Scientia Africana, Vol. 12 (No.1), June 2013. Pp20-28
© College of Natural and Applied Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

ON ESTIMATION METHODS AND TEST FOR PROPORTIONAL HAZARDS ASSUMPTIONS IN SURVIVAL DATA

K. A. Adeleke[a], A. A. Abiodun[b], R. A. Ipinyomi[b]

[a]*Department of Statistics, Obafemi Awolowo University Ile-Ife Osun State.*
[b]*Department of Statistics, University of Ilorin, Ilorin Kwara State.*
[a] *Corresponding author: aadeleke@oauife.edu.ng*

Received: 18-10-12
Accepted:20-11-12

ABSTRACT

This work compared three estimation methods to handle tied survival time data under the semiparametric Cox proportional hazards model framework (the Exact, Breslow and Efron partial likelihood) and also two parametric proportional hazards models (the Exponential and Weibull) which utilized full likelihood estimation method. These methods were described and applied to two datasets, a clinical dataset on breast cancer patients and a dataset on duration of labour before delivery. We also checked for proportional hazards assumptions on some of the covariates used in the analysis. Using Akaike Information Criterion (AIC) for overall model comparison, Efron method had the least AIC value which is an indication of best performance in handling tied observation, whereas Exponential model with highest AIC performed least. On checking the proportionality assumption for the three categorical variables used in the analysis of cancer data, it was observed that the assumption was valid for absence or presence of Lymph Nodes, whereas it was not valid for progesterone receptor and estrogen receptor.

***Key words*:** *Censored data, Proportional hazards model, Akaike Information Criterion, Parametric model, Survivorship function, Partial likelihood.*

INTRODUCTION

Survival analysis encompasses wide varieties of methods for analyzing the timing of events. The prototypical event is death, which accounts for the name given to these methods. Survival time is defined as the time until failure (Pagano and Gauvreau 1993). In clinical studies, the failure being investigated is often death. Survival analysis describes the methodologies used in biostatistics to quantify and describe survival time and to examine the magnitude of differences in survival time. Survival analysis is also appropriate for many other kinds of events, such as criminal recidivism, divorce, child-bearing, unemployment, and graduation from school. Many studies in statistics deal with deaths or failures of components: the numbers of deaths, the timing of death, and the risks of death to which different classes of individuals are exposed. Many studies on proportional hazard model and the prognostic factors have been published by some authors including Wei *et al*.(1989), Seaman and Bird (2001), Bolard et al. (2001), Young et al. (2001), Meisinger et al. (2002) and Bliwise et al. (2002). Cox proportional hazard model (Cox, 1972) has been a popular semiparametric model often used in the analysis of survival data. Cox has stimulated the interest of many statisticians in his path-breaking work on semiparametric approach to modeling hazard function on a set of

explanatory variables. A large number of papers on this model and related areas have been published since 1972. Often times, survival data contains tied observations, and these need be taken care of during analysis. The ideal method of handling ties is the "Exact method of partial likelihood" under Cox proportional hazard model formulation. This is however computationally intensive (Huang and Liu, 2007). The methods by Breslow (1974) and Efron (1977) are much simpler.

This paper is therefore is concerned mainly with comparing the known methods of estimating survival time data with tied observations under Cox proportional hazard model (i.e Exact, Efron and Breslow partial likelihoods) and two commonly used parametric proportional hazard models) which are the Exponential and Weibull models..

MATERIALS AND METHODS

Proportional hazard models are of two forms, namely, Semi-parametric and parametric proportional hazard model. In proportional model, the ratio of the hazard functions for two individuals with prognostic factors or covariates $x_1 = (x_{11}, x_{21}, \dots, x_{p1})'$ and $x_2 = (x_{12}, x_{22}, \dots, x_{p2})$ is a constant (does not vary with time *t*). This means that the ratio of the risk of failure of two individuals is the same no matter how long they survive (Lee and Wang 2003). The hazard ratio of two individuals with different covariates x_1 and x_2 can be expressed as

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)g(x_1)}{h_0(t)g(x_2)} = \frac{g(x_1)}{g(x_2)} \qquad (2.0)$$

which is constant and independent of time.

