RESEARCH PAPER | Statistics | Volume : 3 | Issue : 12 | Dec 2013 | ISSN - 2249-555X

Spline-based Hazards Regression Model for Current Status Data: An Application to Simulated Data on Renal Impairment

KEYWORDS: Current Status, Proportional Hazards, Sieve Maximum Likelihood Estimation, Spline.

Gurprit Grover
Associate Professor, Department of Statistics, University of Delhi, Delhi-110007, India.

Barnali Deka
PhD Scholar, Department of Statistics, University of Delhi, Delhi-110007, India.

ABSTRACT *Regression modeling of current status data involves the unknown baseline cumulative hazard function. Estimation procedures yields non-smooth curves that complicates the process of understanding the behavior of survival function. Here we have proposed a sieve semiparametric maximum likelihood estimation method for the proportional hazards model for current status data. We have flexibly parameterized the unknown baseline cumulative hazards function using monotone splines. The developed estimation procedure has the advantage of being a computationally efficient one to produce smooth estimates of the survival (or hazard) function and regression parameters as well.*

1. INTRODUCTION

Survival modeling seeks to obtain the probability of occurrence of an event of interest, such as the onset of a disease. Usually the onset time is supposed to be known or right censored. However, many medical and epidemiological investigations capture only the status of an event at the time of examination; because the time to the event of interest cannot be observed exactly. These studies give rise to current status or case-1 interval-censored data. Current status data occurs in many studies– carcinogenicity, demographical, econometrics, epidemiological and reliability[1,2]. An important example is the cross-sectional surveys to study the occurrence of certain chronic disease which is rather cost-effective than long-term follow-ups[2,3]. This motivates our present study where we have tried to estimate the hazard of chronic kidney disease (CKD) and the effect of its risk factors.

CKD is a major public health concern of India. The exact prevalence of CKD in India is not clear in the absence any regular national registry; but a community-based program from Chennai and another population-based study from Delhi reported the estimates of impaired kidney function to be 0.86% and 0.79% respectively[4,5]. We lack maintained follow-up data in renal clinics of hospitals in India and also long-term population-based follow-ups. Since renal impairment is a chronic condition, the onset time of damage occurrence cannot be observed exactly; only the current status of renal function is known at the time of examination.

Survival analyses of current status data received attention with the development of various algorithms to compute nonparametric maximum likelihood estimates (NPMLE) of survival function[6-8]. For analyzing the covariate effects, regression models have been proposed under various semiparametric structures[2]. We focus here on the widely used proportional hazards (PH) model. It includes an unknown baseline cumulative hazards function (CHF) i.e. hazard of the event in the absence of covariates. Various methods have been proposed for estimating the regression coefficients and unknown CHF[9-11]. Huang (1996) developed a profile likelihood-based efficient estimation approach with theoretical results[12]. However the estimates of baseline CHF from these methods are non-smooth and estimation procedures are computationally intensive. Non-smoothness of baseline estimates can be tackled with a parametric assumption, which is very restrictive in practical situations. Hence, the flexible parametric models with some finite-dimensional spline functions are widely used to approximate the unknown baseline CHF (or survival function).

In our present study, we propose a semiparametric PH model for current status data using cubic monotone splines to approximate the unknown baseline CHF and to estimate the parameters using a sieve estimation procedure. We use our proposed model to analyze the simulated CKD data. Section 2 describes PH models and our proposed model with sieve estimation. Section 3 illustrates the simulation study performed to evaluate the proposed method and compare it to a likelihood approach. Analyses of CKD data are provided in section 4 and concluding remarks are presented in section 5.

2. METHODS

Proportional Hazards Model

Suppose Y_i be the time to event of interest; T_i the examination (or censoring) time, independent of Y_i; and Z_i a p x 1 vector of covariates for i^{th} subject. Denote δ_i=1 or 0 accordingly when i^{th} event time is left censored ($Y_i \le T_i$) or right censored ($Y_i \ge T_i$). Given the observed data $\{(t_i, \delta_i, Z_i), i=1,2,\ldots,n\}$ and assuming non-informative censoring, the likelihood for current status data is

$$L(\beta, S_0) = \prod_{i=0}^{n} [S(t_i; Z_i)]^{(1-\delta_i)} \{1 - [S(t_i; Z_i)]\}^{\delta_i} \quad (1)$$

Here, $S(.;Z_i)$ is the survival function. Let $\Lambda(.;Z_i)$ denote the CHF of i^{th} subject such that $S(.;Z_i) = \exp(-\Lambda(.;Z_i))$.

