*IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*
*Volume 21, Issue 5, Ver. 1 (May. 2016) PP 22-34*
*e-ISSN: 2279-0837, p-ISSN: 2279-0845.*
*www.iosrjournals.org*

# Non-parametric Survival Analysis for Breast Cancer Using non-medical Data

Dr. Intisar Ali A. Khalil,
*Saudi Arabia, Umluj Area et al[1]Faculty of Sciences-University of Tabuk, P.O. Box 471Umlj College.*

***Abstract:*** *Breast cancer is a major health problem in many parts of the world. Breast cancer has the highest prevalence rate among women in Saudi Arabia. This study aim to review the validity and workability of semi and non-parametric survival models in non-medical data. The methodology of sisterhood method for mortality data collection is used, by designing questionnaire. Non-medical variables; age, residence, weight, family history, fertility and marital status are used to show the differences in survival analysis. A total of 32 female breast cancer cases were studied. The results indicates that premenopausal, fertile cases were more survive and more diagnosis, while obese cases and those who live in Umluj were less survive and more diagnosis. The median survival time is (82.64) with mean age at diagnosis (45.59).The proportion surviving from general 10-year survival study was decreased gradually with the time interval and the hazard rate increase shortly with the time interval.*

المستخلص
يعد سرطان الثدي مشكلة صحية في أجزاء من العالم,سجل أعلى معدل انتشار بين النساء في السعودية. تهدف هذه الدراسة إلى توضيح صلاحية وفاعلية تطبيق نماذج البقاء علي البيانات غير الطبية. استخدامت منهجية طريقة الأخوات لجمع البيانات. استخدمت النماذج شبه وغير المعلمية لتحليل البيانات. المتغيرات غير الطبية: العمر، مكان الإقامة، الوزن، الخصوبة, التاريخ العائلي والحالة الاجتماعية استخدمت لإظهار الاختلافات في البقاء على قيد الحياة. العينة مكونة من 32 من الاناث المصاباتبسرطان الثدي. النتائج بينت أن الاناث الخصبات و الاناث في فترة ما قبل انقطاع الطمث أكثر بقاء علي قيد الحياةواقل اصابة بالمرض، في حين أن حالات السمنة و اللائي يعيشنداخل مدينةأملج كانت أقل بقاءأ على قيد الحياة واكثر اصابة. متوسط البقاءعلي قيد الحياة ( 82.64) شهراو متوسط العمر عند التشخيص (45.6). انخفضت نسبة الباقين على قيد الحياة عند تطبيق الدراسة لمدة 10 سنوات, تدريجيا معالزمن, في حين زادت معدلات الخطورة ببطء مع الزمن.

## I. INTRODUCTION

At an individual level, diagnosis of cancer is generally regarded as a human tragedy. At the level of society, cancer is one of the major chronic diseases, causing a notable amount of health administrative costs. Prognosis and possible cure from cancer are thus important measures of life span which can be assessed by analyzing the survival of cancer patients. Different statistical approaches are used in the literature to analyzing the cancer survival data. The results of survival analysis for cancer patients have been widely presented and reported for different human sub populations of the globe (Woolson, 1981[2]; Kardaun, 1983[3]; Beadle *et al*, 1984; Sedmak *et al*., 1989[4]). McCarty (1974) has mentioned that for adopting any suitable statistical technique for analyzing survival data, it should be assumed that the statistical model embody the evaluation of some natural processes believing that the model is a useful approximation of a real process. Several approaches have been proposed in the literature by Leung *et al.* (1997) [5]and Little and Rubin (2002) for analyzing the survival data. Investigators are often inclined to use conventional statistical methodology for the analysis of survival data. Logistic regression analysis could be applied to quantify the importance of certain covariates in classifying individuals into two groups: those who did or did not experience the event during the period of observation. This

[1]Dr. Maria Zakaria Adam Hashim, Department of Mathematic , Umluj college- university of Tabuk

[2]**Woolson, R.F., (1981).** Rank test and a one sample log rank test for comparing observed survival data to standard population. Biostatistics, 37:687-696.

[3]**Kardaum,(1993).** Statistical analysis of male larynx cancer patients: A case study. Statistical Nederlandica, 37:103-126.

[4]**Sedmak, D. D., T. A. Meineke, D. S. Knechtges and J. Anderson, (1989**): Prognostic significance of cytokeratin-positive breast cancer metastases. Modern Pathol, 2: 519-520.

[5]**Leung, K.M., R.M. Elashoff and A.A. Afifi, (1997).** Censoring issues in survival analysis. Annual Review of Public Health, 18:83-104

approach can result in a considerable loss of information because differences in the timing of event occurrence are not considered. Alternatively, one could use ordinary least squares regression analysis to identify covariates that influence survival times. The major drawback here lies in the fact that survival data are often censored, i.e., they contain observations for which one does not know when the event has occurred. This is either because the corresponding individual is lost from the data set during the study period, or because the study has ended before all individuals have experienced the event. Both causes of censoring occur commonly during epidemiology investigations. While conventional statistical methodology, censored observations would either have to be deleted, or one would have to make certain ad-hoc assumptions. By contrast, the likelihood-based parameter estimation methods used in survival analysis can effectively extract relevant information from both censored and uncensored observations, thereby producing reliable parameter estimates (Allison,1995)[6] and (Le,1997)[7]. Furthermore, survival analysis is the only method that can readily accommodate time dependent covariates, i.e., independent variables whose values change during the course of the study.

## II. RESEARCH PROBLEM AND OBJECTIVES

This study attempted to apply the non-parametric approach to breast cancer survival data in order to show their applicability and workability in non-medical data and to present survival analysis results for breast cancer data. The problem of analyzing censored data is usually referred to as survival analysis, which is a model time to failure or event. Survival analysis is unlike linear regression survival analysis which has binary outcome and also unlike logistic regression, because survival analysis analyses the time to an event.

However, Survival analysis in oncology is complicated since not all patients can be observed for the same period of time (Kosko, 1992)[8].In survival analysis terminology, patients who are observed until they reach the end point (e.g. death) are called uncensored cases while those who survive further than the end of the study or who are lost to follow-up at some point are called censored cases. Traditional methods of analysis of censored data rely on linear models (statistics), statistical methods such as the life-table, the Kaplan-Meier method and regression models such as the Cox Proportional Hazards are typically used to model and predict survival data with the ability to handle censored data. Survival data can be represented statistically by the probability density function, survival function or hazard rate function. Survival statistics indicates a cohort of patients with certain types and stages of cancer and is measured following treatment.

Statistics alone may not be sufficient to predict the future outcome of a particular patient, as no-two patients are exactly alike (Kosko, 1992).The main objective of this study is to review the methodology and workability of semi non-parametric survival models to non-medical data, and their application to breast cancer data in Saudi Arabia.