Proportionality assumption can be checked by using the graphical methods, time dependent covariate stratification and test based on residuals (Schoenfeld Residual or Cox Snell Residual).

Cox Proportional Hazards Model

The Cox (1972) proportional hazard model is a semiparametric hazard model
which can be given as

$$h(t|x) = h_0(t)\exp\left(\sum_{j=1}^{p} b_j x_j\right) = h_0(t)\exp\,(b'x) \qquad (2.1)$$

where $h_0(t\ )$ is the arbitrary hazard function when all covariates are ignored and b= $(b_1, \dots, b_p)$ are the coefficients of covariates which denote covariate effects and they can be estimated from the data.

From (2.1), the logarithm of the hazard ratio is expressed as

$$log\frac{h_1(t)}{h_0(t)} = b_1x_{1i} + b_2x_{2i} + \cdots + b_px_{pi} = \quad b'x_j \qquad (2.2)$$

To estimate the coefficients, $b_1, \ . \ . \ . \ b_p$, Cox (1972) proposed a partial likelihood function based on a conditional probability of failure, assuming that there are no tied values in the survival times.

Estimation Procedure for Survival Times without Ties

Suppose that *m* of the survival times from *n* individuals are uncensored and distinct, and *n-m* are right-censored. Let $t_1 < t_2 < \ldots < t_m$ be the ordered *m* distinct failure time with corresponding covariates vector x=[x_1,x_2,.,.,.,.x_p]. For a particular failure at time t_i, conditionally on the risk set $R_{(ti)}$ (the set of individuals who have not experienced the event of failure by time t_i), the he partial likelihood is

$$L(b) = \prod_{i=1}^{m} \frac{\exp\left(\sum_{j=1}^{p} b_j x_j\right)}{\sum_{l\in R_{ti}} \exp\left(\sum_{j=1}^{p} b_j x_j\right)} \qquad (2.3)$$

Scientia Africana, Vol. 12 (No.1), June 2013. Pp20-28
© College of Natural and Applied Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

At maximum, the partial likelihood estimator $\hat{b}\ of\ b$ is obtained by taking the derivative of natural log of equation (2.3) and equating to zero.

i.e $$\frac{\partial logL(b)}{\partial b} = 0 \qquad (2.4)$$

The covariance matrix of maximum partial likelihood is

$$\hat{V}(\hat{b}) = \widehat{Cov}(\hat{b}\,) = \left(-\frac{\partial^2 logL(b)}{\partial b \partial b^{'}}\right)^{-1} \qquad (2.5)$$

Estimation Procedure for Survival Times with Ties

Tied survival times are commonly observed in practice and Cox's partial likelihood function was modified to handle ties (Breslow, 1974; Efron, 1977). Exact likelihood is often used to handle tied survival data when the number of observations failing at distinct failure times remains small. Exact likelihood calculates all possible orderings at each time for which more than one event is recorded (Peto, 1972). However, there are situations when the numbers of tied observations at certain failure times may be large enough to make such calculations unwieldy or unfeasible, thus exact likelihood becomes less appropriate. Suppose that we let $R(t_i)$ denote the set of individuals whose event or censored times exceed $\boldsymbol{t_i}$ or whose censored times are equal to $\boldsymbol{t}$, and $\boldsymbol{d_i}$ denote the multiplicity of failures at $\boldsymbol{t_i}$, then the exact likelihood is

$$L(b)_E = \prod_{i=1}^{m}\left\{\int_0^{\infty}\prod_{j=1}^{d_i}\left[1-\exp\left(-\frac{\exp(x_j^{'}b)}{\sum_{l\in R(t_i)}\exp(x_l^{'}b)}\right)\right]\right. \qquad (2.6)$$

Let $z_{u_i^*}^{'}$ be the sum of the vectors x_l over the l th individuals who fail at $\boldsymbol{t_i}$, Breslow (1974) provided an approximation to (2.6) as

$$L(b)_B = \prod_{i=1}^{m}\left[\frac{\exp\left(\sum_{j=1}^{p} z_{u_i^*}^{'}b\right)}{\left[\sum_{l\in R(t_i)}\exp\left(\sum_{j=1}^{p} x'_l b\right)\right]^{d_i}}\right] \qquad (2.7)$$

