad \text{..........}(2)$$

In which case the logistic regression model is

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad \text{..........}(3)$$

If some of the independent variables are discrete, nominal scale variables such as sex, ethnic group and so, it is in appropriate to include them in the model as if they were interval scale variables. The number used to represent the various levels of these nominal scale variables are mealy identifiers and

have no numeric significance. In this situation the method of choice is to use a collection of design variables (or dummy variables).

2.3. Testing For the Significance of the Coefficients:

After estimating the coefficients, a first look at the fitted model commonly concerns an assessment of the significance of the variable in the model. This usual involves formulation and testing of a statistical hypothesis to determine whether the independent variable in the model is significantly related to the dependent variable.

The guiding principle with logistic regression in to compare observed values of the response variable to predicted value obtained from the model with and without independent variable. In logistic regression comparison of observed to predict values are based on the log likelihood function define the equation:

$$L(B) = Ln[(B)] = \sum_{i=1}^{n}[Y_i - Ln P_i + (1-Y_i)Ln(1-P_i)] \quad \text{..........} (4)$$

To find the value of β that maximizes L (B_0) we differentiate L (B_1) with respect to β_0 and β_1 set the resulting expressions equal to zero. This equation is as follows:

$$\sum_{i=1}^{n}[Y_i - P_i] = 0 \quad \text{..........} (5)$$

$$\sum_{i=1}^{n} x_i [Y - P_i] = 0 \quad \text{..........}(6)$$

And this called the likelihood equation. And better the likelihood comparison, it is helpful conceptually to think of an observed value of the response variable as also being predicted values resulting from a saturated model.

Saturated is one that contains as many parameters as there are data points. The comparison of the observed to predict values using the likelihood function are based on the following expression:

$$G = -2Ln\left(\frac{\text{Likelihood of the fitted model}}{\text{Likelihood of saturated model}}\right) \text{.....(7)}$$

For purpose of assessing the significance of independent variables we compare the value of D with and without the independent variables in the equation. The change in D due to the inclusion of the independent variables
in the model are obtained as:
G = D (model without the variable) – D (model with the variable) ……. (8)
This statistic plays the same role in logistic regression as the numerator of the partial F test dose in linear regression. Because the likelihood of the saturated model is common to both values of being differenced to compute G, it can be expressed as:

$$G = -2Ln\left(\frac{\text{Likelihood without the variable}}{\text{Likelihood with the variable}}\right) \text{.....(9)}$$

His calculation of the likelihood and likelihood ratio test are standard features of any good logistic regression software. This makes it easy to check for the significance of the addition of new terms to the model.
In summary the method for testing the significance of the coefficient of a variable is logistic regression it uses the likelihood function for a dichotomous outcome variable, (Christopher B. S. ,2006).

2.4. Confidants Interval Estimation:

An important adjunct to testing for significance of the model is calculation and interpretation of the confidence intervals for parameters of interest.
The basis for construction of the interval estimator is the same statically theory we used to formulate the test for significance of the model. In particular the confidence interval estimators for the slop and intercept are based on their respective Wald test. The endpoints of a 100(1-α) %confidence interval for the slope coefficient is:

$$\hat{\beta}_1 \pm Z_{1-\alpha/2}\, SE(\beta_1) \text{..... (10)}$$

And for the intercept they are:

$$\hat{\beta}_1 \pm Z_{1-\alpha/2}\, SE(\beta_0) \text{..... (11)}$$

Where $Z_{1-\alpha/2}$ is the upper 100(1-α /2) % point from the standard normal distribution and SE Denotes a model based estimator of the standard error of the respective parameter estimator.

