Parametric Cox's Model for Partly Interval-Censored Data with Application to AIDS Studies

F. A. M. Elfaki, M. Azram, and M. Usman

***Abstract*—The Parametric Cox's Proportional Hazard Model based on Expectation-Maximization (EM) algorithm for partly interval-censored data is studied. We mean by partly interval-censored data that the observed data include both the exact and interval-censored observations on the survival time of interest. Through the simulation data and real data, we demonstrate that the resulting estimate of regression coefficient and its associated standard error. Our proposal is easily implemented by using SAS software in the present of partly interval-censored data.**

***Index Terms*—About four key words or phrases in alphabetical order, separated by commas. cox model, weibull distribution, partly interval-censored data, HIV/AIDS data, EM algorithm**

I. Introduction

This paper discusses a parametric comparison of survival functions based on incomplete survival data: partly interval censored failure data (Zhao. X, et al., 2008; Kim., J 2003; Peto and Peto, 1972; Huang, 1999). Partly interval censored data arise when the event of interest is observed directly for some subjects, but for the remaining subjects, the event of interest observed belong to the interval time. Also partly interval censored data is an important type of interval-censored, which arise in medical and health studies for example that entails periodic follows-up; see the Framingham Heart Disease Study (Odell et al., 1992) and the Danish Diabetes Study (Ramlau Hansen et al., 1987).

Zhao. et al., (2008) presented a class of generalized log-rank test for partly interval censored failure time data. Kim (2003) used the proportional hazards model for regression analysis of partly interval censored data. Huang (1999) studied asymptotic properties of the nonparametric maximum likelihood estimator of a distribution function based on partly interval censored data. Peto and Peto, (1972) discussed partly interval censored data, treating an exact observation as an interval-censored observation with very short interval.

A general model is adopted in this paper which incorporates most of the widely used life stresses, such as the proportional hazards model. The model can be used for single or multiple stresses. Under this formulation, the model can be either solved as a Proportional Hazards Weibull

Model (PHWM). Several authors proposed methods for analyzing multivariate failure time data on the basis of an assumed parametric model. Namely, Lawless (1983), Basu and Chosh (1980); Ansell and Phillips (1989); Larson and Dinse (1985); Elfaki (2007).

In this paper we are dealing with partly interval censored data to use parametric Cox's model that is; Cox's model with Weibull distribution. The maximum likelihood estimator of the regression parameter and the cumulative hazard function are computed by using EM algorithm

II. Parametric Estimation Procedures

The proportional hazards (PH) regression model is commonly used in the analysis of survival data and, recently, there has been an increasing interest in its application in reliability engineering. The hazard rates of the individuals with different explanatory variables are proportional to each other. Here, there is a baseline hazard, $h_0(t)$, corresponding to the standard condition, and the explanatory variable, z, acting multiplicatively on the baseline hazard, that is, the effect of the covariates is to increase or decrease the hazard.

Let T be a continuous random variable representing an individual's lifetime, and let $z=(z_1,...,z_p)$ be a known vector of regressor variable associated with the individual. Under the proportional hazards assumption, the hazard function of T, given z, is of the form $h(t/z)=h_0(t)g(z,\theta)$, where θ is a vector of unknown parameters. Following Cox's (1972), we will focus on a particular model that is

$$h(t/z)=h_0(t)\exp(z\theta) \tag{1}$$

where $\theta=(\theta_1,...,\theta_p)'$ is a vector of regression coefficients. Model (1) is flexible enough for many purposes. Model (1) requires two model assumptions, namely,

A baseline distribution where the standard condition holds; and

A functional form for the dependency of the lifetime on the covariates, often in terms of parametric model.

In this paper, we examine a fully parametric approach of model (1). The standard approach to inference for the a parametric regression models is the EM algorithm method. Here, if we observe a subject who failed at time t, then, the contribution to the likelihood is $f(t;\theta,z)$, the density function at t. The contribution from a subject censored at t

Manuscript received May 10, 2012, reviewed June 12, 2012.

F. A. M. Elfaki and M. Azram are with Department of Science In Engineering, Faculty of Engineering, IIUM, P. O. Box 10, 50728 Kuala Lumpur, Malaysia (e-mail:faizelfaki@yahoo.com).

M. Usman is with Department of Mathematics, Faculty of Science, Universitas Lampung , Bandar Lampung, Indonesia.

is $R(t;\theta,z)$, the probability of survival (reliability) beyond t. Thus, full likelihood based on the data (t_i,δ_i,z_i), $i=1,2,...,n$, is given by Lawless (1983) and Kalbfleisch and lawless (1988), as follows:

$$L(\theta)=\prod_{i=1}^{n} f(t_i;\theta,z_i)^{\delta_i} R(t_i;\theta,z_i)^{1-\delta_i} \tag{2}$$

where δ_i's are the event indicator variables ($\delta_i=1$ if the ith subject fails; $\delta_i=0$ if the ith subject is censored), θ is a parameter that indexes the density function; and z_i are the covariates for the ith subject.

