RESEARCH PAPER | Mathematics | Volume : 4 | Issue : 1 | Jan 2014 | ISSN - 2249-555X

A Comparative Study of Life Time Models in the Analysis of Survival Data

KEYWORDS: Bone-Morrow transplantation, Deviance, MLE, Weibull, Log-normal, Proportional Hazard, Kaplan-Meier

V. Vallinayagam	S. Parthasarathy	P. Venkatesan
Department of Mathematics, St.Josephes Engineering College, Chennai	Department of Mathematics, SRM University-Ramapuram Campus, Ramapuram, Chennai -89	Department of Statistics, National Institute for Research in Tuberculosis, Chennai

ABSTRACT *In this paper, an empirical comparison is made between two parametric models namely, Weibull and Log-normal and the semi-parametric Cox model in the analysis of survival data. The bone-morrow transplantation data is used for the comparison. The Lognormal model gave better fit than the other models in terms of deviance.*

1. INTRODUCTION

Survival analysis can be described as a branch of statistics which handles with death in biological organisms and failure in mechanical systems. Survival methods are applied for a vast array of social phenomena including births, marriages, divorces, job terminations, promotions, arrests, migrations, and revolutions [1]. That is used to describe, explain, or predict the occurrence and timing of events. This is called as reliability theory or reliability analysis in engineering [1]. Survival analysis focuses on time to event data. In the most general way, it contains techniques of positive valued random variables, such as, time to death, time to onset (or relapse) of a disease etc,. Some methods of survival analysis are purely descriptive (e.g.,Kaplan-Meier estimation of survival functions), but most applications involve estimation of regression models, which come in a wide variety of forms [3]. These models are typically very similar to linear or logistic regression models, except that the dependent variable is a measure of the timing or rate of event occurrence. Traditionally only a single event is considered in survival analysis. Recurring event or repeated event models relax that assumption. The study of recurring events is relevant in systems reliability and in areas of social sciences and medical research. A key feature of all methods of survival analysis is the ability to handle right censoring, a phenomenon that is almost always present in longitudinal data. Right censoring occurs when some individuals do not experience any events, implying that an event time cannot be measured. Introductory treatments of survival analysis for social scientists can be found in Teachman (1983), Allison (1984, 1995), Tuma and Hannan (1984), Kiefer (1988), Blossfeld and Rohwer (2001), and Box-Steffensmeier and Jones (2004). For a biostatistical point of view, see Collett (2003), Hosmer and Lemeshow (2003), Kleinbaum and Klein (2005), or Klein and Moeschberger (2003).

2. MODELS AND METHODS

In this section we discussed about the parametric and semi-parametric models.

2.1: WEIBULL MODEL

The Weibull distribution is mainly used in connection with lifetime applications. It can be used to represent many distributions as a function of the shape parameter . The density function is

$$f(t) = \frac{b}{T^b} x^{b-1} e^{-\left(\frac{x}{T}\right)^b} \qquad (2.1).$$

Greater significance is attached to the distribution function, however, in practical applications: where t = variable, T = Characteristic life and b= Shape parameter, F(t) = frequency, f(t) = probability density for "moment" t.

$$F(t) = 1 - e^{-\left(\frac{t}{T}\right)^b} \qquad (2.2)$$

2.2: LOG-NORMAL MODEL

The Log-normal distribution is a distribution that is asymmetrical on one side and which exhibits only positive values. Many interrelationships in nature have a positive skew, left steep and right flat distribution. An illustrated explanation of a feature with non-symmetrical distribution is that the feature cannot undershoot or overshoot a certain boundary value. A significant example is the distribution of time values that cannot be negative. Logarithms are used to achieve values with approximately normal distribution particularly in the case of distributions that are limited to the left by the value 0. The creation of a Log-normal distribution may be attributed to the fact that many random variables interact multiplicatively. In contrast, the normal distribution is created by the additive interaction of many random variables. The Log-normal distribution is of particular significance in biology and economics applications. The probability density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2} \qquad (2.3)$$

where x = variable $(x \geq 0)$, μ = mean and σ = Standard deviation.