## III. MATERIALS AND METHODS

**Population and sampling:**

This study was planned to investigate the survival analysis of breast cancer among Saudi Arabian's women. Umluj area was selected as a case study for this research. The only health information agency in this area is Al-Hawra Hospital. It is one of the government hospitals that have been established in Tabuk state in the North West region of Saudi Arabia. The hospital has five health units in addition to outpatient clinics and emergency department. There is no unit of oncology and radio therapy treatment, so the hospital registry system shows that all patients that have been suspected to cancer were referred to other hospital in the area such as (Alwagh, Yanbu, and Tabouk). Therefore, the study uses primary data to briefly review the methodological features of the semi and non-parametric survival models. The study deriving such data from female diagnosed with breast cancer using the approach of sisterhood method. A questionnaire designed to collect information about breast cancer patient in the family. A total of 100 questionnaires forwarded to students in Umluj College, to bring information about diagnosed relatives with breast cancer. The response rate of the population is very low, only 36 questionnaires were retained and 4 questionnaires of them are not completed so they excluded. So researchers decide, with the consideration that the approach is no-parametric, to conduct the study using only 32 complete cases, see annex (A) table 1. Student enrolled to Umluj College, were from Umluj and different

[6]**Allison, P.D. (1984).** Event history analysis: regression for longitudinal event data. Beverly Hills, CA: Sagle publication

[7]**Lee et al, (1992);** Statistical methods for survival data analysis. 2nd edition. Wiley, New York. 482pp. (comprehensive, no easy reading).

[8]**Kosko, B (1992).** Neural networks and fuzzy systems. Dynamical systems approach to machine intelligence. 1st ed. Prentice-Hall International Editions,

villages around it. The researcher used the student as population units because the population is closed and very sensitive toward such information.

**The study models: Semi and Non-parametric models:**

The probability of surviving beyond t is

$$S(t) = \Pr(\mathbf{t} > t) \tag{1.1}$$

Because t cannot be negative, s (0) =1

S (t) can be estimated by the Kaplan-Meier method (Kaplan and Meier,1958)[9]

$$\hat{S}(t) = \prod_{j:t_j \geq t} \left[ 1 - \frac{d_j}{N_j} \right] \tag{1.2}$$

Where $N_j$ is the number of cases at risk of an event at time $t_j$ and $d_j$ is the number of event at time $t_j$.

Instantaneous risk that an event occurs in the small interval between t and $t+\Delta t$ *is*

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq t \prec t + \Delta t \,/\, t \geq t\}}{\Delta t} \tag{1.3}$$

A hazard is a rate not a probability.

**One Sample Kaplan-Meier**

If the data were not censored, the obvious estimate would be the empirical survival function

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} I\{t_i \succ t\}, \tag{1.4}$$

Where, *I* is the indicator function that takes the value 1 if the condition in braces is true and 0 otherwise. The estimator is simply the proportion alive at *t*.

**Estimation with Censored**

Kaplan and Meier (1958) [10]extended the estimate to censored data. Let

$$\mathbf{t}_{(1)} \prec \mathbf{t}_{(2)} \prec \ldots \prec \mathbf{t}_{(m)} \tag{1.5}$$

Denote the distinct ordered times of death (not counting censoring times). Let $d_i$ be the number of deaths at $t_{(i)}$, and let $n_i$be the number alive just before $t_{(i)}$. This is the number exposed to risk at time $t_{(i)}$. Then the Kaplan-Meier or product limit estimate of the survivor function is

$$\hat{S}(t) = \prod_{i:t_{(i)} \prec t} \left( 1 - \frac{d_i}{n_i} \right). \tag{1.6}$$

A heuristic justification of the estimate is as follows. To survive to time t you must first survive to $t_{(1)}$. You must then survive from $t_{(1)}$ to $t_{(2)}$ given that you have already survived to $t_{(1)}$.And so on. Because there are no deaths between $t_{(i-1)\ and}\ t_{(i)}$, we take the probability of dying between these times to be zero. The conditional probability of dying at $t_{(i)}$ given that the subject was alive just before can be estimated by $d_i/n_i$. The conditional probability of surviving time $t_{(i)}$ is the complement $1 - d_i / n_i$. The overall unconditional probability of surviving to t is obtained by multiplying the conditional probabilities for all relevant times up to t. the Kaplan-Meier estimate is a step function with discontinuities or jumps at the observed death times. If there is no censoring, the K-M estimate coincides with the empirical survival function. If the last observation happens to be a censored case, the estimate is undefined beyond the last death.

**Non-parametric Maximum Likelihood (NPML)**

The K-M estimator has interpretation as a non-parametric maximum likelihood estimator (NPML). Let $c_i$ denote the number of cases censored between $t_{(i)}$ ant $t_{(i+1)}$, and let $d_i$ be the number of cases that die at $t_{(i)}$. Then the likelihood function takes the form

$$L = \prod_{i=1}^{m} \left[ S(t_{(i-1)}) - S(t_{(i)}) \right]^{d_i} S(t_{(i)})^{c_i}, \tag{1.7}$$

[9]**Kaplan, E. L. and Meier, P. (1958):** Nonparametric estimation from incompletes observations. Journal of American Statistical Association, 53:457-451.

[10]**Kaplan, E. L. and Meier, P. (1958):** Nonparametric estimation from incompletes observations. Journal of American Statistical Association, 53:457-451.

Where the product is over the *m* distinct times of death, and we takes $t_{(0)}=0$ with $S_{(t(0))}=1$. the problem now is to estimate m parameters representing the values of survival function at the death times $t_{(1)}$, $t_{(2)}$, …,$t_{(m)}$.Write $\pi_i = S(t_{(i)})/S(t_{(i-1)})$ for the conditional probability of surviving from $S(t_{(i-1)})$ to $S(t_{(i)})$. Then the likelihood becomes

$$L = \prod_{i=1}^{m} (1-\pi_i)^{d_i} \pi_i^{c_i} (\pi_1 \pi_2 \ldots \pi_{i-1})^{d_i + c_i}. \tag{1.8}$$

Note that all cases which are at $t_{(i)}$ or are censored between $t_{(i)}$ and $t_{(i+1)}$ contribute a term $\pi_j$ to each of the previous times of death from $t_{(i)}$ to $t_{(i-1)}$addition, those who die at $t_{(i)}$ contribute $1-\pi_i$, and the censored cases contribute an additional $\pi_i$. let $n_i = \sum_{j \geq i} (d_j + c_j)$ denote the total number exposed to risk at $t_{(i)}$. We can then collect terms on each $\pi_i$ and write the likelihood as:

$$L = \prod_{i=1}^{m} (1-\pi_i)^{d_i} \pi_i^{n_i - d_i} \tag{1.9}$$

a binomial likelihood. The *m.l.e.* of $\pi_i$ then

$$\hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i} \tag{1.10}$$

The K-M estimator follows from multiplying these conditional probabilities.