2.5. Goodness of Fit Statistic:

Instead of using maximum likelihood estimation we could estimate the parameters by minimizing sum weighted of squares

$$S_w = \sum_{i=1}^{N} \frac{(y_i - n_i\Pi_i)^2}{n_i\Pi_i(1-\Pi_i)} = 0 \text{ ..... .(12)}$$

Since

$$E(y_i) = ni\ \Pi_i$$

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{(y_i - n_i\Pi_i)^2}{n_i\Pi_i(1-\Pi_i)}\right)(1 - \Pi_i + \Pi_i) = S_w \text{.....(13)}$$

When χ^2 is evaluated at the estimated expected frequencies, the statistic is

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{\left(y_i - n_i\hat{\Pi}_i\right)^2}{n_i\hat{\Pi}_i\left(1-\hat{\Pi}_i\right)}\right) \quad \text{.....(14)}$$

This is asymptotically equivalent to

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\Pi_i}\right) + \left(n_i - y_i\right) + \log\left(\frac{n_i - y_i}{n_i - n_i\Pi_i}\right)\right] \quad \text{.....(15)}$$

The asymptotically distribution of D, under the hypothesis that the model is correct, is $D \approx \chi^2(N\text{-}p)$

The choice between D and χ^2 depends on the adequacy of the approximation to the $\chi^2(N\text{-}P)$ distribution. There is some evidence to suggest that χ^2 is often better than D because D is unduly influenced by very small frequencies.

Both the approximations are likely to be poor, however, if the expected frequencies are too small (e.g., less than1).

2.6. Wald Test:

To test the significance of each coefficient (β_i) is measured by the weld statistic, given by

$$\text{Wald} = \left[\hat{\beta}_i / \text{s.e}\left(\hat{\beta}_i\right)\right]^2 \quad \text{.....(16)}$$

This is distributed as chi-square with one degree of freedom in addition

to that Wald statistic is an alternative test which is commonly used to test the significance of individual logistic regression coefficients for each independent variable that is to test the null hypothesis in logistic regression that particular legit (effect) coefficient is zero. For dichotomous independents, the weld statistic is the square ratio of UN standardized legit coefficient to its standard error.

Menard warns that for large legit coefficient standard error is inflated, lowering the weld statistic leading to type tow error. That is, there is flaw in the weld statistic such that very large affects may lead to large standard errors and small weld chi-square values. For models with large legit coefficients or when dummy variable involved, it is better to test the difference using the likelihood ratio test of the difference of the models with and without parameter. Also weld statistic is sensitive to violation of the large-sample assumption of logistic regression. For these reason, the likelihood ratio test of individual model parameter is generally preferred.

2.7. The Likelihood Ratio:

It is a function of log likelihood, because -2LL has approximately a chi-square distribution, -2LL can be used for assessing the significance of logistic regression, analogous to the use of the sum of squared error is ordinary least square (OLS) regression. The -2LL statistic is the likelihood ratio. It is also called goodness of fit, deviance chi-square, scaled deviance, deviation chi-square, DM, or L-square. It reflects the significance of the unexplained variance in the dependent, in the SPSS output, this statistic isfound in the "-2log likelihood" column of the iteration history table or the "likelihood ratio test" table. The likelihood ratio is not used directly in significance, but it is the basis for the likelihood ratio test, which is the test of difference between two likelihood tests. This is the test of the difference between two likelihood ratios (two -2LLs). The likelihood ratio test is based on -2LL (deviance). This likelihood ratio test is a test of the significance of the difference between the likelihood ratio (-2LL) for the researchers model minus the likelihood ratio for a reduced model. It is an alternative to the weld statistic and is also called the log-likelihood test. There are three main forms of the likelihood ratio test,(Elhady, A., 2009).

The following formula of likelihood ratio test is:

$-2 \text{ Log } (L_0 / L_1) = -2 [\text{ Log } (L_0) - \text{ Log } (L_1)] = -2 [L_0 - L_1]$.....(17)

2.8. Model Building:

The goal of building a model is to select those variables that result in a best model with in the scientific context this goal we must have two steps:

1. A basic plan for selecting the variable for the model.
2. Asset of method for assessing the adequacy of the model both in terms of its individual variables and its overall fit,(John, Neter, 1996).

There are several steps one can follow to aid in the selection of variables for a regression model. The selection process should begin with a careful invariable analysis of each variables for nominal, ordinal and continuous

variables with few integer values. We suggest this to be done with a contingency table of outcome (y=0.1) versus the level of the independent variable.

.The likelihood ratio chi-square test with (k-1) df is exactly equal to the value of the likelihood ratio test for the significance of the coefficient for the (k-1) design variable in a un invariable logistic regression model that contains that single independent variable in addition to the overall test. It is usually preferable those variable exhibiting at least a moderate level association along with confidence limit using one of the levels as the reference group.