2.3: PROPORTIONAL HAZARD MODEL

The proportional hazards model was introduced in 1972 by D. R. Cox in order to estimate the effects of different covariates influencing the times to the failures of a system. This model has been employed for different applications in lifetime data analysis. Because of its generality and flexibility, this model was quickly and widely adopted in various fields like biomedical, reliability and economics. Cox's proportional hazard is expressed as

$$f(t, z) = h_0(t)\varphi(\gamma z) \qquad (2.4)$$

where (t) h_0 is the hazard function which is dependent on time only and without influence of covariates.

The positive functional $\varphi(\gamma z)$ is dependent on the effects of different factors, which have multiplicative effect on the hazard function. The proportionality assumptions is.

Volume : 4 | Issue : 1 | Jan 2014 | ISSN - 2249-555X

$$\frac{h(t:z_x)}{h(t:z_y)} = e^{[\gamma(z_x - z_y)]} \qquad (2.5).$$

The hazard at different z values are in constant proportion for all t > 0, hence the name for proportional hazard.

2.4: KAPLAN-MEIER ESTIMATE

Kaplan-Meier estimate is one of the best options to be used to measure the survival fraction (1958). This estimate is also called as "product limit estimate". It involves computing of probabilities of occurrence of event at a certain point of time.

Kaplan-Meier method is a nonparametric approach for survival analysis. It incorporates information from all of the observations, both censored and uncensored by considering survival to any point in time as a series of steps defined by the observed survival and censored times (Hosmer and Lemeshow, 1999) [18].

The survival probability at any particular time is calculated by the formula given below:

$$S_t = \frac{Ns - Nd}{Ns} \qquad (2.6)$$

where Ns = Number of patients living at the start,
Nd = Number of patients died.

where = Number of patients living at the start, = Number of patients died.

For each time interval, survival probability is calculated as the number of patients surviving divided by the number of patients at risk. Patients who have died, dropped out, or move out are not counted as "at risk" i.e., patients who are lost are considered "censored" and are not counted in the denominator. Total probability of survival till that time interval is calculated by multiplying all the probabilities of survival at all time intervals preceding that time (by applying law of multiplication of probability to calculate cumulative probability). Although the probability calculated at any given interval is not very accurate because of the small number of events, the overall probability of surviving to each point is more accurate. There are three important SAS procedures available for analyzing survival data: LIFEREG, LIFETEST and PHREG. Procedure LIFEREG is a parametric regression procedure for modeling the distribution of survival time with set of variables. Procedure LIFETEST is a non-parametric procedure for estimating the survivor function, comparing survival curves, and testing the association of survival time with other variables. Procedure PHREG is semi-parametric procedure that fits the proportional hazard model.

3. DATA BASE

In this section, we have considered the Bone-Morrow transplantation data for empirical comparison. The SAS (Statistical Analysis Software) package was used for calculation [2]. The bone-morrow transplantation data involves 137. The following variables are considered for modeling whose descriptions are given in the table 1.

Table 1: List of variable names

Age-pt	Patient age in years
Age-don	Donors age in years
Sex-pt	Patient sex (1-Male, 0-Female)
Sex-don	Donors sex (1-Male, 0-Female)
Pat-cmv	Patients CMV status (1-Positive, 0-Negative)
Don-cmv	Donors CMV status (1-Positive, 0-Negative)
FAB	It is a way of classification rule
Hosp	Hospital name (1-The Ohio state Uviversity, 2-Alferd, 3St.Vincent, 4-Hahnemann)
MTX	It is a modified classification of FAB
Acut-indi	Acute GVHD indicator
Chro-indi	Chronic indicator
Plate-indi	Platelet recovery indicator
Time	Time t_0

