IJETCR Open Access

ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 2 Issue 4 Page No. 83-87

Design and analysis of accelerated failure time model for cancer data

T. Jai Sankar
Department of Statistics, Bharathidasan University, Tiruchirappalli – 620 023, Tamil Nadu, India
tjaisankar@gmail.com

ABSTRACT
In this study suitable survival model traced out for cancer data. The accelerated failure time (AFT) model is more appropriate when the group differences are seen over a shorter timeframe while in the longer term the probability of remaining event free is similar in the two groups. This is consistent with there being a delay in the event occurring in one group compared to the other but no permanent effect. The presence of such a delay is seen in many therapeutic settings and a range of time to event endpoints the AFT model describes a relationship between the survivor functions of any two individuals. In which explanatory variables measured on an individual are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis. This means that the model can be interpreted in terms of the speed of progression of a disease, an interpretation that has immediate intuitive appeal.

Keywords: Survival analysis, Weibull AFT model, progression of a disease.

INTRODUCTION:

Survival analysis is the study of life times. In clinical and other experimental enquiries, measurements on characteristics, which possibly have influence on lifetime, are also obtained. Such characteristics are called concomitant or explanatory variables or simple covariates. For example, many medical charts contain a large number of patient characteristics, i.e. values of covariates, which are possibly related to the prognosis. A statistical analysis is useful in sorting out the ones that are most closely related to the prognosis. This can be done by introducing models, which represent the influence of concomitant variables. Such models, which deal with the relationship between two variables, a dependent or response variable Y and independent variable's or covariate's X, are known as regression models. Survival analysis is a loosely defined statistical term that encompasses a variety of technique for analyzing positive valued random variables. Typically, the value of the random variable is the time to the failure of a physical component (mechanical or electrical) or the time to the death of a biological unit (patient, animal, cell, etc.).

2.0 Survival Analysis in Cancer Research:

In order to describe completely the experience of cancer in a population, it is necessary to know not only its incidence and mortality, but also the survival of cancer patients. There are three main sources of information about survival, the randomized controlled clinical trial which represents the "Gold standard" for the evaluation of forms of treatment; the hospital based study which aims to provide information about the outcome of treatment in particular settings: population based survival which reflects a broader range of cancer control activities. The field of survival analysis offers a wide range of valuable statistical methods and models for use on data that arise in cancer research from any of the above resources. Survival analysis in cancer research dates back to many decades. Only in the past three decades, much progress has been achieved in all spheres of cancer survival analysis encompassing estimation and testing of survival probabilities and eliciting of prognostic factors for survival in different framework.

2.1 Censoring:

In survival analysis the observations are life times, which can be indefinitely long. So quite often the experiment is so designed that the time required for collecting the data is reduced. Two types of the experiment to reduce the time taken for completing the study.

2.2 Type I censoring:

Let t_c be the pre-assigned fixed number which we call the fixed censoring time. Instead of observing T_1, T_2 …T_n

(the random variables of interest) we can only observe Y_1, Y_2, Y_3 ,…,Y_n where,

$$Y_i = \begin{cases} T_i & if \quad T_i \le t_c \\ t_c & if \quad t_c < T_i \end{cases},$$

Notice that, the distribution of Y has positive mass Pr $\{T>t_c\} > 0$ at $y = t_c$.

2.3 Type II censoring

Let r<n be fixed, and let $T_{(1)} < T_{(2)} < T_{(3)} < \ldots < T_{(n)}$ be the order statistics of $T_1, T_2, \ldots T_n$. Observation cases after the r-th failure so we can observe $T_{(1)}, T_{(2)}, \ldots T_{(n)}$. The full ordered observed sample is,

$$Y_{(1)} = T_{(1)}, Y_{(2)} = T_{(2)}, Y_{(3)} = T_{(3)}, \ldots, Y_{(r)} = T_{(r)}, Y_{(r+1)} = T_{(r+1)}, \ldots, Y_{(n)} = T_{(n)}$$

Both type I and type II censoring arise in engineering applications. In such situations there is a batch of transistors or tubes: we put them all on test at t = 0, and record their times of failure. Some transistors may take a long time to burn out, and we will not want to wait the long to end the experiment. Therefore we might stop the experiment at the pre-specified time t_c, in which case we have Type I censoring, or we might not know beforehand what the value of the fixed censoring time is good so we decide to wait until a pre-specified fraction r/n of the transistors has burned out, in which cases we have Type II censoring.

