ISSN 1 746-7233, England, UK
World Journal of Modelling and Simulation
Vol. 5 (2009) No. 1, pp. 57-62

Exponential semiparametric regression models under random censorship*

Junqiang Yang[1,2†]

[1] Hunan Urban Construction College
[2] School of Mathematics and Computational Science, Xiangtan University, Xiangtan, 411101, China

(*Received December 15 2007, Accepted May 13 2008*)

Abstract. Using the weighted maximum likelihood method, we propose a consistent estimation of parametric portion and nonparametric portion in exponential semiparametric regression models under random censorship. A small Monte Carlo study is carried out to examine the proposed estimation method.

Keywords: semiparametric regression model, weighted maximum likelihood, censoring

1 Introduction

In a regression model the objective is to describe the relationship between a response variables and a set of covariates, the typically used models are parametric or nonparametric models. However, when a model includes both a parametric and a nonparametric components, a semiparametric model is needed. In semiparametric models, the parametric portion includes covariates for which information concerning the functional form of the response–covariate relationship is available and thus used; the nonparametric portion includes covariates where less information is known concerning the functional form of the response–covariate relationship and thus the functional form is not specified. Because if only partial information concerning the functional form of the response–covariate relationship is available, then a completely nonparametric model is inefficient and a completely parametric model may be wrong, it is important to combine the parametric portion and the nonparametric portion in a model.

Sevirini and Staniswalis (1994)[6] used a quasi–likelihood function to estimate the parameters in a semiparametric model. This method of estimation only requires specification of the second–moment properties of the data, rather than specification of the entire distribution.

Hunsberger (1994)[2] used a weighted likelihood (Staniswalis, 1989)[9], sometimes termed a local likelihood (Hastie, 1986)[1], to show that there exists a sequence of consistent estimators for the parametric and nonparametric components of the semiparametric regression model for arbitrary but specified densities of the observations, asymptotic normality and consistency for these estimators are established.

In survival analysis, it is important to consider the relationship of lifetime to other factors. One way to do this is through regression models, in which the dependence of lifetime on concomitant variables is explicitly recognized. Lawless (1982)[3] discussed the parametric regression models for lifetime distribution in detail, but if the relationship between a lifetime and a set of concomitant variables can not be described by the parametric regression model, we should think about the nonparametric regression models and the semiparametric regression models. In this paper, we only discuss the exponential semiparametric regression models, further study on other models will be developed in our future work.

* Research Supported by the National Natural Science Foundation of China (50675185), the Research Fund for the Doctoral Program of China (20070530003).
† Corresponding author. *E-mail address*: xpyjq1668@163.com.

2 Semiparametric exponential regression models

When individuals have constant hazard functions that may depend on concomitant variables, an exponential regression model is appropriate. Parametric exponential regression models have been studied by Glasser (1967), Cox and Snell (1968), Prentice (1973), Lawless (1976,1982), Feigl and Zelen (1965), Zippin and Armitage (1966), Greenberg et al. (1974), and others. For example, the p.d.f. of Z, given x, is

$$f(z|x) = \theta^{-1}(x)\exp(\frac{-z}{\theta(x)}) \quad z \geq 0.$$

In parametric regression models, the most useful functional form for $\theta(x)$ is

$$\theta(x) = \exp(x\beta).$$

Where x and β can be vectors, respectively. The ordinary maximum likelihood method can be relied on to estimate the parameters β.

In this paper, we will discuss semiparametric exponential regression models. The survivor function and p.d.f. of Z, given x and t, are assumed to be

$$S(z|x,t) = \exp\left[-(\frac{z}{\theta(x,t)})\right], \quad t \geq 0, \tag{1}$$

and

$$f(z|x,t) = \theta(x,t)^{-1}\exp(\frac{-z}{\theta(x,t)}), \quad t \geq 0. \tag{2}$$

Here x and t are regression variables and

$$\theta(x,t) = E(Z|x,t) = \exp(x\beta + g(t)).$$

For simplification, we assume $x \in \Re$ and $t \in \Re$, β is a scale.

The model (2) is a proportional hazards model. In addition,it can be viewed as a location–scale model for $Y = \log(Z)$. From (2) the p.d.f of Y, given x and t, is

$$f(y|x,t) = \exp[\{y - (x\beta + g(t))\} - \exp\{y - (x\beta + g(t))\}], \quad -\infty < y < \infty. \tag{3}$$

Alternately, we can write

$$y = x\beta + g(t) + \epsilon, \tag{4}$$

where ϵ has a standard extreme value distribution with p.d.f equal to $\exp(s - \exp(s)), -\infty < y < \infty$.

3 The weighted maximum likelihood method

Suppose that associated with each individual is a lifetime or censoring time Z_i and a regression vector (x_i, t_i), the notation $\delta_i = 1$ and $\delta_i = 0$ will be used to refer to individual i for which Z_i is a lifetime and a censoring time, respectively. We work with log times, $Y_i = \log(Z_i)$, log lifetime Y has p.d.f and survivor functions

$$f(y|x,t) = \exp\{[y - (x\beta_0 + g(t))] - \exp[y - (x\beta_0 + g(t))]\} \tag{5}$$

and

$$S(y|x,t) = \exp\{-\exp[y - (\beta_0 + g(t))]\}, \tag{6}$$

respectively, where β_0 is the true parameter value.

