GLOBAL JOURNAL OF MATHEMATICAL SCIENNCES VOL. 8, N0. 1, 2009: 13 - 17
COPYRIGHT (C) BACHUDO SCIENCE CO. LTD. PRINTED IN NIGERIA. ISSN 1596-6208

BAYESIAN ANALYSIS OF RIGHT CENSORED SURVIVAL TIME DATA

A. A. ABIODUN

(Received 7 October 2008; Revision Accepted 3 February 2009)

ABSTRACT

We analyzed cancer data using Fully Bayesian inference approach based on Markov Chain Monte Carlo (MCMC) simulation technique which allows the estimation of very complex and realistic models. The results show that sex and age are significant risk factors for dying from some selected cancers. The risk of dying from these cancers is observed to progressively increase as age of patients increases. It is also observed that in order to allow for nonlinearity due to metrical covariate age, the semiparametric P-splines model is better than the model that categorizes age into various age groups.

KEY WORDS: Survival time, Censoring, Random effects, Markov Chain Monte Carlo.

1. INTRODUCTION

Analysis of survival or failure times has gained a considerable attention, particularly in the field of medical applications wherefrom the conventional denotation 'survival analysis' arises [Hennerfeind(2006)]. Censoring is one phenomenon that makes survival analysis differ from other analyses. This is a situation of incompleteness in the observed survival data. The most common censoring in survival time data is Right Censoring which occurs when the actual time a subject experiences the event of interest is not known. In this type of censoring, it is assumed for some individuals in the study that there is a time to event T_e and the right censoring time C where the T_e's are assumed to be independently and identically distributed with density function $f(t)$ and survival function $S(t)$. The exact survival time T of any individual will be known if and only if T_e is less than or equal to C. If T_e is greater than C, then the individual is a survivor and the exact survival time is censored at C. Thus the observed time is T = min(T_e,C) and the data for such a design can be represented by pairs of random variables (T, δ),where δ indicates whether the survival time T corresponds to an event (δ=1) or is right censored (δ=0).

An aspect of analysis of survival time data that has gained popularity, especially in medical research is assessing the relationship between survival time and some biological, socio-economic and demographic characteristics that could possibly affect the survival status of patients. One popular regression model formulation that is often used in survival analysis is the Cox (1972) proportional hazards model. The model utilizes the hazard function $\lambda(t)$, also known as the hazard rate or force of mortality which is defined as the probability of experiencing event of failure in the infinitesimally small interval (t, $t+\Delta t$), given that such an event has not been experienced prior to t. It is expressed as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr\langle t \le T \le t + \Delta t \,|\, T > t \rangle}{\Delta t} \quad . \qquad (1.1)$$

1.1. Likelihood for Right Censored Data

The likelihood for censored data is derived by considering the observed survival times t_i. Suppose we have n subjects with subject i observed for a time t_i, if the subject fails at time t_i, then its contribution to the likelihood function (under non- informative censoring) is

A. A. Abiodun, Department of Statistics, University of Ilorin, P.M.B 1515, Ilorin, Nigeria

$$L_i = f(t_i) = S(t_i)\lambda(t_i). \quad (1.2)$$

If the subject is still alive at t_i, all we know under non- informative censoring is that the lifetime exceeds t_i and thus the contribution of such censored observation to the likelihood is

$$L_i = S(t_i). \quad (1.3)$$

Let δ_i be a failure indicator which takes value 1 if subject *i* fails at time t_i and value 0 is subject *i* is censored. Then we write the full likelihood as

$$L = \prod_{i=1}^{n} L_i = \lambda(t_i)^{\delta_i} S(t_i). \quad (1.4)$$

2. COX PROPORTIONAL HAZARDS MODEL FORMULATION

Suppose that the data collected on *n* subjects are denoted by (t_i, δ_i, Z_i), where t_i is time to failure of the *ith* subject, δ_i is the censoring indicator such that for the *ith* subject, δ_i =1 if event of failure occurs to the subject at time t_i and δ_i = 0 if the time is right censored (i.e we observe some value *c* with the knowledge that $t_i > c$) and Z_i is a p- dimensional vector of covariates. Cox (1972) model assumes that the hazard function for the *i-th* subject with covariate value Z_i has the form

$$\lambda(t_i, Z_i) = \lambda_0(t)\exp(Z_i'\gamma), \quad (2.1)$$

where $\lambda_0(t)$ is an arbitrary baseline hazard function and γ is a p- vector of unknown regression coefficients. Model (2.1) is semi-parametric because the dependence function $\exp(Z_i'\gamma)$ is modelled explicitly but no specific probability distribution is assumed for the survival times. Thus γ is only estimable through the partial likelihood estimation procedure.