Under PH assumption, the effect of covariate Z on the cumulative hazard of occurrence of the event by the examination time t can be modeled as

$$\Lambda(t; Z_i) = \Lambda_0(t)\exp(-Z_i'\beta) \quad (2)$$

Here, $\Lambda_0(t)$ is unspecified nondecreasing baseline CHF. So, the log-likelihood is

$$l(\beta, \Lambda_0) = \sum_{i=1}^{n} \{\delta_i \ln(1 - \exp(-\Lambda_0(t_i)e^{Z_i'\beta}) - (1-\delta_i)\Lambda_0(t_i)e^{Z_i'\beta}\} \quad (3)$$

Eq.(3) includes the infinite-dimensional nuisance parameter Λ_0, which cannot be eliminated using partial likelihood as in the case of right censoring. So the maximum likelihood estimates (MLE) of β and Λ_0 have to be derived simultaneously by maximizing the full log-likelihood (3) over (β, Λ_0). Since $\lambda(\beta, \Lambda_0)$ is concave in β (or Λ_0) for any fixed Λ_0 (or β), profile-likelihood procedure of Huang (1996) can be applied[12]. However the estimation of βs along with nonparametric profiling of Λ_0 is computationally quite intensive. Also the non-smoothness of the NPMLEs of Λ_0 further complicates the process of understanding the behavior of the survival (or haz-

Volume : 3 | Issue : 12 | Dec 2013 | ISSN - 2249-555X

ard) function. Therefore, we propose a sieve estimation procedure that involves approximating an infinite-dimensional parameter space by a series of finite dimensional functions.

Sieve Maximum Likelihood Estimation Using Cubic Monotone Splines Smoothing

If we flexibly parameterize Λ_0 in (2) by a linear span of some known basis functions, then the log-likelihood (3) can be thought as sieve log-likelihoods as defined in Geman and Hwang[13]. Then, maximization of these sieve log likelihoods with respect to β and Λ_0 in the linear span would produce the sieve MLEs of (β, Λ_0). Since the linear span consists of finite number of basis functions, the dimensionality of the optimization problem reduces easing the numerical difficulties. The sieve method is a powerful tool in semiparametric survival regression for current status data. It has been used mainly to approximate the baseline CHF or survival function in various survival models for current status data[14-16]. A popular choice for basis functions is splines. Splines are piecewise polynomial functions that are combined linearly to approximate an unknown function on an interval.

Here, we propose to model the baseline CHF, $\Lambda_0(t)$ flexibly using a linear combination of monotone splines (M-splines). An M-spline basis of order k is defined as [17]

$$M_j(x;k)=\begin{cases}\dfrac{k[(x-t_j)M_j(x;k-1)+(t_{j+k}-x)M_{j+1}(x;k-1)}{(k-1)(t_{j+k}-t_j)}, & t_j\le x\le t_{j+k}\\ 0, & \text{elsewhere}\end{cases}$$

$$with \qquad M_j(x;1)=\begin{cases}\dfrac{1}{(t_{j+1}-t_j)}, & t_j\le x\le t_{j+1}\\ 0, & \text{elsewhere}\end{cases}$$

where $t_1,\ldots, t_m$ is a sequence of increasing knots in (0, ∞). To each M-spline, we associate an (Integrated) I-spline basis as

$$I_j(x;k)=\int_0^x M_j(u;k)du \qquad (*)$$

Each M-spline is a piecewise polynomial ok degree k-1, and each associated I-spline is a piecewise polynomial of degree k. M-splines are nonnegative, and so associated I-splines are monotonically nondecreasing. Therefore a linear span of I-spline basis functions can be used to approximate any monotonic function just by constraining the coefficients to be positive.