2.4 Random censoring:

Let $C_1, C_2 \ldots C_n$ be i.i.d. each with d.f. G. C_i is the censoring time associated with T_i. We can only observe (Y_1,δ_1), $(Y_2,\delta_2),\ldots, (Y_n, \delta_n)$ where

$$Y_i = \min(T_i, C_i) = T_i \wedge C_i,$$

$$\delta_i = I(T_i \le C_i) = \begin{cases} 1 & if \quad T_i \le C_i \text{ , that is } T_i \text{ is not censored.} \\ 0 & if \quad T_i > C_i \text{ , that is } T_i \text{ is censored.} \end{cases}$$

Notice that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with some d.f. H. also $\delta_1, \delta_2, \delta_3, \ldots, \delta_n$ contain the censoring information, (in Type I and Type II censoring we also were able to observe to which items were censored, but since it was easy to see which ones those were, we don't need to define the δ_i's explicitly.)

Random censoring arises in medical applications with animal studies or clinical trials, patients may enter the study at different times; and then each is treated with one of several therapies. We want to observe their lifetimes, but censoring occurs one of the following ways,

1. Loss to follow-up: the patient may decide to move elsewhere; we never see him again.

2. Drop out: the therapy may have such bad side effects that it is necessary to discontinue the treatment. Or the patient may still be in contact (he hasn't move), but he refuses to continue the treatment.

3. Termination of the study: The Figure 1 illustrates a possible trail.

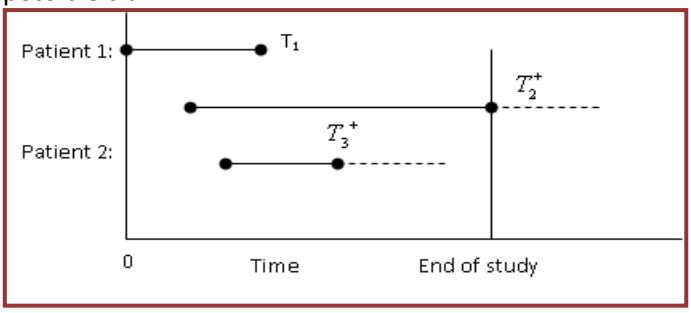

Figure 1. Ways of Random Censoring

Here, patient I entered the study at t=0 and died at T_1 to give an un-censored observation; patient II entered the study, and by the end of the study he has still alive resulting in a censored observation T_2^+; and patient 3 entered the study and was lost to follow-up before the end of the study to give another censored observation T_3^+. With random censoring we will make the crucial assumption that T_i and C_i are independent. Without this assumption few results are available. It seems justified with random entries to the study and randomly occurring losses to follow-up. However, if the reason for dropping out is related to the course of therapy, there may be dependence between T_i and C_i.

2.5 Cancers of Oral Cavity:

Cancers of the oral cavity together are classified based on the UICC norms in this report. The individual sites included in this classification are lip, tongue, (Excluding posterior 1/3 or Base), Gum, Floor of Mouth, Check, Hard Palate and Retro molar Area. Cancers of the Oral Cavity together are ranked within the top five cancers during 2009-2011 between both sexes. They accounted for 6.8% of all male cancers and 5% of all female cancers. The number of cases per year indicated an almost equal sex ratio of 807 females to 1000 males. The average annual ASR was 7 per 100,000 between both sexes during 2009-2011. The cumulative risk of getting cancers of the oral cavity together in their lifetime (0-74 years) was about the same among males (one in 112) and females (one in 120) in Chennai.

The distribution of cases by individual sites and sub-sites comprising the Oral Cavity revealed the following: Among males, Cheek mucosa (36%) was the commonest followed by Anterior 2/3 tongue (28%), Floor of mouth (11%) and Gum (112%). Among females, the order was Cheek mucosa (53%), Anterior 2/3 tongue (16%), Gum (15%), Hard Palate (5%), Lip (4%) and Floor of Mouth (3%). The histological verification of cancer diagnosis was possible in 83%. Squamous cell carcinomas (92%) were the commonest followed by Verrucuous carcinoma (2%), Adenocystic carcinoma (1%) and others (2%). Unspecified category accounted for 3%. The age standardized incidence rate by classified age groups was the highest in the geriatric age group (65+ years) with the peak incidence occurring in the age group of 60-64 years in both sexes.

