Thailand Statistician
July 2007; 5: 81 - 92
http://statassoc.or.th
Contributed paper

Maximum Likelihood Estimator for Semiparametric Transformation Model under General Censorship

Bungon Kumphon*[a], and Prayad Sangngam [b]

[a] Department of Mathematics, Faculty of Science, Mahasarakham University, Mahasarakham 44150, Thailand.

[b] Department of Statistics, Faculty of Science, Silpakorn University, Nakhon Pathom 73000, Thailand.

***Author for correspondence**; e-mail: kbungon@hotmail.com

Received: 2 January 2007
Accepted: 27 April 2007

Abstract

In a semiparametric transformation model, an increasing transformation of the survival time is linearly related to a covariate Z with an error distribution ε . In other words, the survival time T has the property that $\alpha(T) = -\theta z + \varepsilon$ given $Z = z$, where α is an unknown extended real-valued function on R and θ is an unknown constant in $\mathbf{R}^d$. An observation is said to be censored by a general censorship scheme if there are random intervals which, when the observation falls inside them, would hide it. In such cases we get the censoring interval instead of the actual observation. In this paper we consider maximum likelihood estimation of the transformation function α and the regression coefficient θ when the survival time data are subjected to general censorship.

Keyword: censored data, interval censoring, semiparametric model, transformation model.

1. Introduction

Recently, there has been a growing interest in the semiparametric model which is the hybrids between parametric and nonparametric model. For the classical parametric model, it is usually assumed that the dependent variable is functionally dependent on the explanatory or the regressor variable or covariate and unobservable error. One way to examine the relationship between the dependent variable or failure time and the covariate is through a regression model, in which failure time has a distribution that depends upon the covariate. This involves specifying a model for the distribution of survival time, T, given covariate Z. The exponential, Weibull, log-normal and generalised gamma distribution are the most frequently used parametric failure time distribution models. When the circumstances do not support the usage of a fully parametric failure time distribution model, one often turns to a nonparametric or semiparametric method. The main point in using a semiparametric model is that certain of its properties do not depend upon the underlying failure time distribution -e.g. the proportional hazard model due to Cox [2] does not specify the form of the baseline hazard function, $h_0(t)$.

In many situations, it is common to have incomplete data and often, such incomplete observation of the data results from a random censoring mechanism. The type of censoring we consider here, referred to as general censorship, is a generalization of the different types of censoring. Under this scheme some of the data become unobservable when they fall inside a random interval. These intervals can be finite or infinite. Various combinations of finite and infinite intervals give all the different types of censorship such as left censoring, right censoring, double censoring and different cases of interval censoring. For a detailed discussion on this see Jammalamadaka and Mangalam [6].

Let T be the variable of interest, and Z be a covariate, an element of $\mathbf{R}^d$. Let $F(\cdot \mid z) \equiv F_z(\cdot)$ be the distribution function of T given $Z = z$. Let us consider a transformation model

$$F(t \mid z) = \psi(\alpha(t) + \theta z), \qquad (1)$$

where α is an unknown extended real-valued function on $\mathbf{R}$, θ is an unknown constant in $\mathbf{R}^d$ and ψ be the continuous and strictly decreasing function. It is easy to show that (1) is equivalent to the linear transformation model

$$\alpha(T) = -\theta z + \varepsilon, \qquad (2)$$

where ε is a random error with a known distribution. The function α is called the *transformation function* and θ is referred to as the *regression coefficient*. The transformation function α is assumed to satisfy two conditions, α is monotonic increasing and $\lim_{t\to\pm\infty}\alpha(t)=\pm\infty.$ The proportional hazard model and the proportional odds model are special cases of (2) with ε following the extreme-value distribution and the standard logistic distribution, respectively. The generalised odds-rate model also belongs to (2).

Horowitz [5] developed semiparametric estimators of α and θ when the distribution function of the error is unknown. Gorgens and Horowitz [4] extended this technique to censored data. Gorgens [3] developed better estimators of the transformation function and the error distribution. Jammalamadaka and Mangalam [6] provided an algorithm to find the self consistent estimator and showed that the nonparametric maximum likelihood estimator was satisfied the self consistent equation. Our aim in this paper is to estimate the transformation function α and θ when failure times are subjected to general censorship by Maximum Likelihood Method. The motivation for this is that it is a strong generalization of the ordinary regression problem (linear or nonlinear) where the kind of relationship between the variable of interest and the covariate is precisely known. The computation of maximum likelihood estimator is provided in Section 2. Simulation studies in Section 3 and the conclusion of this study is in the last section.