3. Data Analysis and Application:

Descriptive analyses of cancer data were performed by SPSS. The entire variables that were either statistically associated with breast cancer in the multiple logistic regression models were fitted to obtain independent estimates of the risk of breast cancer. Modeling started with all variables followed by sequential deletion according to their statistical importance. Each variable was assessed through the Wald test and models were compared using the likelihood ratio test. Model fitting cannot establish to truth of any particular theory of carcinogenesis, through it may be possible to test particular hypotheses within given general class of models.

3.1. Risk Factor for Female Breast Cancer:

Any woman may develop breast cancer. However, the following risk factors may increase the likelihood of developing the disease. Both genetic and environmental factors are believed to play a role in a woman's risk of developing breast cancer. If either a woman's mother or sister has breast cancer, the woman's risk increases about two to three times. Having both a mother and a sister with breast cancer increases a woman's risk up to six-fold.

If that relative had bilateral breast cancer or was diagnosed at an early age, the risk may be further increased. In small group of families, the patterns of breast cancer incidence seem to be consistent with known patterns of genetic inheritance.

The common risk factors of breast cancer available in this research are: age, region, occupation, ethnic group, morphology, marital status and stage of the disease.

3.2. The Population and Sample Size:

Data were collected from secondary source from NCI, Gezira Cancer Registry from (2005 to 2008). The incidence rate of breast cancer was 30.90% from all malignancies. To determine the sample size was taken of all dead which amounted to 110 deaths and have been taken 110 survivors randomly as control.

3.3. Statistical Analysis:

All data were coded and entered into computer. Editing of the data was done as when needed. All the data were scrutinized for advanced analysis. For this research on breast cancer, the outcome variable was binary. Its value (1) represented the live and (0) the dead. The data were collected for 110 dead and 110 lives, for all the variables (divided into a number of groups) variables measured on breast cancer were tested by using chi-square test. Some of the significant and strong associations were mentions. The preliminary analysis of the data set was

carried out. The comparisons between two groups (dead and live) of data were made for all the independent variables. The categorical independent variables were arranged in tables and chi-square tests were applied to compare the difference between dead and live. The preliminary analysis of the data set was carried
out. The comparisons between two groups (dead and live) of data were made for all the independent variables. The categorical independent variables were arranged in tables and chi-square tests were applied to compare the difference between dead and live.

Region: We note that for the dead group37.3% from Greater Wad Medani, 8.2% from El Hassahesa, 5.5% from El Kamlin, 14.5% from East Gezira, 12.7% from south Gezira, 6.3% from Omalgora, and 15.5 % from El Managil. From the live group, 31.5% from Greater Wad Medani, 15.5% from El Hassahesa, .9% from Alkamline, 14.5% from East Gezira, 13.2% from south Gezira, 10% from Omalgora, and 13.6% from El Managil. The difference between the distributions of the two groups with respect to region is not significance (p-value= .273).

Occupation: We note that from the dead group, 70.8% are not work, 11% are formal work, 9.1% are informal work, and 9.1% are unknown. From the live group, 64.5% are not work, 11.8% are formal work, 17.2 % are Informal work, and 6.5% are unknown. The difference between the distribution of the two groups with respect to occupation is not significance (p-value = .271).

Stage: We note that from the dead group, 20% are in stage1, 13.6% are in stge2, 17.2% are in stage3, and 49.2% are in stage4. From the live group, 31% are in stage1, 26.4% are in stage2, 30% are in stage3, and 12.6% in stage4. The difference between the distribution of the two group with respect stage is significance (p-value = .000).

Ethnic group: From the dead group, 30% from North Sudan, 4.5% from West Sudan, 23.5% from East Sudan, 0% from South Sudan, 36.4% from Central Sudan, and 5.6% are unknown. From the live group, 23,4%from North Sudan , 25,4% from West Sudan, 15%% from East Sudan, 0% from
South Sudan, 17.2% from Central Sudan, and 19% are unknown. The difference between the distribution of the two groups with respect to ethane group is significance (p-value = .000).