2.6 Accelerated Failure Time (AFT) Model:

In many clinical trial applications the accelerated failure time model is often a more realistic model than the proportional hazards model in the analysis of time to event data. The proportional hazards model is appropriate when there is a permanent difference between the groups in the longer term in the context of the follow-up period. The accelerated failure time model is more appropriate when the group differences are seen over a shorter timeframe while in the longer term the probability of remaining event free is similar in the two groups. This is consistent with there being a delay in the event occurring in one group compared to the other but no permanent effect. The presence of such a delay is seen in many therapeutic settings and a range of time to event endpoints the AFT model describes a relationship between the survivor functions of any two individuals. Let T_i be a random variable denoting the failure time for the i^{th} subject, and let X_{i1} , X_{i2} , X_{i3}, …, X_{ip} be the values of p covariates for that same subject. The model is then

$$\log T_i = \beta_0 + \beta_1 x_{i1} + ,...,+ \beta_p x_{ip} + \sigma\varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} S_0(.)$$

where ε_i is a random disturbance term, and β_0 , …, β_p, and σ are parameters to be estimated, $S_0(.)$ is a known baseline survival, T_i is are actual survival time sometimes observed, σ is a scale parameter and $x_{i'}$ are fixed px1 vector of covariates. The σ can be omitted, which requires that the variance of ε_i be allowed to be different from 1. But it is simpler to fix the variance of ε_i at 1 and let σ change. All AFT models are named for the distribution of T rather than the distribution of ε or log T. The reason for allowing different distribution assumptions is that they have different implications for the shapes of hazard function.

3.0 Model Formulation:

The accelerated failure time model is a general model for survival data. In which explanatory variables measured on an individual are assumed to act multiplicatively on the time-scale, and so affect the rate at which an individual proceeds along the time axis. This means that the model can be interpreted in terms of the speed of progression of a disease, an interpretation that has immediate intuitive appeal.

The accelerated failure time model becomes

$$h_i(t) = e^{\beta x_i} h_0(e^{\beta x_i} t).$$

According to the general accelerated failure time model, the hazard function of the i'th individual at time t, $h_i(t)$, is then such that

$$h_i(t) = e^{\eta_i} h_0(e^{\eta_i} t)$$

$$\text{where } \eta_i = \beta_i x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

is the linear component of the model, in which x_{ji} is the value of the j^{th} explanatory variable, X_j, j=1,2,….,p, for the i^{th} individual, i=1,2,……,n. As in the proportional hazards model, the baseline hazard function $h_0(t)$ is the hazard of death at time t for an individual for whom the values of the explanatory variables are all equal to zero.

4.0 Weibull AFT Model:

The survival function for this family is given by

$$\overline{F}(t) = e^{-\lambda t^{\gamma}}; t > 0; \lambda, \gamma > 0.$$

This distribution is a generalization of the exponential distribution. However, it has a hazard rate, which may have different shapes. For γ=1, the distribution has constant hazard (i.e. exponential distribution); for γ>1, it belongs to IFR class and γ<1, to DFR class.

The next step in the specification of the accelerated failure time model is to impose a probability distribution on the survival times. If this is chosen to be the Weibull distribution with scale parameter λ and shape parameter γ, written *W(λ, γ)*, the baseline hazard function is

$$h_0(t) = \lambda\gamma t^{\gamma-1}.$$

The hazard function for the i^{th} individual is given by

$$h_i(t) = e^{\eta i} \lambda\gamma (e^{\eta i} t)^{\gamma-1} = (e^{\eta i})^{\gamma} \lambda\gamma t^{\gamma-1}$$

So that the survival time of this individual has a $W(\lambda e^{\gamma\eta_i}, \gamma)$ distribution. The Weibull distribution is therefore said to posses the accelerated failure time property. The Weibull distribution has both the proportional hazards property and accelerated failure time property; there is a direct correspondence between the parameters under the two models. If the baseline

hazard function is the hazard function of a *W(λ, γ)* distribution, the survival times under the proportional hazards model have a $W(\lambda e^{\eta_i}, \gamma)$ distribution, while those under the accelerated failure time model have a $W(\lambda e^{\gamma\eta_i}, \gamma)$ distribution.