Efron (1977) also introduced an alternative approximation method to (2.6)

$$L(b)_E = \prod_{i=1}^{m}\left[\frac{\exp\left(\sum_{j=1}^{p} z_{u_i^*}^{'}b\right)}{\prod_{j=1}^{d_i}\left[\sum_{l\in R(t_i)}\exp\left(\sum_{j=1}^{p} x'_l b\right)-[(j-1)/d_i]\sum_{l\in u_i^*}exp\left(\sum_{j=1}^{p} x'_l b\right)\right]}\right] \qquad (2.8)$$

The maximum partial likelihood estimators $(\hat{b})$ can be obtained by applying the Newton-Raphson iterated procedure.

Parametric Proportional Hazard Models

The parametric proportional hazard models follows the same pattern of Cox-proportional hazard model only that the baseline hazard at this time assumed to have followed a particular parametric distribution.

The model is given as:

$$h(t/x) = h_o(t)\ exp(b'x) \qquad (2.9)$$

where $h_o(t)$ follows a particular parametric distribution.

The coefficients are estimated by Maximum likelihood unlike in Cox where we use partial likelihood method. The hazard ratio has the same interpretation as in Cox. Example of parametric proportional hazard model includes Exponential and Weibull. Exponential model is often referred to as purely random failure pattern, model. It is characterized by a constant baseline hazard function λ. The model is given as

$$h(t|x) = \lambda exp(b'x), \qquad (2.10)$$

and the Weibull proportional hazard model can be expressed as

$$h(t|x) = \lambda\rho t^{\rho-1} exp(b'x) \qquad (2.1)$$

where λ and ρ are the scale and shape parameters respectively. In Weibull model, if the intercept and slope are roughly estimated as $\log(\lambda)$ and ρ and the lines are parallel, then the proportional hazard model is valid. The Weibull tends to exponential distribution if the shape parameter ρ equals 1. However, if the hazard function increases or decreases monotonically with increasing survival time, then a Weibull distribution could be considered.

Model Comparison

We used Akaike Information Criterion (AIC) for model comparison. It is given by $-2\log L(\hat{b}) + 2pd$, where $\log L(\hat{b})$ is the loglikelihood and pd is the number of effective parameters. Model with smaller AIC value is usually considered a better model.

Application

The methods discussed in this study are applied to two survival time datasets containing tied observations.