Often, survival time data involve identified clusters of subjects according to some unobserved characteristics such that subjects belonging to the same cluster are similar with respect to such characteristics so that the survival times of such subjects are correlated whereas the survival times of subjects belonging to different clusters are independent. One appropriate way of analyzing such data is to use random effect (frailty) model.

$$\lambda(t_i | Z_i, W) = W_c\lambda_0(t)\exp(Z_i'\gamma), \quad (2.2)$$

where W_c is the random effect (frailty) shared by the subjects belonging to cluster *c*

Model in (2.2) can be written as

$$\lambda(t_i) = \exp(\eta_i(t)), \quad (2.3)$$

with $\eta_i(t) = g_o(t) + \gamma' Z_{ic} + b_c)$, where $g_o(t) = \log\lambda_o(t)$ and $b_c = \log(W_c)$

Model expressed in (2.3) assumes that effects of covariates are linear on the log hazards and are thus modelled parametrically as fixed effects. Often, in practical situations, effects of continuous covariates are not linear and thus cannot be adequately modelled as fixed effects. Thus extending Hennerfeind et al (2005), the parametric predictor $\eta_i(t)$ in (2.3) is replaced with a more flexible semiparametric structured additive predictor that incorporates this complexity within the same framework. Thus the Cox type hazard model, (2.1) can be written as

$$\eta_i(t) = g_o(t) + \sum_{j=1}^{p} f_j(x_{ij}) + \gamma' Z_i + b_{ci}, \quad (2.4)$$

where f_j is the nonlinear effect of a continuous covariate x_j,

γ is the vector of usual linear fixed effects,

b_c is the cluster specific random effect (frailty) with $b_{ci} = b_c$ if *i-th* individual is in cluster , *c* =1,…,*C*. Clearly, b_c are usually assumed to be independent realizations from normal or log-gamma distribution with known mean and unknown variance.

3. BAYESIAN INFERENCE

Bayesian analysis requires assignment of priors. Thus for defining priors and developing posterior analysis, the predictor (2.4) needs to be rewritten in generic matrix notation. Thus we express g_o, f_j and *b* as the matrix product of an appropriately defined design matrix ***Z*** which leads to re-expressing (2.4) as

$$\eta = Z_0\beta_0(t) + Z_1\beta_1 + \cdots + Z_m\beta_m + +V\gamma . \quad (3.1)$$

We then assign priors as follows. For fixed effect parameter γ we have assumed diffuse priors i.e. $P(\gamma) \propto const$

The general form of priors for β_j can be cast into the form

$$p(\beta_j|\tau_j^2) \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2r_j^2}\beta_j^{/}K_j\beta_j\right),$$

where K_j is a precision or penalty matrix of rank (K_j) = r_j, which shrinks parameters towards zero or penalizes too abrupt jumps between neighbouring parameters.

For the baseline g_o and non-linear effect f_j of continuous covariate, we assign Bayesian P-splines prior as in Lang and Brezger (2004) and the random effect b_c are assumed to be i.i.d Gaussian. i.e $b_c \sim N(0, \tau_c^2)$.

4. APPLICATION: HOSPITAL ADMISSION OF CANCER PATIENTS

We consider data on cancer patients who were admitted at the University of Ilorin Teaching Hospital (UILTH) from 1999 to 2005. The record of each patient contains information on variables length of stay in the hospital recorded in days, sex, age and outcome which indicates whether the patient is dead or alive. We define survival time as length of stay till event of death occurs while those whose records read “alive” were right-censored because such patients had not died as at the time of the study. Nine types of cancer were selected and the Patients were grouped into nine cancer/tumor types/sites, which include: carcinoma, leukaemia, lymphoma, melanoma, sarcoma, rectum, lung, liver and stomach. Prostate and breast cancers are not included because they are gender related and may possibly introduce gender bias into the analysis.

Fitting variable cancer type as fixed effect requires that we construct eight dummy variables, and this result in eight parameter estimates to be compared to an arbitrarily chosen reference category. A more efficient alternative to this is to fit the cancer type as a random effect (frailty).

At the initial stage, we fitted sex and continuous age as fixed effects with diffuse prior. That is we fitted model

$$\eta = fo(t) + sex\,\gamma_1 + age\,\gamma_2$$

Table 1 shows the posterior estimates, standard errors and the 95% credible intervals. Effects of sex and age when fitted as fixed effects are seen to be significant as the credible intervals do not include zero. To gain more insight into the analysis with respect to gender differences, we fitted models for combined and then male and female differently. Since the assumption of linear effect of metrical covariates such as age on the predictor is too restrictive as discussed in section (2), we consider two widely used alternative ways to allow for non-linearity in the effects of metrical covariates. In the first alternative, we categorize the covariate age by constructing a set of d_j variables $\tilde{z}_j$, $j = 1, \ldots, p$, with one being arbitrarily chosen as a reference category, thereby producing $p-1$ dummies with $\tilde{\gamma}_1, \cdots, \tilde{\gamma}_{p-1}$ parameters to be estimated for the categorized covariate. In the second alternative, which is a more flexible and data driven way, we incorporate age additively in the predictor using smooth regression function $f_j(x_j)$ and then model it nonparametrically using P-splines prior as in Lang and Brezger (2004). In this paper, Sex was coded 1 for male and 0 for female patients. The metrical age was coded into four categories: “less

than 23 years" (reference group), "23-39 years", "40-55 years", and "greater than 55 years". Our research interest thus includes: investigating the effect of categorized age on the risk of dying from cancer for the cancer patients combined and for male and female separate, comparing the two ways described above by considering some hierarchical models, starting from very simple model and progressively increase model complexity. Model comparisons are based on Deviance information criterion (DIC) introduced by Spiegelhalter et al (2002), which is a Bayesian analogue of Akaike information criterion (AIC). The following models are fitted, noting that all models contain baseline effect.