Courtesy: Survival Analysis by John P. Klein and L. Moeschbeger

Table 2: Parameter Estimates of the Models

Models	Weibull			Log-normal			Cox Proportional		
Variable	Coefficient	Std Error	Sig.	Coefficient	Std Error	Sig.	Coefficient	Std Error	Sig.
Age-pt	-.011	.028	.693	-.010	.026	.693	-.001	.021	.947
Age-don	-.244	.026	.360	-.009	.023	.693	.043	.021	.038
Sex-pt	.136	.349	.696	.270	.311	.385	-.099	.262	.706
Sex-don	-.026	.337	.937	-.189	.319	.553	.034	.252	.894
Pat-cmv	.034	.348	.922	.114	.335	.734	-.078	.257	.760
Don-cmv	.279	.353	.429	.284	.333	.393	-.246	.263	.349
FAB	-1.070	.344	.002	-1.013	.328	.002	.654	.254	.010
Hosp	.606	.200	.003	.450	.155	.004	-.566	.164	.001
MTX	-1.049	.444	.018	-1.042	.394	.008	.969	.348	.005
Acut-indi	-.500	.429	.243	-.334	.399	.402	.544	.324	.093
Chro-indi	1.219	.341	0.0	1.237	.319	0.0	-.974	.258	0.0
Plate-indi	1.974	.448	0.0	2.080	.464	0.0	-1.351	.334	0.0
Deviance	407.18			389.58			665.49		

Volume : 4 | Issue : 1 | Jan 2014 | ISSN - 2249-555X

Table 3: The Mean and Median Estimates

		Mean		Median	
Variables		Estimate	Std Error	Estimate	Std Error
Hospital	1	1142.09	125.38	522	82.29
	2	480.47	156.37	162	60.36
	3	1029.48	166.74	1279	848.1
	4	1721.33	184.27	---	---
Sex	Female	1018.01	127.93	469	192.07
	Male	1340.63	129.54	677	291.19
FAB	0	1426.87	123.831	1279	629.39
	1	880.6	146.88	431	100.59

4. RESULTS

The Kaplan-Meier curves for the hospital sex and FAB for bone-morrow data are given in Fig 4.1 to 4.3. The deviance of Weibull distribution is 407.18, Log-normal is 389.58 and proportional Hazard is 665.49. If we compare between the two parametric models namely Weibull and Log-normal, Log-normal model is the best fit for this data because the deviance value of Log-normal is less than the deviance of Weibull. If we compare between the parametric and semi-parametric models, the semi-parametric model that is proportional hazard is not fit for this data because the deviance of proportional hazard is higher than the two other models namely Weibull and Log-normal.

We notice that, among the four hospitals, Hahnemann hospital's estimated value is higher than other three hospitals (see figure 4.2). Also we observed that, Male patient's survival time is better than female patient's survival time (see figure4.3).