Data 1: Data on Breast Cancer Patients

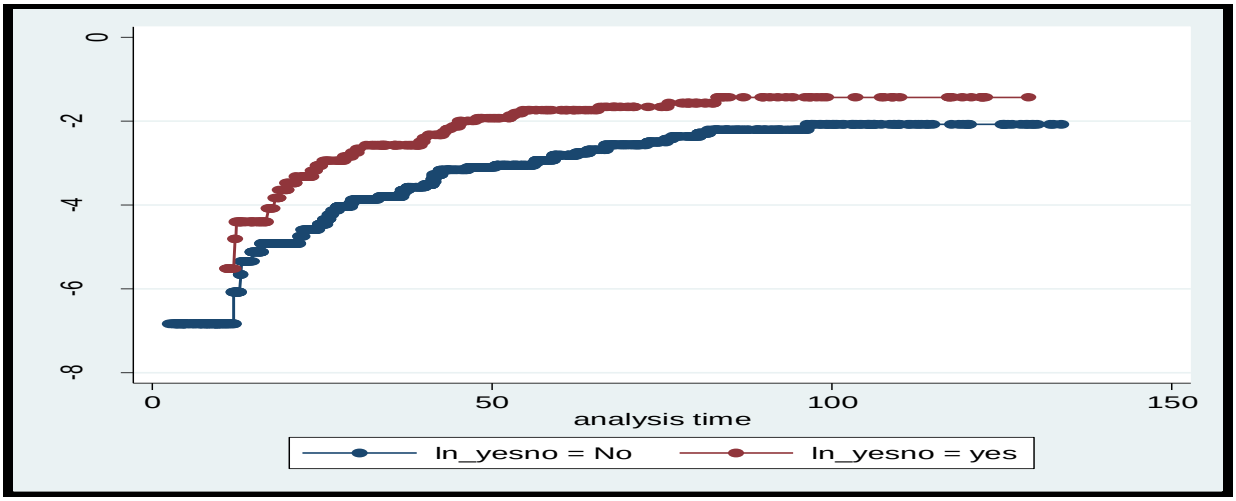
These data are clinical data collected on breast cancer patients (See SPSS 17 data). One thousand, two hundred and seven (1207) patients were listed in the study. During the period of study, 72 of the patients died (failed) due to breast cancer. The event time T of interest is breast cancer free time, which is defined as time in weeks from diagnosis till death. The covariates in the data are: age, pathological size, positive axillary lymph nodes, histological grade, estrogen receptor status (er: 0= negative, 1=positive, 2=unknown), progesterone receptor status (pr: 0= negative, 1=positive, 2=unknown), Time in weeks (Time) and Absence or presence of Lymph Nodes (ln_yesno: 0=No, 1= yes). At the first stage, Proportional Hazards (PH) assumption was checked on the dichotomous/categorical variables estrogen receptor status, progesterone receptor status and absence or presence of lymph nodes. It is observed that proportional hazards assumption holds for only Absence or Presence of Lymph Nodes (Fig, 2A) as the two curves are parallel over time..Figures 2 (b and c) however, show violation of proportional hazard assumption for progesterone receptor and estrogen receptor. As observed, the curves cross at the 10^{th} months as well as at about 10^{th}, 40^{th}, 75^{th} and 90^{th} months for estrogen receptor and 10^{th}, 60^{th} and 85^{th} moths for progesterone receptor. These two covariates are therefore not included in the proportional hazards models fitting. The next stage involved fitting Cox Proportional Hazard Model under the Exact, Breslow and Efron partial likelihoods..

Data 2: Data on Duration of Labour before Child Delivery

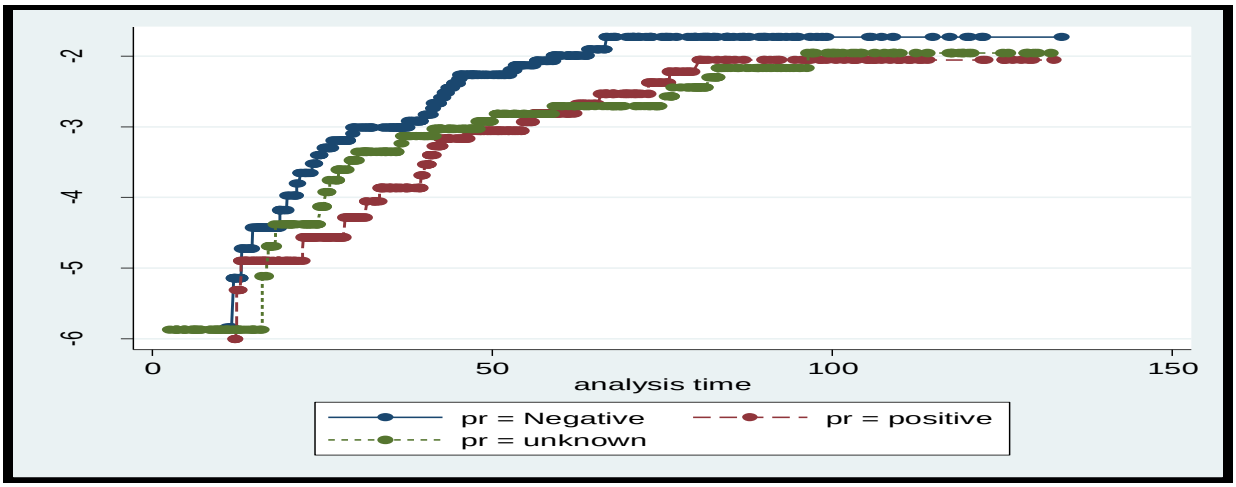
The data on duration of Labour before child delivery for two hundred and ninety women were collected from Federal Medical Centre Lokoja (FMC) Lokoja, Kogi State. The survival time is the time from the onset of labour to the time of delivery of baby (babies), recorded in Hours. Some covariates thought to be associated with labour duration were also collected. These include: age, occupation and religion of the woman under labour. Others are parity (number of previous births), birth type (single or multiple birth), birth weight and sex of the baby. Those who had not delivered as at the time of data collection and those who delivered through caesarian operation were right censored. Since these data were recorded to the nearest hours, there were a number of tied observations.