**Expectation of Life (life tables)**

If $\hat{S}(t(m)) = 0$ then one can estimate $\mu=E(T)$ as the integral of the K-M estimate:

$$\hat{\mu} = \int_0^{\infty} \hat{S}(t)dt = \sum_{i=1}^{m} (t_{(i)} - t_{(i-1)}) \hat{S}(t_{(i)}) \tag{1.11}$$

Can you figure out the variance of $\hat{\mu}$ ?

**Regression: Cox's Model**

Let us consider the more general problem where the researcher has a vector *x* of covariates. The k-sample problem can be viewed as the special where the *x'* are dummy variables denoting group membership. Recall the basis model

$$\lambda(t, x) = \lambda_0(t) e^{x'B}, \tag{1.12}$$

And consider estimation of β without making any assumptions about the baseline hazard $\lambda_0(t)$ . Cox, (1972)[11] proposed fitting the model by maximizing a special likelihood. Let

$$t_{(1)} \prec t_{(2)} \prec \ldots \prec t_{(m)} \tag{1.13}$$

Denote the observed distinct times of death, as before, and consider what happens at $t_{(i)}$. Let $R_i$ denote the risk set at $t_{(i)}$, defined as set of indices of the subjects that are alive just before $t_{(i)}$. Thus, $R_0 = \{1,2,..., n\}$. suppose first that there are no ties in the observation times, so one and only person subject failed at $t_{(i}$

$$\frac{e^{x'_{j(i)}\beta}}{\sum_{j \in R_i} e^{x'_j \beta}} \tag{1.14}$$

and does not depend on the baseline hazard $\lambda_0(t)$ .

Cox proposed multiplying these probabilities together over all distinct failure times and treating the resulting product

$$L = \prod_{i=1}^{m} \frac{e^{x'_{j(i)}\beta}}{\sum_{j \in R_i} e^{x'_j \beta}}. \tag{1.15}$$

[11]**Cox, D. R (1972**): Regression models and life tables. Journal of the Royal Statistical Society Series B, 34:187-220.

As if it was an ordinary likelihood. Cox (1975)[12] calls this a "conditional likelihood" because it is a product of conditional probabilities. Kalbfleisch and Prentice**(1973)**[13]consider the case where the covariates are fixed over time and showed that L is the marginal likelihood of ranks of the observations, obtained by considering just the order in which people die and no the actual times at which they die.

**Tests of Hypotheses:**

As usual, there are three approaches to testing hypotheses about $\hat{\beta}$ :

-**Likelihood Ration test**: given two nested models, the researcher treats twice the difference in partial log-likelihoods as a $\chi^2$ statistic with degrees of freedom equal to the difference in number of parameters.

-**Wald test**: using the fact that approximately in large samples $\hat{\beta}$ has a multivariate normal distribution with mean β and variance-covariance matrix var( βa = $I^{-1}$(ββ) Thus, under $H_0 : \beta = \beta_0$, the quadratic form

$$(\hat{\beta} - \beta_0)' \text{var}^{-1}(\hat{\beta})(\hat{\beta} - \beta_0) \sim \chi^2_p \qquad (1.16)$$

where *p* is the diminution of *β*. This test is often used for a subset of β.

-**Score Test**: using the fact that approximately in large samples the score $u_{(\beta)}$ has a multivariate normal distribution with mean 0 and variance-covariance matrix equal to information matrix. Thus, under $H_0 : \beta = \beta_0$, the quadratic form

$$u(\beta_0)' I^{-1}(\beta_0) u(\beta_0) \sim \chi^2_p \qquad (1.17)$$

Note that this test does not require calculating the M.L.E. $\hat{\beta}$ . One reason for bringing up the score test is that unlike the k-sample case the score test of $H_0 : \beta = 0$ based on Cox's models happens to be the same as the Mantel-Haesnszel log-rank test. All the three tests are asymptotically equivalent. The equality of the normal approximations depends on sample size, the distribution of cases over the covariate space, and the extent of censoring.

## IV. RESULTS

This section contains a reviewing of results of application of the semi and non-parametric survival models to the data.

**3.1. Life table estimates of patient survival:**

A life table for the 32 patients, which is constructed for annual intervals and uses the actuarial assumption, is shown in Table 6. The first column in Table (6) gives the index (*i*) for the interval, followed by the number of patients alive at the start of the interval ($l_i$), the number of deaths (irrespective of cause) during the interval ($d_i$), the number of censorings during the interval ($w_i$), and the effective number of patients at risk during the interval (*l-i*).

The effective number of patients at risk is given by ($l_i - wi/2$). The notation $w_i$ is used for the number of censored observations since these are sometimes referred to as 'withdrawals' or 'patients withdrawn alive'. The column labeled pi contains the estimated conditional survival rates of surviving each interval among the patients alive at the start of the interval ($p_i = 1 - d_i/l\_ i$). When estimated from a life table, the conditional survival rates are known as interval-specific survival rates. When the intervals are annual, as in Table 6, they are referred to as annual survival rates. The column labeled 1*pi* contains estimates of the cumulative survival rates from diagnosis to the end of the $i^{th}$ interval, calculated as the product of the interval-specific survival rates for each interval from 1 to *i*. The 'cumulative' prefix is often dropped from the term 'cumulative survival rate' and we will sometimes follow this practice in the text. The remaining two columns in the life table contain the cumulative expected (*1p∗ i*) and relative (*1ri*) survival rates, which will be described later in this section. Life tables are a descriptive procedure for examining the distribution of time-to-event variables. The researcher also can compare the distribution by levels of a factor variable. The basic idea of life tables is to subdivide the period of observation into smaller time intervals. Then the probabilities from each of the intervals are estimated. Using

[12]**Cox, Biometrika(1975**): Partial likelihood) 62 (2):Journal of the Royal Statistical Society, Series B, 62 (2):269-276

[13]**Kalbfleisch, J. D. & Prentice, R. L. (1973).** Marginal likelihoods based on X-Cox's regression and life model. Biometrika 60, 267-278.

IBM SPSS version19.0[14]to apply the life table method for table (1) data in Annex (A1) as follows: The approach is to divide the period of observation into a series of time intervals and estimate survival for each interval.
The intervals need not be of equal length, although they frequently are. Cancer registries often record survival time only in completed years, rather than months or days, so it is common to construct life tables using annual intervals.