Morphology: For the dead group, 40% are Infiltrating duct carcinoma, 41% are carcinoma, 0% are adino carcinoma, 5.5% are Infiltrating duct-lobur-mixed with other type, 3.7% are neoplasm malignant, 1.8% are Polly tumor malignant, 5.5% are Paget disease, and 2.5% are unknown. For the live group, 48.3% are Infiltrating duct carcinoma, 40% are carcinoma, 1.8% are adino carcinoma, 4.5% are Infiltrating duct-lobur-mixed with other type, 3.7% are neoplasm malignant, .9% are Paget disease, .9% are Polly tumor malignant, and no anyone are unknown. The difference between the distribution of the two groups with respect to ethnic group is significance (p-value = .198).

Diagnoses: For the dead group, 3.8% are Clin-invest/ult sound, 1.8% is Clinic only, 17.2% are Cytology, 0% is Histology of metastases, 74.5% are Histology of primary, and 2.7% are unknown. For live group, 0% is Clin-investlult sound, 0% are Clinic only, 22.7% are Cytology, 1.8% is Histology of metastases, 69.2% are Histology of primary, and 6.3% are unknown. The difference between the distribution of the two groups with respect to basis of diagnoses is not significance (p-value = .059)

Marital status: For dead group, 8.2% are single, 70.9% are married, 3.6% are widow, 0% is divorce, and 17.3% are unknown. For live group, 6.4% are single, 53.6% are married, 0% is widow, 12.7% are divorce, and 27.3% are unknown. The difference between the distribution of the two groups with respect to marital status is significance (p-value = .000).

Age group: (age groups according to their menstruation status; group1<40 year, group2 years between 40-50 year and group 3 >50 year).

For dead group, 39% are in group1, 21% are in group, 2, 40% are in group3. For the live group, 39% are in group1, 20% are in group2, 41% are in group3. The difference between the distribution of the two groups with respect to age group is not significance (p-value = .983).

Table (1) Distributions of dead and live according to Region, Occupation, Stage, Ethan group, Morphology Basis of diagnoses, Marital status, and age.

Factor	Dead		live		total		p-value
	F	%	F	%	f	%	
Region							
Madani alkobra	41	37.3	35	31.5	76	34.5	
Alhasahia	9	8.2	17	15.5	26	11.8	
Alkamline	6	5.5	1	.9	7	3.2	
East gezira	16	14.5	16	14.5	32	14.5	.273
South gezira	14	12.7	15	13.6	29	13.2	
Omalgora	7	6.3	11	10	18	8.2	
almanagle	17	15.5	15	13.6	32	14.5	
Occupation							
Not work	79	70.8	71	64.5	150	68.2	
Formal work	11	11	13	11.8	24	10.9	
Informal work	10	9.1	19	17.2	29	13.2	.271
unknown	10	9.1	7	6.5	17	7.7	
stage							
Stage1	22	20	34	31	56	25.5	
Stage2	15	13.6	29	26.4	44	20	
Stage3	19	17.2	33	30	52	23.6	.000
Stage4	54	49.2	14	12.6	68	30.9	
Ethan group							
North Sudan	33	30	25	23.4	59	26.8	
West Sudan	5	4.5	28	25.4	33	15	
East Sudan	26	23.5	16	15	42	19.1	
South Sudan	0	0	0	0	0	0	
Middle Sudan	40	36.4	19	17.2	59	26.8	.000
unknown	6	5.6	21	19	27	12.3	
Morphology**							
Infiltrating duct carcinoma	44	40	53	48.3	97	44.1	
carcinoma	45	41	44	40	89	40.5	
Adina carcinoma	0	0	2	1.8	2	.9	
Infiltrating duct-lobur-mixed with other type	6	5.5	5	4.5	11	5	
Neoplasm malignant	4	3.7	4	3.7	8	3.6	
Pllode tumor malignant	2	1.8	1	.9	3	1.4	
Paget disease	6	5.5	1	.9	7	3.2	.198
unknown	3	2.5	0	0	3	1.4	
Basisof diagnoses							
Clin-invest/ult sound	4	3.8	0	0	4	1.2	
Clinic only	2	1.8	0	0	2	.9	
Cytology	19	17.2	25	22.7	44	20	
Histology of metastases	0	0	2	1.8	2	.9	
Histology of primary	82	74.5	76	69.2	158	71.7	
Unknown	3	2.7	7	6.3	10	4.8	0.59
Marital status							
Single	9	8.2	7	6.4	16	7.3	
married	78	70.9	59	53.6	137	62.3	