Scientia Africana, Vol. 12 (No.1), June 2013. Pp20-28
© College of Natural and Applied Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

(a) PH test for lymph nodes status



(b) PH test for progesterone receptor



(c) PH test for estrogen receptor

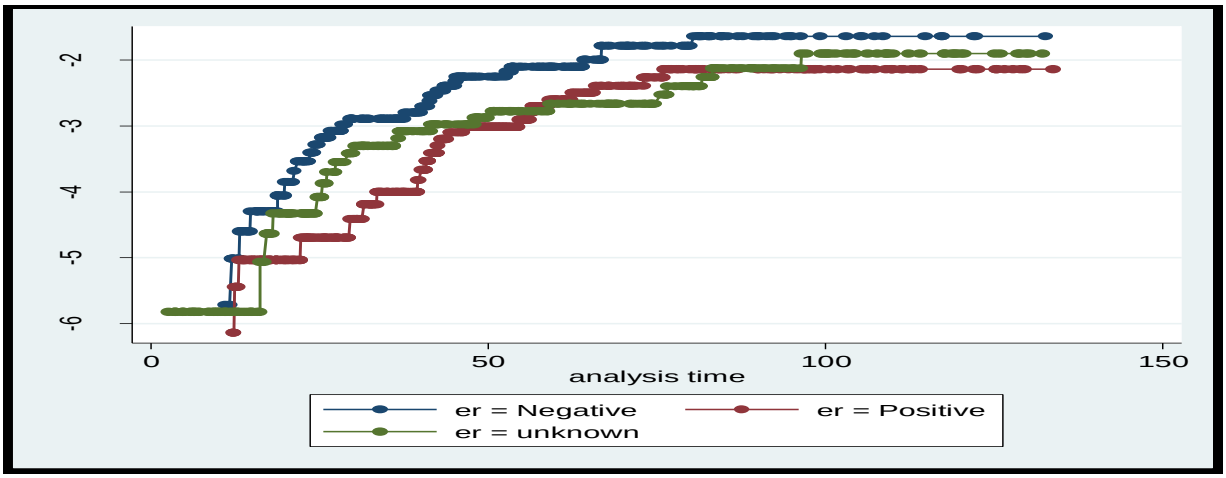


Figure 1: Test for Proportional Hazard assumption

Proportional hazard assumption was found to hold (results not presented) for three covariates including: Mother's age (0 if <25 years; 1 if 25-34 years and 2 if ≥ 35 years), Baby weight (0 if < 3 kg and 1if ≥3 kg) and Birth type (0 if single birth and 1 if multiple birth).The methods for handling ties earlier discussed (the proportional hazards model under Exact, Efron and Breslow) were then applied at the first stage and the results were also similar as in the breast cancer dataset and discussions were also based on the results for Cox model under Efron and the Parametric Exponential and Weibull models.

RESULTS

Results of the Brest Cancer Data

The estimated hazard ratios $\exp(\hat{\beta})$ of the breast cancer data are presented in Table 1. As observed, all estimates based on Breslow and Efron likelihood are identical to those based on exact likelihood, but Efron gave results that are much closer to the exact results in terms of estimated regression coefficients and AIC than Breslow. Also since it is the least computationll intensive, we fitted and compared it with parametric models in the second stage. The parametric models involved under proportional hazards framework are Exponential and Weibull models. The estimated regression coefficients, P-values and the standard errors for the breast cancer data are presented in Table 2. The AIC values are also presented for model comparison. From the table, it is observed that the results are similar for semi parametric Efron and the two parametric models (Exponential and Weibull). As observed, only Positive axilliary lymph nodes have significant influence on the risk of breast cancer.