Let T_{min} and T_{max} be the boundary knots over the observation time axis. Assign m distinct internal knots within [T_{min}, T_{max}], placed as $0\le T_{min}<\tau_1<\tau_2<\ldots<\tau_q<T_{max}$. Considering M-spline basis functions of degree k-1 on each of q+1 subdivisions, the baseline CHF $\Lambda_0(t)$ can be approximated over the whole space [T_{min}, T_{max}] by

$$\Lambda_0(t)=\sum_{j=1}^{m}\alpha_j I_j(t) \qquad (4)$$

Here, m = q+k; $I_j(t)$ is defined by (*); and α's should be nonnegative to ensure that $\Lambda_0(t)$ is nondecreasing. In our study, we place the boundary knots (T_{min} and T_{max}) at the minimum and maximum of observation times respectively. The internal m knots are placed at equally spaced quantiles of the observed times in [T_{min}, T_{max}]. Sieve MLEs of (β, Λ_0) are computed by replacing $\Lambda_0(t)$ in (3) by (4) and maximizing the sieve log likelihood with respect to $\theta = (\alpha_0, \alpha_1, \gamma_1,\ldots, \gamma_m, \beta)$ by using Newton-Raphson method or the function optim in **R**.

Computational Algorithm:

- S1:(Obtain initial values) A set of suitable initial values for θ can be generated applying standard Cox PH model to given data, considering the $T_i|\delta_i=1$ as exact failure time. Generate estimates of β and Λ_0. Then starting values of (α_0, α_1, $\gamma_1,\ldots,\gamma_m$) may be computed by applying the spline model (4) to these estimates of Λ_0 under least square regression setup and with desired number of knots.
- S2: Using Newton-Raphson method or optim in **R**, compute the estimates of θ.
- S3: Estimate the baseline CHF $\Lambda_0(.)$.
- S4: Since $\Lambda_0(.)$ is a nonnegative and monotonically nondecreasing function, we may apply either Pool Adjacent Violators Algorithm (PAVA) to the estimates of $\Lambda_0(.)$ in Step-3 to ensure the monotonicity[6].

3. SIMULATIONS

We performed a simulation study to illustrate the empirical behavior of our sieve estimator. The true event times (Y_i) were procured from a PH model with $\Lambda_0(t)$ following Weibull(3,6). The examination times (T_i) were generated independently from an exponential distribution on the interval (0, 10) with an appropriate parameter so that censoring rates are around 20% and 50%. Two covariates Z_1 (binary) and Z_2 (continuous) were generated from Bernoulli(0.5) and Normal(0, 0.5^2) distributions respectively. The true parameter values were taken as β_1=1 and β_2=0.5. Then the i^{th} observation (t_i, δ_i), where δ_i is the censoring indicator, was generated as: (i) Sample y_i, and t_i from their specified distributions. (ii) If $y_i \le t_i$ then set δ_i = 1 and (t_i, δ_i = 1) is a left censored observation. Else a right censored observation (t_i, δi= 0) is obtained.

We simulated 1500 replications with sample sizes 100, 400 and 1000 in each set. We considered cubic splines (k=3) to allow adequate smoothness with 12 internal knots (q=12) which were placed at equally spaced quantiles within minimum and maximum of observation times. We did apply our model for knots- 7, 10, 12, 15 and 20 no. of knots. However, for 15 and 20 knots there was no improvement in the AIC values. So we report our results for 12 knots. We applied the cubic spline-based sieve method along with the profile-likelihood method by Huang[12].

Table-1: Simulation Results on the Regression Parameters for the Proposed Spline-based Sieve MLE Method and Profile-likelihood Approach