**Table (1): The Construction of life table (10-years survival time)**

| Start Time(i) | Number entering this interval ($l_i$) | Number withdrawn from this interval ($w_i$) | Number exposed to risk ($l_i$-$w_i$/2) | Number of terminal events | Proportion terminating | Proportion surviving | Cum proportion surviving at End | S.E of Cum surviving | Probability density | S.E of probability density | Hazard Rate | S.E of Hazard Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 32 | 3 | 30.500 | 0 | .00 | 1.00 | 1.00 | .00 | .000 | .000 | .00 | .00 |
| 20 | 29 | 2 | 28.000 | 3 | .11 | .89 | .89 | .06 | .005 | .003 | .01 | .00 |
| 40 | 24 | 2 | 23.000 | 1 | .04 | .96 | .85 | .07 | .002 | .002 | .00 | .00 |
| 60 | 21 | 2 | 20.000 | 7 | .35 | .65 | .56 | .10 | .015 | .005 | .02 | .01 |
| 80 | 12 | 0 | 12.000 | 9 | .75 | .25 | .14 | .07 | .021 | .005 | .06 | .02 |
| 100 | 3 | 0 | 3.000 | 2 | .67 | .33 | .05 | .05 | .005 | .003 | .05 | .03 |
| 120 | 1 | 0 | 1.000 | 1 | 1.00 | .00 | .00 | .00 | .000 | .000 | .00 | .00 |
| Median survival time (82.64) month | | | | | | | | | | | | |

If survival time is known only in completed years, the exact number of person-months at risk cannot be calculated. When constructing life tables from such data, it is assumed that censoring occurs uniformly throughout the interval, so each of the censored patients is assumed to be at risk for half of the interval (an assumption known as the actuarial assumption). The median survival time from Life Table in this study is estimated as 82.64 month rank in the time interval 80 up to 90 month, the total number of female entering this interval is 12, no patients were withdrawn from this interval by the event of death. Therefore, the numbers exposed to risk of death in this interval is equal 12 with 9 terminal events so the proportion of terminating is 0.75; the proportion surviving through this interval is (1-0.75=0.25) with hazard rate equal to 0.06. The proportion surviving from general 10-year survival study was decreased gradually with the time interval and the hazard rate increase shortly with the time interval.

### 3.2. Estimating Survival Using Kaplan-Meier Model

The 10-year Survival analysis (1425-1435):The simplest measure of patient survival is the proportion of patients who survive at least *t* years following diagnosis. This is known as the *observed survival rate*. For example, the 1-year observed survival rate is estimated as the proportion of patients who are alive one year subsequent to diagnosis.Among the patients who survived the first year, we may be interested in the proportion who survives a further year.

**Table (2): Kaplan Meier Survival Analysis Results**

| | No. of Cases | Median survival time/month | Mean age at diagnosis | Log rank (p-value) |
|---|---|---|---|---|
| Overall | 32 | 84 | 45.6 | - |
| Menopausal Status | | | | |
| Pre | 25 | 76 | 40.4 | |
| Post | 7 | 60 | 64.3 | |
| Marital Status | | | | 0.176 |
| Married | 27 | 96 | 48.9 | |
| Not Married | 5 | 60 | 29.4 | |
| Weight | | | | |
| Over weight | 18 | 84 | 48 | |
| Normal | 14 | 96 | 40 | |
| Place of Residence | | | | |
| Umlj | 13 | 60 | 46.9 | |
| Outside Umlj | 19 | 84 | 44.7 | |
| Family History | | | | |
| Yes | 7 | 96 | 47.0 | 0.336 |
| No | 20 | 84 | 42.2 | |
| Don't Know | 5 | 60 | 45.9 | |
| Having Children | | | | |
| Yes | 21 | 96 | 53.5 | |
| No | 11 | 60 | 30.5 | |

[14]IBM SPSS version 19.0: IBM software for the Statistical Packaged for Social Sciences Edition 20

This estimate is known as a *conditional survival rate*, since it is conditional on surviving one year. Conditional survival rates are useful for showing how the mortality due to the disease varies. Survival rates calculated from diagnosis are known as *cumulative survival rates*, although the cumulative is often omitted.

**3.3. The Survival Analysis using Cox's Regression Models**

The likelihood-ratio, Wald, and score chi-square statistics are asymptotically equivalent tests of the omnibus null hypothesis that all of the $\beta$*'s* are zero. In this instance, the test statistics are in close agreement, and the hypothesis is soundly rejected.The initial estimation in based on log likelihood is114.568.The Cox regression model results shows that place of resident, age , marital status and weight have a negative regression coefficients which indicates that they reduce the hazard of breast cancer by 4.9%, 45.1 %, 126.4% and 61.3% respectively. Family history, fertility and menopausal status have a positive regression coefficient which indicates that greater value of these variables increase the hazard of breast cancer by 36.6%, 132.4% and 155.4% respectively.

**Table (3): Results from Cox regression Model**

| | ***B*** | ***SE*** | ***Wald*** | ***df*** | ***Sig.*** | ***Exp(B)*** | ***95.0% CI for Exp(B)*** | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | ***Lower*** | ***Upper*** |
| ***Age*** | -.049 | .060 | .648 | 1 | .421 | .953 | .846 | 1.072 |
| ***Resident*** | -.451 | .584 | .598 | 1 | .439 | .637 | .203 | 2.000 |
| ***Marital*** | -12.639 | 97.54 | .017 | 1 | .897 | .000 | .000 | 3.5E77 |
| ***Fertility*** | 13.248 | 97.55 | .018 | 1 | .892 | 567112 | .000 | 6.1E88 |
| ***Weight*** | -.613 | .543 | 1.276 | 1 | .259 | .542 | .187 | 1.569 |
| ***History*** | .366 | .559 | .427 | 1 | .513 | 1.441 | .482 | 4.312 |
| ***Menopause*** | 1.554 | 1.460 | 1.134 | 1 | .287 | 4.732 | .271 | 82.692 |

The overall goodness of this model was calculated as follows;

$$R^{2}_{M} = 1 - \exp\left[\frac{2}{32}(92.651 - 92.462)\right] = 0.00194$$

This results shows that the above model was perfectly adequate model to fit the breast cancer survival data because it has a very small value of $R^2$. The relative hazard of this model is $\frac{h(t)}{h(0)}$ = (1.290114) and the log relative hazard is $\ln \frac{h(t)}{h(0)}$ =-.0.2547.

**3.3.2. Cox Regression Diagnostics**

In order to chick the Cox regression diagnostics the researcher uses all the data as an interval censoring study. First wee checks for non-proportional hazard assumption using Schoenfeld, **(1982)**[15] residual, then repeats the estimation by fitting with strata here she used a year at diagnosis as strata variable. Also she is detecting Influential Observationusing *dfbeta* and detecting non-linearity using Martingale residual. As for a linear or generalized-linear model, it is important to determine whether a fitted Cox-regression model adequately represents the data.

**3.3.3. Checking for Non-Proportional Hazards :**A departure from proportional hazards occurs when regression coefficients are dependent on time that is, when time interacts with one or more covariates. Tests and graphical diagnostics for interactions between covariates and time may be based on the scaled Schoenfeld residuals from the Cox model. The formula and rationale for the scaled Schoenfeld residuals are complicated the details are available in (Hosmer and Lemeshow, 1999[16], or Therneau and Grambsch, 2000)[17].