2. The Computation of Maximum Likelihood Estimator

Let $T_i\ ,i=1,\ldots,n$, be a sequence of independent identically distributed (i.i.d.) random variables with distribution function F . Let $\left(U_i,V_i\right),\ \ i=1,\ldots,n$ be the open interval which represent the censoring mechanism, consisting of n pairs of i.i.d. extended real-valued random variables $\left(U_i,V_i\right),$ such that $P\left(U_i<V_i,\forall i\right)=1$. The i th observation is said to be censored if $T_i\in\left(U_i,V_i\right)$ and let $\delta_i=I\left[T_i\notin\left(U_i,V_i\right)\right]$, where $I(\cdot)$ is the indicator function, so that $\delta_i=0$ if the i th observation is censored and $\delta_i=1$ if it is uncensored. We assume that T_i and $\left(U_i,V_i\right),$ are independent given the concomitant variable Z .

The density function of T is given by $f\left(t\right)=\psi'\left(\alpha\left(t\right)+\theta z\right)\alpha'\left(t\right)$, so the likelihood and the log-likelihood functions in the presence of censoring are given by

$$L_n\left(\theta,\alpha\right)=\prod_{i=1}^{n}\left[f\ \left(t_i\right)\right]^{\delta_i}\left[F\ \left(v_i-\right)-F\left(u_i\right)\right]^{1-\delta_i}$$

and

$$\log L_n=\sum_{i=1}^{n}\left\{\begin{array}{l}\delta_i\left(\log\left[\psi'\left(\alpha\left(t_i\right)+\theta z_i\right)\right]+\log\alpha'\left(t_i\right)\right)\\+\left(1-\delta_i\right)\log\left[\psi\left(\alpha\left(v_i-\right)+\theta z_i\right)-\psi\left(\alpha\left(u_i\right)+\theta z_i\right)\right]\end{array}\right\}$$

The value of this expression depends on the function α and its derivative only at the jump points and it can be made arbitrarily large by making $\alpha'\left(t_i\right)$ as large as we want without affecting the values of $\alpha\left(t_i\right)$. Consequently, we work with a discretized version of the likelihood function where $\alpha'\left(t\right)$ is replaced by a jump size a_i and $\alpha\left(t_i\right)$ by $A_i=\sum_{k=0}^{i}a_k$. Thus we find the function that maximises this modified log-likelihood among all α's such that α is an increasing step function that is a constant for $t<t_1$ and jumps at t_i.

We replace any empty censoring interval (a censoring interval that contains no uncensored observations) by its midpoint as an uncensored observation, group the censored and uncensored observations separately and reorder the uncensored observation in the ascending order. Let n_1 be the size of the exact data and $n_2=n-n_1$ be the size of the censored data. Let $a_0=\alpha\left(t_1-\right)$ and a_i be the jump size at t_i for $i=1,\ldots,n_1$. Then the modified version of the log-likelihood is given by

$$l_n\left(\theta,a\right)=\sum_{i=1}^{n_1}\left\{\log\left[\psi'\left(A_i+\theta z_i\right)\right]+\log a_i\right\}$$
$$+\sum_{j=1}^{n_2}\log\left[\psi\left(\sum_{k:t_{(k)}\le v_j}a_k+\theta z_j\right)-\psi\left(\sum_{k:t_{(k)}\le u_j}a_k+\theta z_j\right)\right]$$

which is to be maximised under the constraint that all of the a_i's except a_0 are non-negative. If $\hat{a}_i$, $i=0$ to n_1 maximises $l_n\left(\theta,a\right)$, then the function $\hat{\alpha}$ defined as an increasing step function with jumps $\hat{a}_i$ at t_i and value a_0 for $t<t_1$ is the MLE of α.

Let $\Theta \subset \mathbf{R}^d$ be the d-dimensional parameter space of θ. Let θ_0 and α_0 denote the true value of the parameters and let F_0 be the true conditional distribution function of T given Z. The MLE's of α and θ, $\hat{\alpha}$ and $\hat{\theta}$, are obtained by maximising $l_n(\theta, a)$ over $\Theta \times A$, where $A = \mathbf{R} \times \mathbf{R}^{+n_1}$, the set of $n_1 + 1$ dimensional vectors whose coordinates are all non-negative except the first. Under some mild assumptions, it can be shown that $l_n(\theta, a)$ is strictly concave for each n as in proposition 2.1, and that it is bounded above. It therefore has a unique maximiser, and the maximiser $(\hat{\theta}, \hat{a})$ can be obtained by equating the first derivative to zero and using the multivariate Newton-Raphson algorithm.