Model I: $\eta = f_o(t) + b_c$ (random effect) - Null Model

Model 2: $\eta = f_o(t) + f_{age}$ (metrical age)

Model 3: $\eta = f_o(t) + f_{age} + b_c$ (metrical age with random effect)

Model 4: $\eta = f_o(t) + (23-39)\gamma_1 + (40-55)\gamma_2 + > 55\gamma_3$ (categorical age)

Model 5: $\eta = f_o(t) + (23-39)\gamma_1 + (40-55)\gamma_2 + > 55\gamma_3 + b_c$ (categorical age with random effect)

5. RESULTS

Results for the analyses are presented in table 2, showing fixed effects of age of patients for the combined, male and female and in table 3, showing the hierarchical models under the categorized age and age fitted by P-splines.

Table 1: Effect of sex and continuous age fitted as fixed effects

Covariate	Posterior mean	Std. error	2.5 Quant.	97.5 Quant.
Sex	-0.4100	0.667	-1.219	-0.226
Age	0.0088	.0062	0.0023	0.0208

Table 2: Fixed effect of age for the combine, male and female patients

(a) Combined

Covariate(Age)	Posterior mean	Std. error	2.5 Quant.	97.5 Quant.
23-39 years	0.290	0.396	0.485	1.082
40-55 years	0.418	0.375	0.324	1.189
>55 years	0.633	0.358	0.061	1.347

(b) Male

Covariate(Age)	Posterior mean	Std. error	2.5 Quant.	97.5 Quant.
23-39 years	0.443	0.465	0.644	1.204
40-55 years	0.534	0.475	0.324	1.295
>55 years	0.657	0.482	0.431	1.338

(c) Female

Covariate(Age)	Posterior mean	Std. error	2.5 Quant.	97.5 Quant.
23-39 years	0220	0.413	0.536	1.109
40-55 years	0.423	0.387	0.349	1.332
>55 years	0.545	0.438	0.393	1.281

The results in table 2 a,b and c are the posterior means, standard errors and the quantiles of fixed effects of the categorized age for combined, male and female patients. It is observed that the risk of dying from cancer increases with age for both combined and both sexes separately. For example, in the combined data, patients in age group 23-39 years have a risk of exp(0.290) which is 1.33 times that of patients in the reference category (less than 23 years).

The results are in the same direction for males and females, though the risks are relatively much higher for male than their female counterpart. For example, when the risk for male patients in age category 40-55 is 1.70 times those in the reference category, it is 1.52 for the females.

Table 3: DIC for the various models for combine, male and female patients

Model	DIC		
	Combined	Male	Female
M1	707.177	300.124	413.924
M2	740.315	306. 260	421.026
M3	708.331	288.941	408.092
M4	748.043	309.906	427.728
M5	708.474	307.876	415.682

It is observed in Table 3 that all the models fitted are best for the male patients alone and worst for the combined data as revealed by the values of the DIC which is least for the males and highest for the combined. It is also observed that the P-splines models for age are better than models with categorized age as the DIC values are seen to be smallest for the later than the former throughout for the combined, male and female, and we also observe that the data really contains random effect (frailty) and that models that take this into account are better than those that ignore it.

6. CONCLUSION

In the analysis of data on hospital admission for the cancer patients under study, results show significant differences among age groups with respect to the risk of dying from the selected cancer considered. Results of Deviance information criterion (DIC) also reveal that when we allow for non-linearity in the effects of metrical covariate age, the nonparametric model using P-splines prior as in Lang and Brezger (2004) is preferred over the model that categorize age.

Software Package: All analyses in this paper have been done using BayesX, a public domain software package for performing complex full and empirical Bayesian inference is available at http://www.stat.uni-muenchen.de/~lang/BayesX.

Limitation of the study: The major caveat to be considered when interpreting the result is about patient's age which is self reported. Most often, self reported age by patients may not be their true age. Despite this limitation, the study strength is significant.

REFERENCES

Cox, D. R., 1972. Regression models and Life-tables (with discussion). J. Roy. Statist. Soc. Ser. B. 34: 187-220.

Hennerfeind, A., 2006. Bayesian nonparametric regression for survival and event history data. P.hD at http://www.stat.uni-muenchen.de

Hennerfeind, A., Brezger, A. and Fahrmeir, L., 2005. Geoadditive Survival Models. SFB Discussion Paper 414. Available at htt://www.stat.uni-muenchen.de/sfb386/papers/disp/paper414.pdf.

Lang, S. and Brezger, A., 2004. Bayesian P-splines. J. Comput. Graph. Statist. 13: 183-212

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and vander Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. B, 65: 583-639.