[15]**Schoenfeld, D (1982):** Partial residuals for the proportional hazards model, Biometrika 69, 551{55}.

[16]**Hosmer, D. W. & Lemeshow, S. (1999).** Applied survival analysis. Regression modeling of time to event data. NY: John Wiley & Sons widely used.

[17]**Therneau, T.M. and Grambsch, P.M.(2000):** Modeling survival data: extending the Cox model. Springer.

The scaled Schoenfeld residuals comprise a matrix, with one row for each record in the data set to which the model was fit and one column for each covariate. Plotting scaled Schoenfeld residuals against time, or a suitable transformation of time, reveals un-modeled interactions between covariates and time. One choice is to use the Kaplan-Meier estimate of the survival function to transform time.
A systematic tendency of the scaled Schoenfeld residuals to rise or fall more or less linearly with (transformed) time suggests entering a linear-by-linear interaction (i.e., the simple product) between the covariate and time into the model.
A test for non-proportional hazards can be based on the estimated correlation between the scaled Schoenfeld residuals and (transformed) time. This test can be performed on a per-covariate basis and also cumulated across covariates. In this study the initial estimated of Cox model for the entire data ad follows;

**Table (4): Cox Regression Results for Entire Data.**

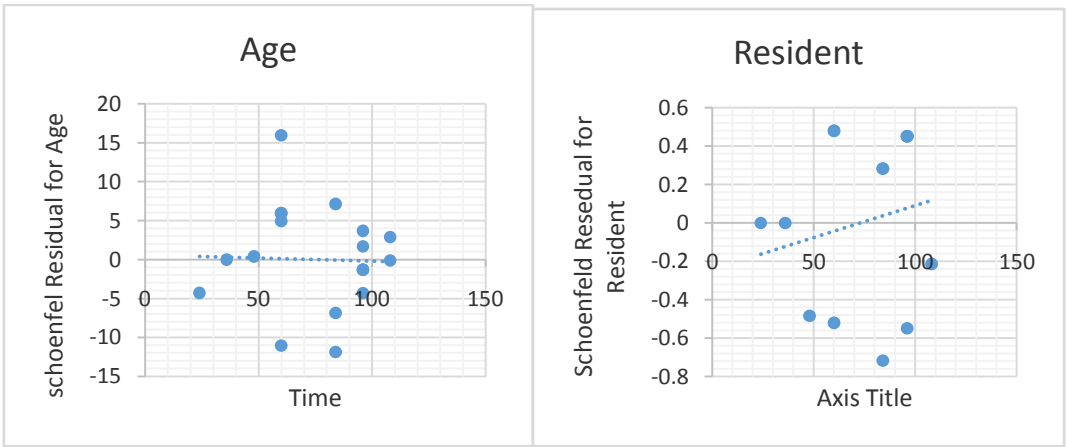
| T | $\hat{\rho}$ | SE | $\chi^2$ | **P > z** | Lower | Upper |
|---|---|---|---|---|---|---|
| Age | -.049 | .060 | .648 | .421 | .846 | 1.072 |
| Resident | -.451 | .584 | .598 | .439 | .203 | 2.000 |
| Marital | -12.639 | 97.6 | .017 | .897 | .000 | 3.50 |
| fertility | 13.248 | 97.6 | .018 | .892 | .000 | 6.169 |
| Weight | -.613 | .543 | 1.276 | .259 | .187 | 1.569 |
| History | .366 | .559 | .427 | .513 | .482 | 4.312 |
| Menopause | 1.554 | 1.460 | 1.134 | .287 | .271 | 82.692 |

It is conceivable that a variable with a non-significant coefficient in the initial model nevertheless interacts significantly with time, starting with the original model we present in the following table a test for non-proportional hazard.

**Table (5): Tests for non-proportional hazards in this model are as follows**

| T | $\hat{\rho}$ | SE | $\chi^2$ | **P > z** | Lower | Upper |
|---|---|---|---|---|---|---|
| Age | -.079 | .084 | .878 | .349 | .784 | 1.090 |
| Resident | -.444 | .614 | .523 | .470 | .193 | 2.137 |
| Marital | -1.185 | 1.265 | .877 | .349 | .026 | 3.649 |
| Weight | -.091 | .604 | .023 | .881 | .279 | 2.986 |
| History | .529 | .560 | .894 | .344 | .567 | 5.084 |
| Menopause | .751 | 1.846 | .165 | .684 | .057 | 78.934 |

$\hat{\rho}$ is the estimated correlation between the scaled Schoenfeld residuals and transformed time. Under the null hypothesis of proportional hazards, each $\chi^2$ test statistic is distributed as $\chi^2$ with the indicated degrees of freedom.Thus, the tests for all covariate in the original model and in the new test were statistically not significant as is the global test for non-proportional hazards.



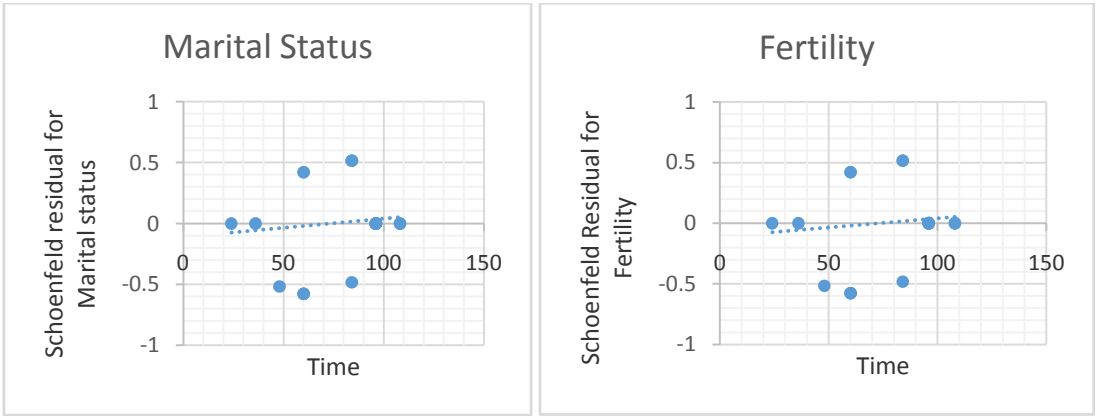



**Figure (1): Plots of scaled Schoenfeld residuals against transformed time for the covariates Age, residence, marital status and Fertility**.

Figure (1) shows plots of scaled Schoenfeld residuals vs. the covariates race, residence, menopause and age in the 1st study. The line on each plot is a smoothing SP-line (a method of nonparametric regression); the broken lines give a point-wise 95-percent confidence envelope around this fit. The tendency for the effect of all variables is to rise with time.