Proposition 2.1 *Let $\psi(t)$ be a differentiable distribution function on $\mathbf{R}$. If $\log(\psi'(t))$ is strictly concave in t, then $\log(\psi(v) - \psi(u))$ is strictly concave in u and v.* Proof is shown in appendix**.**

3. Numerical Simulation

A simulation study was performed to measure the performance of maximum likelihood estimator. The regression coefficient θ is estimated by $\hat{\theta}_n$, the transformation function α is estimated by $\hat{\alpha}_n = \hat{A}$ and the distribution function F by $\hat{F}_n = \psi(\hat{\alpha}(t) + \hat{\theta}z)$. The performance of $\hat{F}_n$ is more important than $\hat{\alpha}_n$ (see Cheng [1]) and can be measured by $\left\| \hat{F}_n - F_0 \right\| = \max_t \left\| \hat{F}_n(t) - F_0(t) \right\|$. The model for this study was the proportional hazards model where ε has a standard extreme value distribution yielding $P(\varepsilon < t) = 1 - \exp(-\exp(t))$. The dimension of the regression coefficient θ is taken to be 1, the true values of θ, $\theta_0 = 0$, is considered and the true transformation function is chosen to be $\alpha_0(t) = \log t$. The values of the covariate Z are randomly generated from a standard normal distribution. The random variables U_i and V_i are produced according to the level of censoring, which is set by letting U_i and V_i as the minimum and the maximum of c independent exponential random variables, for $c = 2$. The sample size n is set to be 100 and each combination of all these factors was replicated 300 times. The implementations of data generation and computation were written in the statistical software package S-Plus. The non-linear optimisation routine *Nonlinear Minimization subject to Box Constraint* (NLMINB) was used for numerical maximisation.

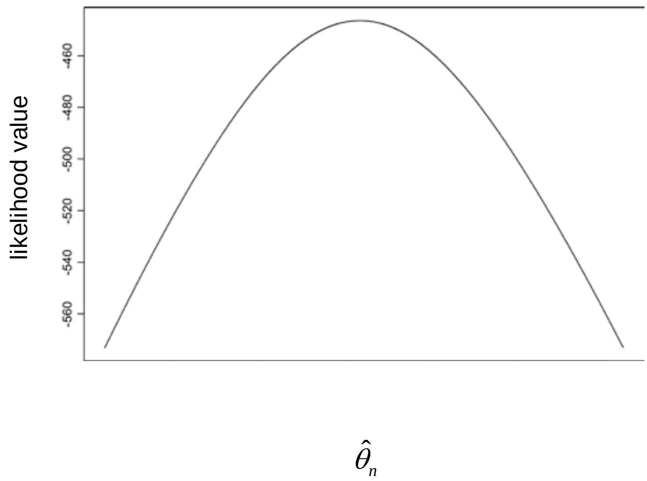

Figure 1. Log-likelihood function for a simulation data.

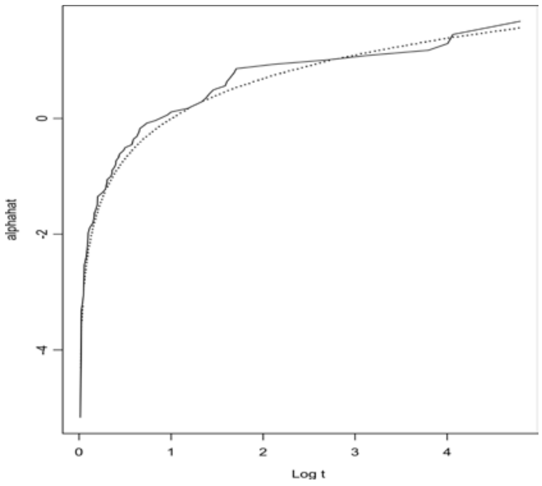

Figure 2. The estimated value of α versus the true value of α .

Figure 1 shows the log-likelihood function versus the regression coefficient $\hat{\theta}_n$. It maximised at $\hat{\theta}_n$ = 0.0032 which is the value of our maximum likelihood estimator and the estimated value of α versus the true value of α is shown in Figure 2. The solid line is the estimated value of α and the dotted line is the true value of α. The results show that 30.89 % of them being censoring, the estimated value $\hat{\theta}_n$ is 0.0032 with variance 0.0120 and the maximum distance $\left\|\hat{F}_n - F_0\right\|$ = 0.0920 which are fairly good.