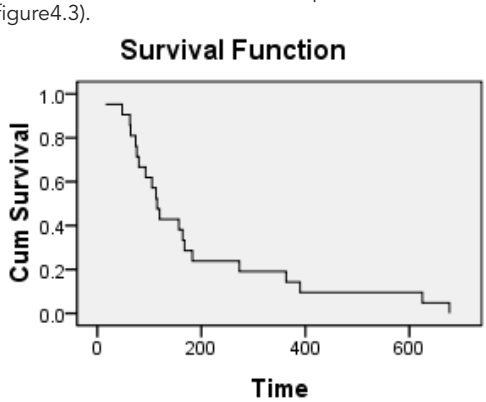

Figure4.1: survival time curve for Hahneman Hospital

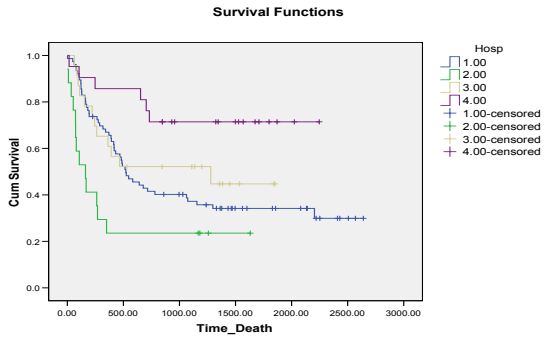

Figure 4.2: Survival time curves for Hospitals

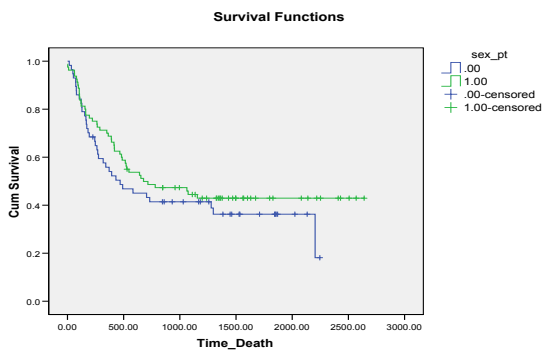

Figure 4.3: Survival time curves for sex

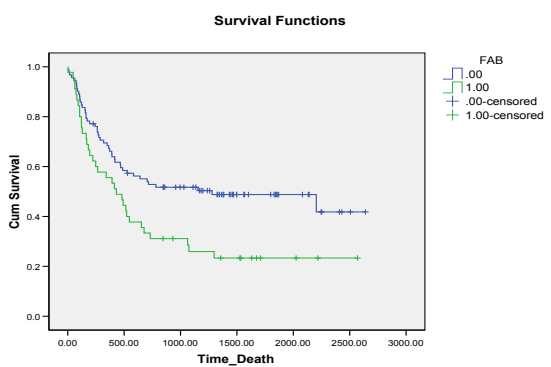

Figure 4.4: Survival time curves for FAB

The regression coefficients for the covariate along with deviance are given in table 2 for the Weibull, lognormal and Cox models

4. CONCLUSION

We have referred an article "Cure Models for Estimating Hospital-Based Breast Cancer Survival" they have documented the utility of a mixture model to estimate the cure fraction and compare it with other approaches [18]. The variables analyzed were tumor stage, postoperative pathology of pathologic tumor residue (TR: negative or positive) and pathologic nodal status (PN: negative or positive). Lognormal kernel's deviance was least when compared with exponential and Weibull distributions. The deviance of the non PH cure model was the least of all the models in this study.

Also we have referred an article "Comparison of Five Survival Models: Breast Cancer Registry Data from Ege University Cancer Research Center"[7], in that article, Gompertz distribution is the best fit distribution based on the lowest AIC value, by comparing Weibull, Gamma, Log-logistic, Log-

Volume : 4 | Issue : 1 | Jan 2014 | ISSN - 2249-555X

normal and Gompertz distributions.

In our article, in the parametric models, Log-normal distribution has the lowest deviance value than the deviance value of Weibull distribution. So we can conclude that, among the parametric models, Log-normal distribution is best fit model for this data. If we compare the deviance of the semi-parametric model proportional hazard with the deviance of the parametric models Weibull and Log-normal distributions, proportional hazard has the highest deviance value so we can conclude that this model is not fit for this data.

From the Kaplan-Meier estimator for the variables Hospital, patient sex and FAB. From the Table 3, we can conclude that, Hahnemann Hospital's patient survival time is more than the other three hospitals. Hahnemann hospital's patient survival time is three times of Alferd hospital's patient survival time. If we compare patient sex wise, we can conclude that male patient's survival time is higher than female patient's survival time. If we compare FAB wise FAB grade 4 or 5 and AML is lower survival time than other FAB classifications survival time.