The effect of Age on the hazard of re-offending is initially positive, but this effect decrease with time and eventually becomes negative (by 60 months). The effect of residence is initially negative, but eventually becomes positive (by 60 months). The effect of marital status is initially negative but eventually it becomes positive (by 70 months). The effect of fertility is initially negative, but this effect increase with time and became positive (by 80 month). The re-specified model shows an evidence of non-proportional hazards; the global test $\chi^2 = 39.7$ and p-value =0.000.

### 3.3.4. Fitting by Strata

An alternative to incorporating interactions with time is to divide the data into strata based on the values of one or more covariates. Each stratum may have a different baseline hazard function, but the regression coefficients in the Cox model are assumed to be constant across strata. An advantage of this approach is that we do not have to assume a particular form of interaction between the stratifying covariates and time. There are a couple of disadvantages, however: The stratifying covariates disappear from the linear predictor into the baseline hazard functions. Stratification is therefore most attractive when we are not really interested in the effects of the stratifying covariates, but wish simply to control for them. When the stratifying covariates take on many different (combinations of) values, stratification which divides the data into groups is not practical. We can, however, recode a stratifying variable into a small number of relatively homogeneous categories. In this study we divided age into two categories: those 54 years old or less; those 54 years old. Race and residence are statistically not significant with age, but stage at diagnosis and parity are statistically significant with age at diagnosis. Fitting by strata for the covariates stratified be age.

**Table (6): Fitting the stratified Cox model to the data**

| | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| Age | -.049 | .060 | .663 | 1 | .415 | .952 | .846 | 1.071 |
| Resident | -.462 | .585 | .624 | 1 | .430 | .630 | .200 | 1.983 |
| Marital | -12.299 | 93.127 | .017 | 1 | .895 | .000 | .000 | 8.476E7 |
| Fertility | 12.898 | 93.130 | .019 | 1 | .890 | 399571.4 | .000 | 7.481E8 |
| Weight | -.605 | .543 | 1.241 | 1 | .265 | .546 | .188 | 1.583 |
| History | .378 | .563 | .451 | 1 | .502 | 1.459 | .484 | 4.397 |

The initial log likelihood for the estimation in this model is 107.17. The stratified model includes six covariates that affect the hazard of breast cancer significantly. In model age, residence, marital and weight has negative regression coefficient which indicates that they reduced the hazard of breast cancer by 4.9%, 46.2%, 123% and 60.3% respectively, when we assume other covariates constant in the model.

Fertility and Family history estimated to increase the hazard of breast cancer by 128.9% and 37.8% respectively. The new log relative hazard is (-.139) which indicate that the stratified model

### 3.3.5. Detecting Influential Observations

As in linear and generalized linear models, we don't want the results in Cox regression to depend unduly on one or a small number of observations. Approximations to changes in the Cox regression coefficients attendant on deleting individual observations (*dfbeta*), and these changes standardized by coefficient standard errors (*dfbetas*), can be obtained for the Cox model.

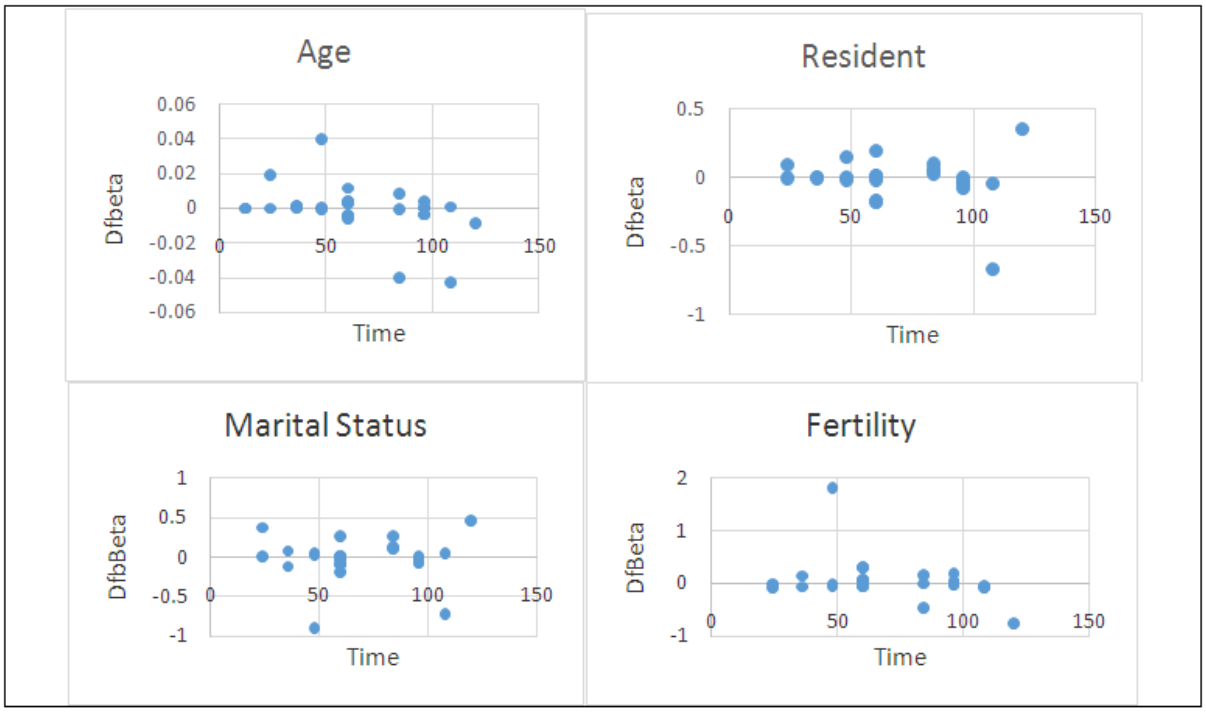



**Figure (2): *dfbeta* for Age, Residence, Marital and Fertility.**

Figure (2) shows index plots of *dfbeta* for the four covariates, Age, residence, marital status and Fertility in the stratified Cox model. All of the *dfbeta* are small relative to the sizes of the corresponding regression coefficients.

### 3.3.6. Detecting Nonlinearity

Other kind of Cox-model residuals, called martingale residuals, are useful for detecting nonlinearity in Cox regression. Plotting residuals against covariates, in a manner analogous to plotting residuals against covariates from a linear model, can reveal nonlinearity in the partial relationship between the log hazard and the covariates. The martingale residuals shown in Fig (2) are slightly skewed. This might be attributed to the single failure outcome feature of the Cox model. Presently, the estimated mode, median, and mean martingales are -0.85, 0, and -0.125, respectively. The estimated measure of skew-ness was approximately 3. We see an indication of a lack of fit of the model to individual observations.