Table 1: Hazard Ratios for Exact, Breslow and Efron likelihood

	Semi-parametric PH (Cox) Models		
Variable	Exact	Breslow	Efron
Age	0.9865	0.9865	0.9865
Pathsize	1.0012	1.0012	1.0012
Pos_lynode	1.0910	1.0909	1.0910
Histgrad	1.1355	1.1354	1.1355
Presence	1.5249	1.5252	1.5249

Table 2: Regression coefficients P-values, Standard errors and the AIC for Efron Partial likelihood, Exponential and Weibull models

	Cox Model (Efron)		Exponential Model		Weibull Model	
Variable	Coef. (P>z)	Std. Err.	Coef. (P>z)	Std. Err.	Coef. (P>z)	Std. Err.
Age	0.0135 (0.157)	0.0094	0.01509 (0.113)	0.0096	0.0133 (0.165)	0.0095
Pathsize	0.0012 (0.776)	0.0043	0.0013 (0.749)	0.0047	0.0010 (0.804)	0.0044
Pos_lynode	0.0871 (0.001)	0.0286	0.0832 (0.002)	0.0289	0.0926 (0.001)	0.0286
Histgrad	0.1271 (0.347)	0.1510	0.1201 (0.374)	0.1524	0.1173 (0.382)	0.1511
Presence	0.4221 (0.146)	0.4423	0.4170 (0.153)	0.4426	0.3870 (0.183)	0.4423
AIC	559.98		699.36		578.02	

Scientia Africana, Vol. 12 (No.1), June 2013. Pp20-28
© College of Natural and Applied Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Table 3: Results of Analysis of Labour Duration Data

Variable	Cox Model (Efron)		Exponential Model		Weibull Model	
	Coef. (P>z)	Std. Err.	Coef. (P>z)	Std. Err.	Coef. (P>z)	Std. Err.
Age 25-34	0.1262 (0.026)	0.1500	0.1185 (0.031)	0.1601	0.0171 (0.024)	0.1601
Age ≥ 35	0.1833 (0.210)	0.2382	0.2710 (0.119)	0.2539	0.2473 (0.131)	0.2430
Weight	-0.2107 (0.016)	0.1080	-0.1982 (0.013)	0.1388	-0.1864 (0.018)	0.1395
Birthtype	0.1584 (0.061)	0.3545	0.1595(0.059)	3700	0.1430 (0.043)	0.3693
AIC	1539.92		1647.67		1554.78	

The hazard ratios for Efron, Exponential and Weibull models are exp(0.0871)=1.0913 (P-value=0.001), exp(0.0832)=1.0868 (P-value=0.002) and exp(0.0926)= 1.0971 (P-value=0.001) respectively. This implies that after adjusting for all other covariates, for every unit increase in positive axillary lymph nodes, the risk of death due to breast cancer increases by 9.1% for Efron, 8.7% for Exponential and 9.7% for Weibull. The presence of lymph node (after adjusting for other covariates) with hazard ratio exp(0.4221)= 1.5249 (P-value=0.146) under Efron method shows that the risk of dying due to breast cancer by those with lymph nodes is 1.5 times those without lymph nodes. This is however not significant. The results are similarly interpreted for Exponential and Weibull models. Comparing Efron, Exponential and Weibull, it is discovered that Efron has the smallest standard errors for all estimated regression coefficients, showing best performance. This is followed by the Weibull model whereas Exponential model has largest standard errors showing the worst performance. Overall model comparison is done using Akaike Informatioin Criterion (AIC). As observed, Efron likelihood has the least AIC value (559.98) which shows the best performance in handling tied observation, whereas Exponential model with AIC of 699.36 performed worst.

Results of the labour duration data

Table 3 presents the estimated regression coefficients, the standard errors and the P-values as well as the DIC for comparing the models. It is observed that the regression coefficients for the parametric models are not remarkably different from that of Efron likelihood. Since the main objective of the study is to compare estimation methods, we do not discuss the influence of the covariates on the hazard function. However, as observed from the table, the standard errors for each covariate effects are generally least for Efron and highest for Exponential model. Also from the AIC values, best performance are observed from Cox model under Efron likelihood (DIC=1539.82), followed by Weibull model (DIC=1554.78) and worst for Exponential model (DIC=1647.67).