Sample Size	True β	Proposed Method				Profile Likelihood Method			
		Bias	SSE	ESE	CR (%)	Bias	SSE	ESE	CR (%)
Censoring Rate = 20%									
100	0.5	0.057	0.261	0.272	94.3	0.043	0.249	0.252	93.5
	1	0.106	0.238	0.252	94.5	0.092	0.211	0.219	94.3
400	0.5	-0.019	0.218	0.234	94.9	-0.210	0.220	0.238	95.3
	1	-0.021	0.183	0.188	94.7	-0.056	0.200	0.213	94.9
1000	0.5	0.002	0.118	0.120	96.4	0.011	0.133	0.140	95.1
	1	0.007	0.097	0.111	95.6	0.015	0.114	0.136	93.8
Censoring Rate = 50%									
100	0.5	0.072	0.362	0.358	94.3	0.133	0.348	0.339	94.1
	1	-0.054	0.364	0.368	94.8	-0.061	0.344	0.337	93.7
400	0.5	0.047	0.288	0.291	94.5	0.101	0.315	0.317	93.9
	1	0.027	0.274	0.268	95.2	0.049	0.324	0.328	94.5
1000	0.5	-0.011	0.216	0.211	95.7	-0.063	0.247	0.260	94.3
	1	-0.016	0.209	0.218	95.2	-0.038	0.220	0.223	94.3

Volume : 3 | Issue : 12 | Dec 2013 | ISSN - 2249-555X

Table-2: Simulation Results on the MaxMSE (Mean Square Error) of the Estimates of $\Lambda_0(t)$ based on 1500 data Sets from the Proposed Spline-based Sieve MLE Method and Profile-likelihood Approach

Sample Size	Proposed Method		Profile Likelihood Method	
	20%	50%	20%	50%
100	0.0065	0.0063	0.0078	0.0080
400	0.0054	0.0055	0.0067	0.0067
1000	0.0039	0.0042	0.0059	0.0062

Table-1 summarizes various operating characteristics of the estimates of βs from the two simulation studies. The bias is the difference between the average of 1500 point estimates and the true value; ESE the average of the estimated standard errors; SSE the sample standard deviation of the 1500 point estimates. The coverage rates (CR) of true βs by the 1500 95% confidence intervals (CI) have also been reported. Following observations are noted from the simulation results: (i) Small biases are observed for the two estimators. However our proposed spline-based method yields lesser bias for all the parameter estimates; (ii)The ESEs of the proposed approach are close enough to the SSEs; and (iii) The CPs agree with the nominal value 0.95 very well.

Next we estimated the baseline CHF $\Lambda_0(t)$ at some pre-specified equally-spaced quantiles of the event times at each replication and then calculated the mean square errors (MSE) of these estimates. Table-2 shows the maximum of these local MSEs (MMSE) for both methods under the same simulation setup as above. Though both the methods produced quite good estimates of $\Lambda_0(t)$, as shown by smaller MMSE values, spline-based sieve estimates indicate better estimation than the profile likelihood method.

4. APPLICATION

The natural history of CKD has a prolonged asymptotic period, followed by the progression to end stage renal disease (ESRD) or renal failure, requiring dialysis. Substantial loss of kidney function might occur before clinical symptoms become apparent, if not detected timely. A patient with CKD suffers from progressive deterioration of renal function. Stages of CKD are defined according to the kidney function marker- glomerular filtration rate (GFR). Estimate of GFR (eGFR) is obtained by various formulas that use serum creatinine values and other anthropometric parameters. The eGFR (in ml/min per 1.73m^2) is then used to classify subjects into K/DOQI stages of CKD– 1. ≥90 (Normal); 2. 60-89 (Mild); 3. 30-59 (Moderate); 4. 15-29 (Severe) and 5. ≤15 (ESRD)[18].

Our event of interest is the renal impairment (CKD stage 3 onwards) or low eGFR (eGFR ≤ 60 ml/min/1.73m^2). Since renal impairment is a chronic condition, the onset time of damage occurrence cannot be observed exactly; only the current status of renal function is known at the time of examination. However, applications of survival analyses require follow-up data, whether retrospective or prospective. We lack maintained follow-up data in renal clinics of hospitals in India and also large population-based follow-ups. However, some studies had conjured up estimates of prevalence and other epidemiological parameters of CKD in India [4,5,19-22]. In order to apply our proposed method, we generated a hypothetical data set, comprising of 1500 cases, simulated on the basis of information reported in some of studies. Reported prevalences (in %) of low GFR in India are- 0.86, 1.39, 0.79, 13.3, 3.02 and 17.4 [4,19,6,21,23,24]. We calculated the average of these estimates to derive an empirical estimate of prevalence of low GFR in adult population at 6%.