Demographic data available for analysis included age, sex, marital status, educational level, clinical extent of disease, type of treatment given, vital status and survival duration. Data describing hospitalization experiences were captured from calendar year 2010 to 2012 and collected for 2581 oral cavity cancer patients' data

5.0 Numerical calculations

Table 1. Analysis of Weibull AFT Model

Effect	DF	Wald Chi-square	Pr > Chi-square	**Estimate (β)**	**SE**	95 % Confidence Limits		-2 Log Likelihood
						Lower	**Upper**	
Intercept	-	3.15	0.0757	-3.1520	1.7748	-6.6306	0.3266	7899.1100
Age	1	49.49	<0.0001	-0.1503	0.0214	-0.1922	-0.1084	
Sex	1	1.07	0.3..9	0.0570	0.0551	-0.0510	0.1650	
Marital Status	1	0.00	0.9965	-0.0002	0.0452	-0.0899	0.0885	
Education	1	2.66	0.1028	0.0465	0.0285	-0.0094	0.1024	
Clinical Extent of Disease	1	57.37	<0.0001	-0.5953	0.0786	-0.7494	-0.4413	
Type of Treatment Given	1	23.24	<0.0001	-0.0073	0.0015	-0.0103	-0.0043	
Scale	-	-	-	1.1189	0.0176	1.0850	1.1539	
Shape	-	-	-	0.8937	0.0140	0. 8666	0.9217	

The analysis of the covariates age, clinical extent of disease and type of treatment given are found to have impact in the survival rate of the patients (Table 1). So these factors are suggested as the covariates for estimating the survival rate of oral cancer patients. The covariates sex, marital status, mother tongue, religion and education are found to have no impact in survival rate of the patients.

6.0 CONCLUSIONS:

Random censoring capabilities and the baseline option is a powerful tool for handling early departure causing incomplete data for subjects during the study period. The Weibull AFT model is one of the most powerful model for identify the covariates for estimating vital status of oral cavity cancer patients. The factors age, education, clinical extent of disease and type of treatment given are the influence factors for the response variable vital status. These influence factors are considered for independent variables in this model. The factors age, education, clinical extent of disease and type of treatment given which are possibly related to the prognosis. Those who have the worst level of the disease have very lesser survival rate. Similarly, those who have higher age group have less survival rate. Those who have the lower age group have more survival rate. The types of treatments are also having impact in the survival rate of the patients.

REFERENCES:

1. Bradburn, M.J, Clark, T.G, Love, S.B, Altman, D.G. (2003), "Survival Analysis Part II: Multivariate data analysis - an introduction to concepts and methods", *British Journal of Cancer* 89 (89): 431–436,
2. Collett, D. (1991) Modeling Binary Data, Chapman and Hall, London.
3. Collett, D. (1994) Modeling Survival Data in Medical Research, Chapman and Hall, Madras.
4. Collett, D. (2003), Modelling Survival Data in Medical Research (2nd ed.) CRC press, ISBN 1-58488-325-1
5. Cox, David Roxbee; Oakes, D. (1984), Analysis of Survival Data, CRC Press, ISBN 0-412-24490-X
6. Daniel, W.W. (1999) Biostatistics: A foundation for Analysis in the Health Sciences, John Wiley and Sons, New York.
7. Deshpande, J.V. and Purohit, S.G. (2001) Survival Hazard and Competing Risks, *Current Science*, 80, 9, 1191-9.

8. Escobar, L.A. and Meeker, W.Q. (1992) Assessing influence in regression analysis with censored data, *Biometrics*, 48, 507-528.

9. Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression, Wiley, New York.

10. Hougaard, Philip (1999), "Fundamentals of Survival Data", *Biometrics* 55 (1): 13–22,

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

12. Kaplan, E.L. and Meier, P. (1958) Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, 457-481.

13. Kay, R. (1984) Goodness of fit methods for the proportional hazards model, *Revue Epidemiologie et de Sante Publique*, 32, 185-98.

14. Kleinbaum, D.G. (1996) Survival Analysis: A Self - Learning Text, Springer, New York.

15. Lin, D.Y. and Wei, L.J. (1991) Goodness of fit tests for the general Cox regression model, *Statistica Sinica*, 1, 1-17.

16. McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models, 2nd edn., Chapman and Hall, London.

17. Miller, R.G. (1981) Survival Analysis, Wiley, New York.

18. Martinussen, Torben; Scheike, Thomas (2006), Dynamic Regression Models for Survival Data, Springer, ISBN 0-387-20274-9

19. Prentice, R.L., Kalbfleisch, J.D., Peterson, A.V., Jr., Flournoy, N.S., Farewell, V.T. and Breslow, N.E. (1978) The analysis of failure times in the presence of competing risks, *Biometrics*, 34, 541-54.

20. Swaminathan, R. (2002) Some Statistical Models in Cancer Survival and Their Applications, Ph.D. Thesis, Madras University.

21. Wei, L.J. (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis, *Statistics in Medicine*, 11, 1871-9.

22. Weissfeld, L.A. and Schneider, H. (1990) Influence diagnostics for the Weibull model fit to censored data, *Statistics and Probability Letters*, 9, 67-73.