Taking the natural logarithm of equation (2) simplifies the optimization. The log-likelihood function is given by Lawless (1983) and Kalbfleisch and lawless (1988) as follows:

$$l=\ln(L)=\sum_{i=1}^{F}\ln[f(T_{F,i}]+\sum_{j=1}^{S}\ln[R(T_{S,j}] \tag{3}$$

where T_F is the exact time to failure and T_S is the censored time to failure. The model will be formulated in such a way that equation (3) will be a function of the parameters by expressing the probability density function (pdf) and survival (reliability) functions in terms of these parameters

A. The PH Weibull Model

The Weibull distribution is commonly used for analyzing lifetime data. Also, can be used as the underlying life distribution. In other words it is assumed that the baseline failure rate in equation (1) is parametric and is given by the Weibull distribution. In this case, the baseline failure rate is given by:

$$h_0(t)=\eta\alpha^{-1}(t/\alpha)^{\eta-1}\exp[-(t/\alpha)^{\eta}] \tag{4}$$

where α is the scale parameter depending on z and η is the shape parameter. In fact, η does not depend on z implies proportional hazards for lifetimes and constant variance for log lifetimes of individuals. This assumption is reasonable in many situations, as discussed by Peto and Lee (1973) and Pike (1966).

The PH failure rate then becomes,

$$h(t/z)=\frac{\eta}{\alpha}(\frac{t}{\alpha})^{\eta-1}e^{\sum_{j=0}^{m}\theta_j z_j}\exp\left[-t^{\eta}.e^{\sum\limits_{j=0}^{m}\theta_j z_j}\right] \tag{5}$$

It is often more convenient to define an additional covariate $z_0=1$, in order to allow the Weibull scale parameter raised to the beta (shape parameter) to be included in the vector of regression coefficients. The PH failure rate can then be written as:

$$h(t/z)=\beta.t^{\beta-1}e^{\sum_{j=0}^{m}\theta_j z_j}\exp\left[-t^{\beta}e^{\sum\limits_{j=0}^{m}\theta_j z_j}\right] \tag{6}$$

The survival (reliability) function can be derived as,

$$R(t,z)=e^{-\int_0^t\lambda(t,z)du}=e^{-t^{\beta}.e^{\sum_{j=0}^{m}\theta_j z_j}} \tag{7}$$

where, $\lambda(t,z)$ is failure rate of the model (1).

The pdf can be obtained by taking the partial derivative of the reliability function given by equation (7) with respect to time.

The survival function and the Weibull pdf can then be substituted into equation (3). This yields the likelihood function for PHW model, as follows:

$$l=\sum_{i=1}^{F}\ln\left(\beta.T_{F,i}^{\beta-1}\exp\left(\sum_{k=0}^{m}\theta_k z_{i,k}\right)\exp\left(-T_i^{\beta}e^{\sum\limits_{j=0}^{m}\theta_j z_{ij,j}}\right)\right)-\sum_{k=1}^{S}T_{S,k}^{\beta}.e^{\sum\limits_{j=0}^{m}\theta_j z_j} \tag{8}$$

Solving the parameters that maximize equation (8) will yield the parameters for the PHW model, which are obtained by simultaneously solving the following partial derivatives $\frac{\partial l}{\partial\beta}=0$, $\frac{\partial l}{\partial\theta}=0$.

III. Simulation Studies

A small simulation study was conducted to evaluate the finite sample performance of our proposed method. Our simulation set-up is similar to that in Kim (2003). We generated data from exponential distribution with $h_0(t)=1$ under our proposed model $h(t/z)=h_0(t)\exp(z\theta)$. Examination times were generated to make the proportions of left, interval, and right-censored observations about equal. The sample size n is the sum of the number of exact data n_1 and the number of interval-censored data n_2. Following Kim (2003) we consider the range as; (25, 25) and (40, 10) for a sample of size 50, and (50, 100), (50, 150) and (50, 200) for the sample size 150, 200 and 250 respectively. We refer to (t,z) and (t_1,z_1) as the original data and exact data respectively. Table I show our results compared with one obtained by Kim (2003). For each sample, we obtained the bias and the mean standard error (which is not addressed here). The estimation based on the exact data and original complete data. The results look similar to the one obtained by Kim (2003).

TABLE I: Comparison Results Obtained by our Proposed Model With Kim (2003) From Simulation Data from 1500 Replication.

	Kim (2003)			Proposed Model		
(n_1,n_2)	Biases			Biases		
	$\hat{b}_E$	$\hat{b}_O$	$\hat{b}_{OC}$	$\hat{\theta}_E$	$\hat{\theta}_O$	$\hat{\theta}_{OC}$

(25, 25)	0.258	0.103	0.062	0.242	0.112	0.073
(40, 10)	0.081	0.061	0.058	0.072	0.074	0.052
(50, 100)	0.079	0.030	0.019	0.088	0.042	0.022
(50, 150)	0.075	0.023	0.011	0.082	0.034	0.023
(50, 200)	0.074	0.014	0.008	0.085	0.017	0.012

where $\hat{\theta}_E$ estimated bias from exact data; $\hat{\theta}_O$ estimated bias from observed data; $\hat{\theta}_{OC}$ estimated bias from original complete data