widow	4	3.6	0	0	4	1.8	.000
divorce	0	0	14	12.7	14	6.4	
unknown	19	17.3	30	27.3	49	22.3	
Age group							
Group1	43	39	43	39	86	39	
Group2	23	21	22	20	45	20.5	.983
Group3	44	40	45	41	89	40.5	

Source: SPSS software output.

*f: frequency. **: Wittekind. L. H. Sobinard CH. (2002).

Multiple logistic regression models were applied to research the significance of each characteristic towards breast cancer. Conditional logistic regression was used to calculate the matched odds ratio and 95% confidence intervals for a Multiple analysis .The independent variables with p-value < 0.25 were selected for the multiple logistic regression models because use of the traditional level (such as 0.05) in logistic regression often fails to identify variables known to be important(Mickey and Greenland, 1989).

At first we create a reduce models (model without predictor), then full model (model with predictors). The different between the steps is the predictors that included or not. By default, statistical analysis programs in logistic regression analysis are run in two steps. The first step, called step 0, includes no predictors and just intercept. Often, this model is not interesting to researchers. The second step 1 and this is the model with predictors in this case; it is the full model that was specified in the logistic regression command. And we depend on this step in our analysis. From the chi-square test for characteristic of the study population some variable are significance and the other are insignificance.

From all the above variables stage, marital status, region, occupation, morphology, age group, basis of diagnosis and ethnic group can be conceded as a risk factor of breast cancer.

Multiple logistic regressions fit by the all variables. The test of the model from table (3.2), the interaction terms observed at p-value < 0.25 were discussed below from table (3.3), and finally the goodness of fit test shown in table (3.4), and multiple logistic regressions fit by significance variable. The test of the model from table (3.5), the estimated coefficient shown in table (3.6), and finally the goodness of fit test shown in table (3.7).

Multiple logistic regression fit by the all variables, region, occupation, ethnic group, basis of diagnoses, stage, marital status, morphology, and age group.

From table (3.2) below presents the test of the model. The -2log likelihood = (304.985), (276,810) for Null and Full model respectively. Chi-square = (28.175) with p-value = (.001). Which it means the overall significance of the model statistically is significance. From this result model can be put into consideration and its result is acceptable.

From table (3.3) including all variables because their clinical important to fitting multiple model. Table (3.3) presents the result of fitting this model.

For multiple regressions model including all variables the important variables are tribe, stage and marital status. If p-value < 0.25 as level of significance then morphology can be used as important variables to develop breast cancer and it consider as a risk factor for breast cancer.

The Odd-Ratio for the stage shows that, those women in the late stages are (1.66) times more likely to dead compared with those who have early stage. Those who are single are (1.262) times more likely to dead compared with those who are married. 95% confidence interval indicated that the value of OR ranged between (.462, 0.785), (.998, 1.598) for stage and marital status respectively.

Table (3.4) below show that the hosmer-lemeshow test done to test the goodness of fitting: H-L = (15.589), with p-value = (.049). So as there there is a difference between the observed and predicted values the model is not fitting the data.

Table (2) Test of the model

Model	-2logliklhood	Chi-squire	p-value
Null	304.985	28.175	.001
Full	276.810		

Source: SPSS software output.

Table (3) The estimated coefficients of multiple logistic regression model for all variable

Effect	B	S.E	Wald	Sig	OR	95.0% CI for OR
Region	.067	.070	.924	.336	1.070	.932- 1.227
Occup	-.112	.164	.464	.496	.894	.649 -1.233
Tribe	-.028	.084	.116	.734	.972	.825-1.145
Morph	-.157	.099	2.511	.113	.854	.703-1.038
Basis	-.003	.148	.001	.982	.997	.745-1.333
Stage	-.507	.135	14.029	000	.602	.462-0.785
Mstatus	.233	.120	3.764	.050	1.263	.998-1.598
Age group1	.026	.400	.016	.992	1.006	.467-.250
Age group2	-.014	.403	.001	.972	.986	.447-2.173
Age group3	.027	.401	.004	.947	1.027	.468-2.253

Source: SPSS software output.