REFERENCE [1]. Allison P.D, Hardy M, Bryman A (2013), "Event History Analysis". SAGE Publications, 46 | [2]. Allison P.D. (2010), "Survival Analysis Using SAS A Practical Guide." 2nd Edition, Cary NC: SAS institute. | [3]. Breslow, N. E. (1975), "Analysis Of Survival Data Under The Proportional Hhazards Models." International Statistical Review, 43; 45-7, | [4]. Cox, David R. (1972), "Regression models and life-tables." Journal of the Royal Statistical Society, Series B, 34 187-220. | [5]. Curt- Ronniger. (2012), "Reliability Analysis With Weibull." 12th Edition. | [6]. David D. Hanagal and Alok D. Dabade. (2013), "A Comparative Study At Shared Frailty Models For Kidney Infection Data With Generalized Exponential Baseline Distribution." Journal of Data Science, 11, 109-142. | [7]. David G. Kleinbaum and Michael Kline. (2005), "Survival analysis: A Self Learning Text." Springer, 2nd Edition. | [8]. Elvan Akturk Hayat, Aslisurer, Burak Uyar, Omar Dursun, Mehmet N. Orman, Gul Kitapcioglu, MD. (2010), "Comparison Of Five Survival Models: Breast Cancer Registry Data From Ege University Cencer Research Center."Turkiye Klinikleri Journal of Medical Sciences, 30, 1665-74. | [9]. Hosmer DW, Lemeshow S (1999). "Applied Survival Analysis: Regression Modeling of Time to Event Data." New York, Wiley | [10]. John P. Klein and Melvin L. Moeschberger. (2003), "Survival analysis: Techniques For Censored And Truncated Data." Springer 2nd Edition. | [11]. Johnson, Norman L, Kotz, Samuel, Balakrishnan N. (1994), "14: Lognormal distributions, Continuous univariate distributions." Wiley series in Probability and Mathematical Statistics 1011: Applied Probability and Statistics New York, 2nd Edition, John Wiley and Sons. | [12] John Fox. (2002),"Cox proportional-Hazards regression for survival data Appendix to An R and S-plus companion to applied regression.", Will Rios. | [13]. Khard Limpert EC, Werner A. Stahel, and Markus ABBT. (2001), "Log-normal Distributions Across The Sciences: Keys and Clues." Biosciences, 51(5). | [14]. Manish Kumar Goel, Pradeep Khanna, Jugal kishore. (2010), "Understanding Survival Analysis: Kaplan-Meier Estimate." International Journal Of Ayurveda Research, 1; 274-278. | [15]. Mara Tableman and Jong sung Kim. (2003), "Survival analysis using S: Analysis Of Time To Event Data." CRC press. | [16]. Papoulis Pillai. (2002), "Probability Random Variables, And Stochastic Process." Tata Mc Graw-Hill 4th Edition. | [17]. Rockette H, Antle C, Klimbko LA. (1974), "Maximum Liklihood estimation with the Weibull model."Journal Of The American Statistical Association, 69, 246-249. | [18]. Ranganathan Rama, Rajaraman Swaminathan, Perumal Venkatesan. (2010), "Cure Models for Estimating Hospital-Based Breast Cancer Survival." Asian Pacific journal of Cancer Prevention, 11. | [19]. Turnbull B.W and Weiss, L. (1978), "A Likelihood Ratio Statistic For Testing Goodness Of Fit With Randomly Censored Data." Biometrics 34, 364-375. | [20]. Vilijandas Bagdonavicius, Mikhail Nikulin. (2010), "Accelerated Life Models Modeling And Statistical Analysis." CRC press. | [21]. Vaupel, J. W, Manton, K. G, and Stallard. E. (1979), "The Impact Of Heterogeneity In Individual Frailty On The Dynamics Of Mortality." Demography, 16, 439-452. | [22].Wenge Guo.(2011),"Survival Analysis." Springer. | [23]. Klein JP, JD Rizzo, M.J Zhang and N Keiding. (2001), "Statistical Methods for the Analysis and Presentation of the Results of Bone Marrow Transplants. Part 1: unadjusted analysis."Division of Biostatistic, Medical College of Wisconsin, Milwauukee, Wi, USA, 28: 909-15 |