DISCUSSIONS

This study compared estimation methods using parametric and semiparametric models under proportional hazards framework when there are tied observations in survival data. Three parameter estimation methods commonly used to handle survival data with ties were considered under Cox proportional hazard model framework. These are Exact, Breslow and Efron likelihood. Exponential and Weibull models were also considered from the parametric model formulation. We analyzed a clinical dataset on

breast cancer patients from SPSS 17 and a dataset on labour duration before child delivery, collected from Federal Medical Centre, Lokoja. For the breast cancer datasei, regression estimates were similar under all the estimation methods. Only Positive axilliary lymph nodes had significant influence on the risk of breast cancer. As observed, increase in positive axillary lymph nodes increased the risk of death due to breast cancer. The standard errors of the regression coefficients were smallest under Efron partial likelihood method of estimation while Exponential model had the largest standard errors for the regression coefficients. This showed best performance for Efron partial likelihood estimation method and worst performance for Exponential model. Overall model comparison was done using Akaike Informatioin Criterion (AIC). As observed, Efron likelihood had the least AIC value which shows the best performance in handling tied observation, whereas Exponential model with highest AIC performed worst. On checking for the proportionality assumption for the three categorical variables used in the analysis, it was observed that the assumption was valid for Absence or presence of Lymph Nodes whereas it was not valid for progesterone receptor and estrogen receptor. The results of analysis for labour duration data was similar to those obtained for breast cancer data. In conclusion, two survival datasets containing tied observations were analyzed in this study, using three proportional hazards models Cox model under semiparametric model framework and also Exponential and Weibull models under parametric model framework. It was observed that Cox model with Efron likelihood performed best and Exponential was worst. A possible extension of this study is to examine these methods under various sample sizes and percentage of censoring.

REFERENCES

Bliwise, D. L., Kutner, N. G., Zhang, R., and Parker, K. P. (2002). Survival by Time of Day of Hemodialysis in an Elderly Cohort. Journal of the American Medical Association, 286(21), 2690—2694.

Bolard, P., Quantin, C. P., Esteve, J., Faivre, J., and Abrahamowicz, M. (2001). Modeling Time-Dependent Hazard Ratios in Relative Survival: Application to Colon Cancer. Journal of Clinical Epidemiology, 54 (10) 986—996.

Breslow, N. E. (1974) Covariance analysis of censored survival data. Biometrika 30 , 89.100.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society. Series B 34 187.220 survival. J Chron Dis; 8:53:457-481.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. J. Am. Statist. Assoc. 72, 557-65

Lee, E.T. and Wang, J.W. (2003). Statistical Methods for survival Data Analysis. 3rd Edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

Huang, X.and Liu, L. (2007). A Joint Frailty Model for Survival and Gap Times Between Recurrent Events. Biometrics, 63(2), 389-397.

Kalbfleisch J. D. and R. L. Prentice (1980), The Statistical Analysis of Failure Time Data, Wiley, New York..

Scientia Africana, Vol. 12 (No.1), June 2013. Pp20-28
© College of Natural and Applied Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Meisinger, C., Thorand, B., Schneider, A., Stieber, J., Doring, A., and Lowel, H. (2002) Sex Differences in Risk Factors for Incident Type 2 Diabetes Mellitus: the MONICA Augsburg Cohort Study. Arch Intern Med, 162, 82—89.

Pagano M, Gauvreau K. (1993). Principles of Biostastics. 1st ed. Belmont, Calif: Wadsworth; 445-468.

Peto, R. (1972). Contribution to discussion of paper by D. R. Cox. J. R. Statist. Soc. B 34, 205-7.

Seaman, S. R., and Bird, S. M. (2001). Proportional Hazards Model for Interval-Censored Failure Times and Time-Dependent Covariates: Application to Hazard of HIV Infection of Injecting Drug Users in Prison. Statistics in Medicine, 20(12),

Wei, L. J., D. Y. Lin, and L. Weissfeld (1989). Regression analysis of multivariate failure time data by modeling marginal distributions. Journal of the American Statistical Association 84, 1065–1073

Young, E. M., and Fors, S. W. (2001). Factors Related to the Eating Habits of Students in Grades 9—12. Journal of School Health, 71, 483—488.