Table (4) Goodness of fit test

	Chi-squire	p-value
Hosmer- lemeshow	15.589	.049

Source: SPSS software output.

Then comes the multiple logistic regressions fit by the significant variables of tribe, stage, and marital status.

Table (3.5) below, presents the test of the model where the -2log likelihood = (304.985), (281,002) for Null and Full model respectively. Chi-square = (23.983) with p-value = (.000). These indicate that the overall significance of the model is statistically significant, and the result model can be put into consideration.

Table (5) Test of the model

Model	-2logliklhood	Chi-squire	p-value
Null	304.985	23.983	.000
Full	281.002		

Source: SPSS software output

From table (3.6) including significant variables .Table (3.6) presents the result of fitting this model. For multiple regression models the values of Wald statistic suggest that two variable are significant namely, stage and marital status with p-values = (.000), (.074) respectively.

The OR for the stage shows that, those who are at late stages are (.65) times more likely to die compared with those who are at early stages. Those who are single are (1.223) times more

likely to die compared with those who are single. 95% confidence interval indicated that the value of OR ranged between (.478, .792), (.981, 1.526) for stage and marital status respectively. From table (3.7) below that hosmer- lemeshow test done to test the
Goodness of fitting. H-L = (13.060) with p-value = (.071) so there is no difference between observed and predicted values, and the model is fitting the data.

Table (6) The estimated coefficients of multiple logistic regression o model for Variables with highly significance

Effect	B	S.E	Wald	Sig	OR	95.0% CI for OR
Tribe	-.014	.082	0.031	0.860	0.986	0.840, 1.157
Stage	0.486	0.129	14.278	0.000	0.615	0.478, 0.792
M status	.201	0.113	3.197	.074	1.223	0.981, 1.526

Source: SPSS software output.

Table (7) Goodness of fit test

	Chi-squire	p-value
Hosmer-lemeshow	13.060	.071

Source: SPSS software output

4. Conclusion and Recommendations:

Chi-square test was applied to establish the characteristics of the study population. The results showed that the tribe, marital status and stage are the significance variables.
The logistic regression technique is most commonly used for analysis of epidemiological studies. Logistic regression models have been applied to
this case–control study and the results obtained from various models are compared.
After analysis of the data the region, occupation, morphology, basis of diagnoses and age group are not significant variables and these variables cannot play an important role to develop breast cancer among Sudanese women.
In multiple logistic regression models the research explain that stage, ethnic group and marital status as important variables to develop breast cancer and are considered as risk factors of breast cancer.
The study made the following recommendations**:**

- The awareness about the dangers of the disease is still very weak, so community, government, civil society, and non-governmental organizations should play an important role to raise such awareness.
- Early detection should be integrated into primary health care
- Early diagnosis and screening, specifically targeted towards breast cancer should be carried out.
- In principle, the case – control data used in this study, together with the cancer registry data, can be used to construct models of absolute risk.

REFERENCES

1. Christopher, B. S. (2006). Introduction to Mathematical Modeling of Crop Growth: How the Equations are Derived and Assembled into a Computer Model, Ph D Dissertation, Faculty of Agriculture, University of Putra Malaysia.
2. Davied, W. Hosmer and Stanley Lemeshow (2000).Applied Logistic Regression, Second Edition, Wiley, Inc, New York
3. Elhady, A, (2009). Application of Logistic Regression to Estimate Children Breastfeeding, unpublished M.Sc Dissertation, Department of Applied Statistics and Demography, University of Gezira.
4. John, Neter, Michaeld Kutner, Christopher J.N. Achtsheim and William Wassermah (1996), "Applied Linear Statistical Models", Fourth Edition.
5. Mickey, J., & Greenland, S. (1989). A Study of the Impact of Confounder Selection Criteria on Effect Estimation. American Journal of Epidemiology, 129: pp 125 – 137.
6. Wittekind L. H. Sobinard CH. (2002), Classification of Malignant Tumor, six editions, International Union against Cancer (UICC).