IV. Illustrative Example

The proposed method is illustrated HIV/AIDS of hemophiliacs who were treated in two hospitals in Sudan. They were 550 patients of the study who were at risk for HIV infection through the contaminated blood factor. At the end of the study, there were 550 patients found to be HIV infected, but the infection times were interval-censored. Among them 120 progressed to AIDS (or related symptoms). The patients were classified into either the heavily treated group or lightly treated group according to the amount of blood received (when treated for hemophilia). The goal here is to investigate the possible association between the treatment and the AIDS incubation time. We code the covariate $z_i = 0$ or $z_i = 1$ if the ith patient was lightly or heavily treated. To see the effect of covariates on development of complications, we fitted our proposed model that is proportional hazards regression model with Weibull distribution. Appling the procedures described in section 2, we obtained the result as shown in Table II, or as $(\hat{\theta}_1, \hat{\theta}_2) = (0.621, 0.632)$, the Wald test statistic for testing $(\theta_1, \theta_2) = (0,0)$ is 3.481 and P-value for testing $(\hat{\theta}_1, \hat{\theta}_2) = (0,0)$ is 0.0272. We conclude that the covariates do not have a significant different. However, it is confirmed that the heavily treated group had a significantly higher risk of the onset of AIDS after HIV infection.

TABLE II: Estimate of PHWM Based on EM Algorithm for Sudan HIV/ADIS Data

	$\hat{\theta}_E$	$\hat{\theta}_O$	$\hat{\theta}_{OC}$
Heavily Treat Group	0.621	0.610	0.632
Lightly Treat Group	0.632	0.620	0.644

V. Conclusion

The parametric Cox's proportional hazards regression model with Weibull distribution based on EM algorithm has been used successfully to investigate the causes of failure for HIV infection. EM algorithm was used to estimate the parameters of the model. Through the simulation studies, we find that our approach show similar result as the one obtained by Kim (2003). The simulation studies strongly support the generalized missing information principle in a parametric context and use of the generalized profile information for non-identically distributed samples. From the real data set we find that the covariates do not have a significant different. Fixing the gender at diagnosis, a male has a lower hazard rate than female. Fixing age, very young patients have a lower hazard rate than relatively young patients. Even with many exact observations (550), the additional interval-censored observations (126) help to give a more accurate estimate of the regression parameter. However, it is confirmed that the heavily treated group had a significantly higher risk of the onset of AIDS after HIV infection

Acknowledgement

This work was partially supported by project with code number 11-036-0184

References

[1] X. Zhao, Q. Zhao, J. Sun, and S. J. Kim, "Generralized log-rank test for partly interval-censored failure time data," *Biometrical Journal*, vol. 3, pp. 375-385, 2008.

[2] J. S. Kim, "Maximim likelihood estimation for the proportional hazards model with party interval-censored data," *J R. Statist.* Soc. 2003, Series B65, pp. 489-502.

[3] R. Peto and J. Peto, *Asymptotically Efficient Rank Invariant Test Procedures*. J R. Statist. Soc. 1972, Series A135, pp. 187-220.

[4] J. Huang, "Asymptotic properties of nonparametric estimation based on party interval- censored data," *Statistica Sinica*. 1999, vol. 9, pp. 501-519.

[5] P. M. Odell, K. M. Anderson, and R. B. D' Agostino, "Maximum likelihood estimation for interval-censored data using a weibull based accelerated failure rank invariant test procedures," *J R. Statist.* Soc. 1992, Series A135, pp. 185-207.

[6] H. Ramlau-Hansen, N. C. Jespersen, and P. K. Anderson, *Life Insurance for Insulin Dependent Diabetics. Scandinavian Actuarial Journal*, vol. 39, pp. 19-36, 1987.

[7] M. G. Larson and G. E. A. Dinse, "Mixture model for the regression analysis of competing risks data," *Applied Statistics*. vol. 34, no. 3, pp. 201-211, 1985.

[8] F. A. M. Elfaki, N. A. I. Daud, and M. Y. Ibrahim, "Abdu Allah and M. Usman. Competing risks for reliability analysis using Cox's model. Engineering Computations," *International Journal for Computer-Aided Engineering and Software*, Issue 3, vol. 24, pp. 373-383, 2007.

[9] D. R. Cox, "Regression models and life tables (with discussion)," *J R. Statist.* Soc. vol. 34, pp. 187-220, 1972.

[10] R. R. Peto and P. Lee, "Weibull Distributions for Continuous Carcinogens is Experiments," *Biometrics*. 1973, vol. 29, pp. 457-470.

[11] M. C. Pike, "A method of analysis of a certain class of experiments in carcinogensis," *Biometrics*. 1966, vol. 22, pp. 142-161.

[12] J. D. Kalbfleisch and J. F. Lawless, "Estimation of reliability in field-performance studies," *Technometrics*. 1988, vol. 30, pp. 365-388.

[13] J. F. Lawless, "Statistical methods in reliability," *Technometrics*. 1983, vol. 25, pp. 305-335.