In our simulated population, every subject is examined once to determine the status of kidney function and captured only the current status of renal function at the time of survey. So age (in years) of a subject at examination was considered as the observation time. We generated observation time from truncated Normal (54, 12.73) within the interval (20, 85). Time to renal impairment was generated from an exponential distribution so that rate of the event onset is 6%. Since information on Sex, Diabetes Mellitus and Hypertension were only available; we generated these covariates only through Bernoulli distribution. We generated several data sets sequentially. The finally selected data set was the one where descriptive characteristics were closer to the reported studies. Refer [25].

We applied our cubic spline-based sieve PH model to estimate the hazard of renal impairment and the effect of main risk factors – Sex, Diabetes Mellitus (DM) and Hypertension(HTN). Different number of knots (4, 7, 10, 15), placed at equidistant points between minimum and maximum of observation times, were considered. However, we reported the results of the model with 7 knots only; because it produced the lowest AIC and adding more knots did not improve the model likelihoods.

Table-3:Cox PH Regression model for Renal Impairment. (Spline-based Sieve Estimation)

Risk Factors	Estimates	Hazard Ratio (HR)	95% CI
Sex (Female)	1.15	3.17	2.79 – 4.20
Diabetes Mellitus (Yes)	0.43	1.54	1.57 – 2.61
Hypertension (Yes)	0.70	2.01	1.92 – 3.81

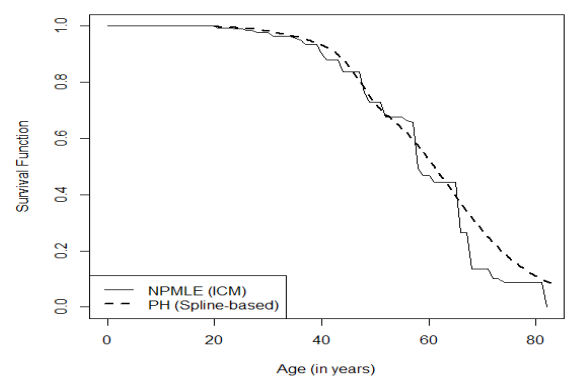

Figure-1: Estimated Cumulative Survival Probabilities of Renal Impairment by Spline-based Sieve Method

Table-3 reports the results of sieve estimation of spline-based Cox PH regression model. Females are thrice as likely to experience renal impairment as their male counterparts (HR=3.17; 95% CI: 2.79-4.20). A diabetic person is at significantly higher risk of developing kidney dysfunction (HR=1.54; 95% CI: 1.57-2.61) and observation conjured in case of a hypertensive is akin to that of diabetics (HR=2.01; 95% CI: 1.92-3.81).

Figure-1 shows NPMLE of survival function using ICM method and the estimated survival function from the cubic monotone spline-based PH model. It shows that our smooth estimates are close to the npmles, except for those aged 70 years and above. This may be because of sparseness of observation for that age group; as our simulation data used average age of individuals as 54 years and also the reported cross-sectional studies from India presented similar scenarios for the elderly population.

5. CONCLUDING REMARKS

This paper has proposed a sieve maximum likelihood estimation procedure for semiparametric proportional hazards

Volume : 3 | Issue : 12 | Dec 2013 | ISSN - 2249-555X

model for current status data. Use of the cubic monotone spline model (4) to approximate $\Lambda_0(t)$ in (2) has produced estimates of survival (hazard) functions, which are smooth, graphically more appealing and meaningfully interpretable. Also, the estimation procedure gets largely simplified due to a finite number of parameters required to define (4). Our simulation studies demonstrated that the proposed method has reasonably satisfactory performance. The choice of the number of knots, m, has been determined by the AIC-type algorithm. However, an adaptive approach may be used allowing unknown number and locations of the knots.

Our estimation is fast and does not require a monotone nonparametric algorithm in every iteration, unlike profile likelihood method[12]; and also yields regression estimators without encountering the problem of slow convergence. Since numerical stability of solutions is of practical importance, our method can be useful from computational perspective.

Our proposed approach was used to estimate the effect of important risk factors of CKD in India. Our survival analysis of simulated data has provided evidence of increased risk of developing chronic renal impairment within various risk groups, which are comparable to the results of other studies. Refer [25]. Use of cubic monotone splines to model the unknown cumulative hazard of renal impairment seems reasonable; as the smooth estimate of survival function, computed using our proposed sieve estimation method, are close to NPMLEs.