We also tried out various other values of $\alpha_0(t) = t,\ \theta_0 = 1,$ different level of censoring with $c = 1, 2, 3$ and 5, and the genearlised odds-rate (GOR) model with coefficient λ equal 0.5 ($P(\varepsilon < t) = 1-(1+\lambda\exp(t))^{-\frac{1}{\lambda}}$, $\lambda > 0$). Note that when λ equal 1, the GOR model gives the proportional odds model. The results are reported in Table 1 and Table 2. In all cases, the conclusions were similar. As for the performance of $\hat{F}_n$, the values of $\left\|\hat{F}_n - F_0\right\|$ performed fairly well but they were influenced by the rate of censoring. As the rate of censoring increased, the error increased, but continued to be within acceptable limits. Hence our method of parameter estimation of both θ and F performs fairly well.

4. Conclusions

Maximum likelihood estimation for the regression coefficients and the transformation function are carried out for a semiparametric transformation model where survival time data are subjected to general censorship. The multivariate Newton-Raphson method was used for optimization and in our entire simulation studies global maximum was attained. The ML estimates performed fairly well in the sense that the estimated values were close to the true values of the parameter.

Table 1. Simulation results for the estimated values of $\alpha_0\left(t\right) = \log t$ in the GOR model with $\lambda = 0.5$.

$\theta_0 = 0$	**Percent of censoring**							
	r=0		**r=29.35**		**r=44.69**		**r=59.17**	
	mean	**var**	**mean**	**var**	**mean**	**var**	**mean**	**var**
$\hat{\theta}_n$	0.0054	0.0210	- 0.0035	0.0223	0.0027	0.0269	0.0155	0.0319
$\left\Vert \hat{F}_n - F_0 \right\Vert$	0.0814	0.0007	0.0915	0.0008	0.1170	0.0016	0.2128	0.0065
$\theta_0 = 1$	**Percent of censoring**							
	r=0		**r=26.00**		**r=36.14**		**r=52.35**	
	mean	**var**	**mean**	**var**	**mean**	**var**	**mean**	**var**
$\hat{\theta}_n$	1.0374	0.0346	1.0248	0.0303	1.0270	0.0336	1.0058	0.0357
$\left\Vert \hat{F}_n - F_0 \right\Vert$	0.0895	0.0009	0.1010	0.0011	0.1267	0.0019	0.2278	0.0071

Table 2. Simulation results for the estimated values of $\alpha_0\left(t\right) = t$ in the GOR model with $\lambda = 0.5$.

$\theta_0 = 0$	**Percent of censoring**							
	r=0		**r=14.96**		**r=22.67**		**r=29.23**	
	mean	**var**	**mean**	**var**	**mean**	**var**	**mean**	**var**
$\hat{\theta}_n$	0.0016	0.0231	0.0058	0.0210	- 0.0023	0.0236	- 0.0005	0.0221
$\left\Vert \hat{F}_n - F_0 \right\Vert$	0.0783	0.0007	0.0869	0.0007	0.0978	0.0009	0.1517	0.0021
$\theta_0 = 1$	**Percent of censoring**							
	r=0		**r=16.36**		**r=23.98**		**r=31.49**	
	mean	**var**	**mean**	**var**	**mean**	**var**	**mean**	**var**
$\hat{\theta}_n$	1.0315	0.0295	1.0255	0.0340	1.0315	0.0339	0.9818	0.0337
$\left\Vert \hat{F}_n - F_0 \right\Vert$	0.0857	0.0008	0.0901	0.0008	0.1022	0.0009	0.1570	0.0022

Appendix

Proof of the Proposition.

Proposition 2.1 *Let* $\psi(t)$ *be a differentiable distribution function on* $\mathbf{R}$ *. If* $\log(\psi'(t))$ *is strictly concave in* t *, then* $\log(\psi(v)-\psi(u))$ *is strictly concave in* u *and* v *.*

Proof: Let $g(t)=\log(\psi'(t))$. Then g is strictly concave by the assumption and $\psi'(t)=e^{g(t)}$. Now,

$$\psi(v)-\psi(u)=\int_u^v e^{g(x)}dx$$

$$\frac{\partial}{\partial u}\log(\psi(v)-\psi(u))=\frac{-\psi'(u)}{\psi(v)-\psi(u)}$$

$$\begin{aligned}\frac{\partial^2}{\partial u^2}\log(\psi(v)-\psi(u)) &= -\frac{[\psi(v)-\psi(u)]\psi''(u)+(\psi'(u))^2}{[\psi(v)-\psi(u)]^2} \\ &= -\frac{e^{g(u)}}{\int_u^v e^{g(x)}dx}\left[g'(u)+\frac{e^{g(u)}}{\int_u^v e^{g(x)}dx}\right]\end{aligned}$$