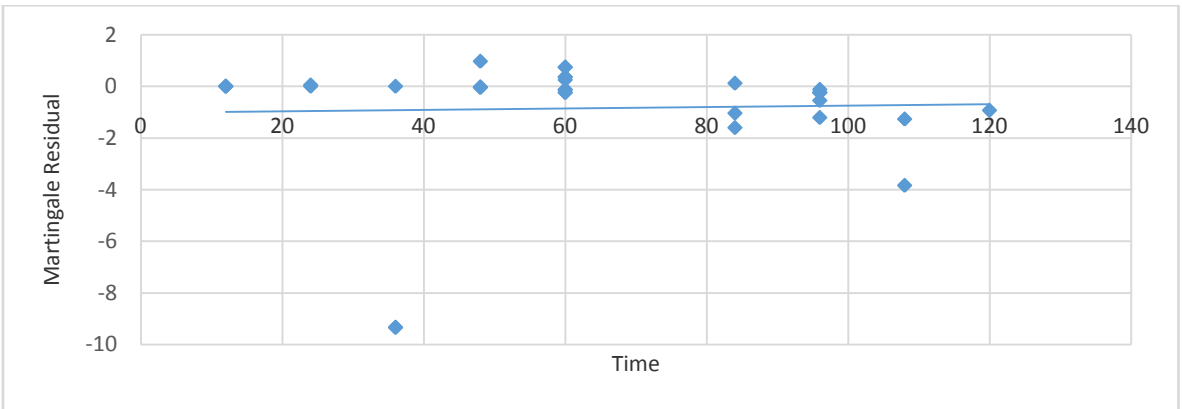



**Figure (3): The Martingale residuals for the model.**

## V. DISCUSSION

The median survival time from Life Table in this study is estimated as 82.64 month rank in the ىلtime interval 80 up to 90 month, the total number of female entering this interval is 12; no patients were withdrawn from this interval by the event of death. Therefore, the numbers exposed to risk of death in this interval is equal 12 with 9 terminal events so the proportion of terminating is 0.75; the proportion surviving through this interval is (1-0.75=0.25) with hazard rate equal to 0.06. The proportion surviving from general 10-year survival study was decreased gradually with the time interval and the hazard rate increase shortly with the time interval. Kaplan Meier model results show that,10-year survival analysis results indicate that the median and mean survival time for the patient in the sample is high; it about 7 year out of 10 years, with mean age at diagnosis 45.59 (3, 72). Also the results indicate that fertile woman are more survived and less diagnosed compare to infertile women. High weight cases were less survived and more diagnosed compare to normal weight cases. Considering the residence of cases, women live in Umluj were more surviving and less diagnosis compare with those live outside Umluj. All these results are confirmed by using the tests of equality. In North America and Europe the incidence in women younger than 40 ranges from 8% to 15%, compared to 28.1% in our experience. It could be argued that our figures, being sample data, reflect a referral bias, but the National Cancer Registry figures for 1994 show that 30% of breast cancer is in those younger than 40 years. Some studies indicate different ages at menarche, weight at menopause, and varying estrogen levels as contributory factors to the difference in incidence between countries. In the U.S., localized breast cancer accounts for 58% of the cases. In our study pre-menopausal women were found to be more survive than post-menopausal women.

The semi-parametric model results (The Cox regression model ) indicates that, age , place of residence, marital status and weight have a negative regression coefficients which indicates that they reduce the hazard of breast cancer by 4.9%, 45.1 %, 126.4% and 61.3% respectively. Family history, fertility and menopausal status have positive regression coefficients which indicate that greater value of these variables increase the hazard of breast cancer by 36.6%, 132.4% and 155.4% respectively. Also results shows that the above model was perfectly adequate model to fit the breast cancer survival data because it has a very small value of $R^2$ (0.00194), with log relative hazard = 0.2547.

The workability of the Cox regression model to non-medical data were confirmed by checking the violation of the assumption of proportional hazard, influential data and nonlinearity in the relationship between the log-hazard and the covariates, and all of them were confirmed.

## VI. OBSTACLES

The lack of clinical information for breast cancer in the area is the main obstacle that face the researcher, after some trails for accessing clinical information from other source, researcher decide to use un-clinical data. The major obstacle that was encountered in this research study was the difficulty in persuading invited students to actually participate in the study, researcher used to conduct an in-depth interview to encourage them to participate in the study. The small sample size (32),led to calculate the life table for the whole sample only, the researcher can't calculate life tables for the rest of covariates in the study.

## VII. CONCLUSIONS

The main goal for this study is to reviews the survival models and examines their workability in non-medical data. To collect the non-medical data, researcher used the approach of sisterhood method of data collection; this approach is a very useful in collecting maternal mortality data. It has been used to calculate maternal mortality ratio in many demographic studies. Information about age, residence, marital status, weight and family history of disease are collected using questionnaire. Application of life table, as an actuarial method, was done. Kaplan-Meier estimator has been used to; show the differences in survival time for each variable in the study. Results from K-M model indicate there is a statistical evidence of difference in the survival for the fertility variable, for the other variables there was no evidence for statistical difference. All these results are confirmed by using the tests of equality. Cox regression model has been applied to test the effect of each covariate in the hazard rate. Results shows that age, place of residence, marital status and weight have a negative effect, while family history, fertility and menopausal status have positive effect on the hazard of breast cancer. Also results shows that the above model was perfectly adequate model to fit the breast cancer survival data. Finally it concluded that the survival models can be applied to non-medical data using primary data (questionnaires).

## VIII. RECOMMENDATIONS

To establish treatment centers of breast cancer in different states in Saudi Arabia especially the peripheral ones, with trained staff who know how to train females for early detection of breast cancer. Facilitate the access of cancer registry data and let it available for researcher, through a partnership of universities and

research centers with the cancer registry center. Activate the roles of civil societies, and research centers in the awareness of breast cancer and early detection. Encourage future researcher to conduct a population-based survival analysis study using primary available data. And further studies to establish survival analysis of breast cancer using parametric approach. Also further studies to establish the actual prevalence and risk factors associated with breast cancer in Saudi Arabia. Develop an indirect survival models to overlap the problem of age miss reporting.

## IX. ACKNOWLEDGEMENT

The authors would like to acknowledge financial support for this work from the Deanship of Scientific Research (DSR), University of Tabuk –Tabuk, Saudi Arabia, under grant no. s,35/139/1435 .

.

## REFERENCES

[1]. **Allison, P.D. (1984).** Event history analysis: regression for longitudinal event data. Beverly Hills, CA: Sagle publication.

[2]. **Anderson, P. K.** and **GILL, R. D. (1982):** Cox's regression model for counting processes: a large sample study. Ann. Statist. 10 1100]1120,

[3]. **Cox, D. R (1972**): Regression models and life tables. Journal of the Royal Statistical Society Series B, 34:187-220.