REFERENCE

1. Jewel NP, van der Laan MJ. Current status data: review, recent developments and open problems. In: Advances in Survival Analysis. Elsevier, Amsterdam 2004; pp 625-642. | 2. Sun J. The Statistical Analysis of Interval-Censored Failure Time Data. 1st Ed New York: Springer 2006. | 3. Keiding N, Begtrup K, Scheike TH, Hasibeder G. Estimation from Current-Status Data in Continuous Time. Lifetime Data Analysis 1996; 2:119-129. | 4. Mani MK. Prevention of chronic renal failure at the community level. Kidney Int 2003; 63(suppl 83):86-89. | 5. Agarwal SK, Dash SC, Irshad M, Raju S, Singh R, Pandey RM. Prevalence of chronic renal failure in adults in Delhi, India. Nephrol Dial Transplant 2005; 20:1638-1642. | 6. Ayer M., Brunk HD, Ewing GM, Reid WT and Silverman E. An empirical distribution function for sampling with incomplete information. Ann Math Statist 1955; 26:641–647. | 7. Turnbull BW. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. J R Statist Soc B 1976; 38:290–295. | 8. Groeneboom P, Wellner JA. Information Bounds and Nonparametric Maximum Likelihood Estimation. New York: Birkhauser 1992. | 9. Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics 1986; 42:845–854. | 10. Shiboski SC. Generalized Additive models for Current Status Data. Lifetime Data Analysis 1998; 4:29–50. | 11. Mongoué-Tchokoté S, Kim, JS. New statistical software for the proportional hazards model with current status data. Comput Statist and Data Anal 2008; 52:4272–4286. | 12. Huang J. Efficient estimation for the proportional hazards model with interval censoring. Ann Statist 1996; 24:540–568. | 13. Geman A, Hwang C. Nonparametric maximum likelihood estimation by the method of sieves. Ann Statist 1982; 10:401–414. | 14. Rossini A, Tsiatis AA. A semiparametric proportional odds regression model for the analysis of current status data. J Am Statist Assoc 1996; 91:713–21. | 15. Martinussen T, Scheike TH. Efficient Estimation in Additive Hazards Regression With Current Status Data. Biometrika 2002; 89:649-658. | 16. Xue H, Lam KF, Li G. Sieve Maximum Likelihood Estimator for Semiparametric Regression Models with Current Status Data. J Am Statist Assoc 2004; 99(466):346-356. | 17. Ramsay JO. Monotone regression splines in action. Statistical Science 1988; 3:425–441. | 18. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification and stratification. Am J Kidney Dis 2002; 39: 1-266. | 19. Mani MK. Experience with a program for prevention of chronic renal failure in India. Kidney Int 2005; 67(suppl 94): 75-78. | 20. Modi GK, Jha V. The incidence of end-stage renal disease in India: A population-based study. Kidney Int 2006; 70(12):2131-2133. | 21. Singh NP, Ingle GK, Saini VK, Jami A, Beniwal P,Lal M, Meena. Prevalence of low glomerular filtration rate, proteinuria and associated risk factors in North India using Cockcroft-Gault and Modification of Diet in Renal Disease equation: an observational, cross-sectional study. BMC Nephrology 2009; 10:4 (doi:10.1186/1471-2369-10-4). | 22. Dash SC, Agarwal SK. Incidence of chronic kidney disease in India. Nephrol Dial Transplant 2005; 21:232–233. | 23. Varma PP, Raman DK, Ramakrishnan TS, Singh P, Varma A. Prevalence of early stages of chronic kidney disease in apparently healthy central government employees in India. Nephrol Dial Transplant 2010; 25:3011-3017. | 24. Agarwal SK, Srivastava RK. Chronic Kidney Disease in India: Challenges and Solutions. Nephron Clin Pract 2009; 111:197–203. | 25. Grover G, Deka B. Modeling the Risk of Renal Impairment using Current Status Chronic Kidney Disease Data: A Simulation-based Analysis. Submitted for publication in Indian Journal of Applied Research 2013.