In order to show strictly concavity, we need to show that

$$g'(u)+\frac{e^{g(u)}}{\int_u^v e^{g(x)}dx}>0$$

If $g'(u)\geq 0$ there is nothing to show, so assume $g'(u)<0$

For $x>u$,

$$\frac{g(x)-g(u)}{x-u}=g'(x^*) \qquad \text{for some } x^*\in(u,x).$$

As g is strictly concave, g' is strictly decreasing and hence $g'(x^*)<g'(u)$.

Therefore,

$$g(x) \quad < \quad g(u)+g'(u)(x-u) \quad \text{for some} \quad x>0.$$

$$\int_u^v e^{g(x)}dx \quad < \quad \int_u^v e^{g(u)+g'(u)(x-u)}\,dx$$

$$= \quad \frac{e^{g(u)}\left[e^{(v-u)g'(u)}-1\right]}{g'(u)}$$

$$< \quad \frac{e^{g(u)}}{g'(u)}$$

Consequently,

$$g'(u)+\frac{e^{g(u)}}{\int_u^v e^{g(x)}dx}>0 \quad \text{as} \quad g'(u) \quad < \quad 0$$

Thus $\log\left(\psi(v)-\psi(u)\right)$ is strictly concave in u.

Now, we will prove that $\log\left(\psi(v)-\psi(u)\right)$ is strictly concave in v.

$$\frac{\partial}{\partial v}\log\left(\psi(v)-\psi(u)\right)=\frac{\psi'(v)}{\psi(v)-\psi(u)}$$

$$\frac{\partial^2}{\partial v^2}\log\left(\psi(v)-\psi(u)\right) \quad =\frac{\left[\psi(v)-\psi(u)\right]\psi''(v)-\left(\psi'(v)\right)^2}{\left[\psi(v)-\psi(u)\right]^2}$$

$$=\frac{e^{g(v)}}{\int_u^v e^{g(x)}dx}\left[g'(v)-\frac{e^{g(v)}}{\int_u^v e^{g(x)}dx}\right]$$

In order to show strictly concavity, we need to show that

$$g'(v)+\frac{e^{g(v)}}{\int_u^v e^{g(x)}dx}<0$$

If $g'(v) \le 0$, there is nothing to show, so assume $g'(v) > 0$.

For $x < v$,

$$\frac{g(v)-g(x)}{v-x} = g'\left(x^*\right) \text{ for some } x^* \in (x, v).$$

As g is a strictly concave, g' is strictly decreasing and hence $g'\left(x^*\right) > g'(v)$.

Therefore,

$$g(x) < g(v) - g'(v)(v-x) \quad \text{for some } x > 0$$

$$\int_u^v e^{g(x)} dx \quad < \quad \int_u^v e^{g(v)-g'(v)(v-x)} dx$$

$$= \quad \frac{e^{g(v)}\left[1-e^{-(v-u)g'(v)}\right]}{g'(v)}$$

$$< \quad \frac{e^{g(v)}}{g'(v)}$$

Consequently,

$$g'(v) + \frac{e^{g(v)}}{\int_u^v e^{g(x)} dx} < 0 \quad \text{as } g'(v) \quad > \quad 0.$$

Thus $\log\left(\psi(v) - \psi(u)\right)$ is strictly concave in v.

□

References

[1] Cheng, Y.-C., *Estimation in semiparametric transformation models with doubly censored data,* Ph.D. thesis, Department of Statistics, Rutgers University, 2002.

[2] Cox, D. R., Regression models and life tables (with discussion), *J. Roy. Statist. Soc. Ser. B*,1972; 34: 187-220.

[3] Gorgens, T., Semiparametric estimation of censored transformation models, *J. Nonparametr. Statist,* 2003; 15: 377-393.

[4] Gorgens, T., and Horowitz, J.L., Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable, *Econometrica,* 1999; 90: 155-191.

[5] Horowitz, J.L., Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, *Econometrica,* 1996; 64: 103-137.

[6] Jammalamadaka, S.R., and Mangalam, V., Nonparametric estimation for middle censored data, *J. Nonparametr. Statist,* 2003; 15: 253-265.