[4]. **Cox, Biometrika(1975**): Partial likelihood) 62 (2):Journal of the Royal Statistical Society, Series B, 62 (2):269-276

[5]. **Hosmer, D. W. & Lemeshow, S. (1999).** Applied survival analysis. Regression modeling of time to event data. NY: John Wiley & Sons widely used.

[6]. **Kalbfleisch, J. D. & Prentice, R. L. (1973).** Marginal likelihoods based on X-Cox's regression and life model. Biometrika 60, 267-278.

[7]. **Kalbfleisch, J. D. & Prentice, R. L. (1980).** The statistical analysis of failure rate data. NY: John Wiley.

[8]. **Kaplan, E. L. and Meier, P. (1958):** Nonparametric estimation from incompletes observations. Journal of American Statistical Association, 53:457-451.

[9]. **Kardaum, (1993).** Statistical analysis of male larynx cancer patients: A case study. Statistical Nederlandica, 37:103-126.

[10]. **Kosko, B (1992).** Neural networks and fuzzy systems. Dynamical systems approach to machine intelligence. 1st ed. Prentice-Hall International Editions,

[11]. **Lee et al, (1992);** Statistical methods for survival data analysis. 2nd edition. Wiley, New York. 482pp. (comprehensive, no easy reading).

[12]. **Leung, K.M., R.M. Elashoff and A.A. Afifi, (1997).** Censoring issues in survival analysis. Annual Review of Public Health, 18:83-104.

[13]. **NCR, (1992):** Saudi Arabia cancer registry report

[14]. **Prentice, R. L. & Farewell, B. T. (1986).** Relative risk and odds ratio regression. Annual Review of Public Health 7: 335-338.

[15]. **Prentice, R. L. and Cai, J. (1992).** Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* **79**, 495-512.

[16]. **Prentice, R. L., Williams, B.J. and Peterson, A.V. (1981).** On the regression analysis of multivariate failure time da ta. *Biometrika* **68**, 373-379.

[17]. **Schoenfeld, D (1982):** Partial residuals for the proportional hazards model, Biometrika 69, 551{55}.

[18]. **Sedmak, D. D., T. A. Meineke, D. S. Knechtges and J. Anderson, (1989**): Prognostic significance of cytokeratin-positive breast cancer metastases. Modern Pathol, 2: 519-520.

[19]. **Shane, S. & Foo, M. D. (1999):** New firm survival: institutional explanations for new franchisor mortality, Management Science, 45(2), pp. 142-159.

[20]. **Statistical program for social sciences (SPSS, 19).** SPSS advanced models 20 Chicago, IL: Author (www.spss,com).

[21]. **Therneau, T.M. and Grambsch, P.M.(2000):** Modeling survival data: extending the Cox model. Springer.

[22]. **Woolson, R.F., (1981).** Rank test and a one sample log rank test for comparing observed survival data to standard population. Biostatistics, 37:687-696.

**Annex (A1):**

**Table (1): Primary Breast Cancer Data obtained by questionnaire, from Umluj Area**

| Case No. | Age | Date 1 | Sex | Res. | Marital status | Children | Relation | Weight | Family history | Status | Date 2 | Time year |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 47 | 1427 | F | Out | yes | yes | Ants | Obese | Yes | | 1435 | 8 |
| 2 | 62 | 1429 | F | Out | yes | yes | Mother | Non | Yes | Dead | 1434 | 5 |
| 3 | 32 | 1430 | F | In | No | No | Sister | Non | Yes | | 1435 | 5 |
| 4 | 37 | 1432 | F | Out | yes | No | Ants | Non | No | | 1435 | 3 |
| 5 | 49 | 1430 | F | Out | yes | Yes | Mother | Obese | No | | 1435 | 5 |
| 6 | 59 | 1430 | F | Out | yes | Yes | Ants | Obese | No | | 1435 | 5 |
| 7 | 54 | 1428 | F | Out | yes | Yes | Ants | Obese | No | Dead | 1432 | 4 |
| 8 | 52 | 1427 | F | In | yes | Yes | Mother | Obese | No | | 1435 | 8 |
| 9 | 48 | 1426 | F | In | yes | Yes | Mother | Obese | No | | 1435 | 9 |
| 10 | 44 | 1427 | F | Out | yes | Yes | Ants | Non | No | | 1435 | 8 |
| 11 | 67 | 1430 | F | In | yes | Yes | Non | Non | Don't know | Dead | 1434 | 4 |
| 12 | 45 | 1431 | F | In | yes | yes | Non | Non | Don't know | | 1435 | 4 |
| 13 | 3 | 1429 | F | In | No | No | Non | Non | Don't know | Dead | 1430 | 1 |
| 14 | 13 | 1433 | F | Out | No | No | Sister | Non | No | Dead | 1434 | 1 |
| 15 | 51 | 1426 | F | In | yes | yes | Ants | Non | No | | 1435 | 9 |
| 16 | 46 | 1428 | F | In | yes | yes | Other | Obese | No | | 1435 | 7 |
| 17 | 32 | 1428 | F | Out | No | No | Other | Obese | No | | 1435 | 7 |
| 18 | 26 | 1433 | F | Out | yes | No | Sister | Obese | No | | 1435 | 2 |
| 19 | 50 | 1427 | F | Out | yes | yes | Other | Obese | No | | 1435 | 8 |
| 20 | 48 | 1430 | F | In | No | No | Other | Obese | Don't know | | 1435 | 5 |
| 21 | 66 | 1430 | F | In | yes | yes | Grand | Obese | No | Dead | 1433 | 3 |
| 22 | 72 | 1431 | F | In | yes | yes | Mother | Obese | No | Dead | 1432 | 1 |
| 23 | 45 | 1425 | F | Out | yes | yes | Ants | Non | No | | 1435 | 10 |
| 24 | 62 | 1429 | F | out | yes | yes | Grand | Non | No | Dead | 1431 | 2 |
| 25 | 27 | 1428 | F | out | No | No | other | obese | No | | 1435 | 7 |
| 26 | 48 | 1430 | F | In | No | No | Other | Obese | Don't know | | 1435 | 5 |
| 27 | 47 | 1427 | F | Out | yes | yes | Ants | Obese | Yes | | 1435 | 8 |
| 28 | 62 | 1429 | F | Out | yes | yes | Mother | Non | Yes | Dead | 1434 | 5 |
| 29 | 32 | 1430 | F | In | No | No | Sister | Non | Yes | | 1435 | 5 |
| 30 | 37 | 1432 | F | Out | yes | No | Ants | Non | No | | 1435 | 3 |
| 31 | 49 | 1430 | F | Out | yes | Yes | Mother | Obese | No | | 1435 | 5 |
| 32 | 47 | 1427 | F | Out | yes | yes | Ants | Obese | Yes | | 1435 | 8 |