The likelihood function for a censored sample based on n individuals is

$$L(\beta,\theta)=\prod_{i=1}^{n}[f(y_i|x_i,t_i)]^{\delta_i}[S(y_i|x_i,t_i)]^{1-\delta_i}.$$

As discussed by Hunsberger (1994), let the parameter $\lambda_i = x_i\beta_0 + g(t_i)$, then $x\beta_0$ is the parametric portion, with β_0 being the unknown parameter to be estimated that relates the covariate x to the response. Here g is the nonparametric portion of the model, with the only assumption on g that it be a smooth function of t with $\nu \geq 2$ continuous derivatives. Several assumptions are made that allow an association between x and t (Rice 1986; Speckman 1988). Assume the regression model $x_i = r(t_i) + \eta_i$ where $r(t)$ is a smooth function with ν continuous derivatives and η_i are independent random error terms with $E[\eta_i] = 0$ and $E[\eta_i^2] = \sigma^2$. Now λ_i can be rewritten using the model for the x's to obtain $\lambda_i = \eta_i\beta_0 + h(t_i)$, where $h(t_i) = r(t_i)\beta_0 + g(t_i)$ is the portion that depends on t. The main result of this research is to estimate β_0 and $h_i = h(t_i)$ $(i = 1,\ldots,n)$ in the exponential semiparametric regression model by maximizing the weighted likelihood function

$$\begin{aligned}WL(\beta,\theta)&=\frac{1}{n}\sum_{i=1}\sum_{j=1}\{w(\frac{t_i-t_j}{b})/\sum_{j=1}^{n}w(\frac{t_i-t_j}{b})\}\{\log[f(y_j;\beta,\theta_i)]^{\delta_j}[1-F(y_j;\beta,\theta_i)]^{1-\delta_j}\}\\&=\frac{1}{n}\sum_{i}WL(\beta,\theta_i)\end{aligned}\tag{7}$$

with respect to β and θ, where $\theta = [\theta_1,\ldots,\theta_n]'$. Here θ_i is used to indicate the function of parameters of h_i. Expressions are given in terms of log lifetime Y and its $p.d.f.$ (5). Suppose that independent observations $(y_i, x_i, t_i), i = 1,\ldots,n$ are available, where y_i is either a log lifetime or a log censoring time; $\delta_i = 1$ and $\delta_i = 0$ denote the individuals for which y_i is a log lifetime and a log censoring time, respectively.

Throughout this article the sum is assumed to be from 1 to n. $WL(\beta,\theta)$ depends on the unobserved η_i's, which can be estimated. In $WL(\beta,\theta_i)$, $w(.)$ is a kernel that assigns zero weights to the observations Y_j that correspond to t_j outside a neighborhood of t_i. The neighborhood is defined by the bandwidth b. The estimates of β_0 and $h_i = h(t_i)$ $(i = 1,\ldots,n)$ are found by choosing the $\hat{\beta}$ and $\hat{h_i}$ to simultaneously maximize $WL(\beta,\theta)$ with respect to β and θ.

To understand the motivation for the weighted likelihood function, one can refer to the approaches of Staniswalis (1989) and Hunsberger (1994). First we examine $WL(\beta,\theta_i)$, this can be seen as the portion estimating $h_i = h(t_i)$ by using the Nadaraya–Watson estimator. The kernel governs those observations are used to estimate $h_i = h(t_i)$. That is, because only the observations Y_i with t_j close to t_i have information about h_i, only the y_j close to the t_i of interest are used to estimate h_i. The summation over i uses all of the individual $WL(\beta,\theta_i)$ to estimate β_0, because all of the observations contain information about β_0. The weighted likelihood function can be written as

$$\begin{aligned}WL(\beta,\theta)&=\frac{1}{n}\sum_{i}\sum_{j}w_{i,j}[\delta_j\log f(y_j|\eta_j,t_i)+(1-\delta_j)\log S(y_j|\eta_j,t_i)]\\&=\frac{1}{n}\sum_{i}\sum_{j}w_{i,j}\{\delta_j[y_j-(\eta_j\beta+h(t_i))]-\delta_j\exp[y_j-(\eta_j\beta+h(t_i))]\\&-(1-\delta_j)\exp[y_j-(\eta_j\beta+h(t_i))]\}\\&=\frac{1}{n}\sum_{i}\sum_{j}w_{i,j}\{\delta_j[y_j-(\eta_j\beta+h(t_i))]-\exp[y_j-(\eta_j\beta+h(t_i))]\},\end{aligned}\tag{8}$$

where $w_{i,j} = w(\frac{t_i-t_j}{b})/\sum_{i=1}^{n}w(\frac{t_i-t_j}{b})$. A Newton–Raphson algorithm is used to approximate $\hat{\beta}$ and $\hat{h}$. Now η is unobservable but can be estimated as follows: $\hat{\eta} = x - \hat{r}(t)$, with $\hat{r}(t)$ being the nonparametric kernel estimate of $r(t)$. In this paper the Nadaraya–Watson estimator is used. This is defined as

$$\hat{r}(t,b)=\sum_{i}w(\frac{t-t_i}{b})X_i/\sum_{i}w(\frac{t-t_i}{b}).$$

The first and second derivatives of $WL(\beta,\theta)$ with respect to β and θ are

$$\frac{\partial WL(\beta,\theta)}{\partial\beta} = \frac{1}{n}\sum_i\sum_j w_{i,j}\{-\delta_j\eta_j + \eta_j\exp[y_j - (\eta_j\beta + h(t_i))]\}, \tag{9}$$

$$\frac{\partial WL(\beta,\theta)}{\partial\theta_i} = \frac{1}{n}\sum_i\sum_j w_{i,j}\{-\delta_j + \exp[y_j - (\eta_j\beta + h(t_i))]\}, \tag{10}$$

$$\frac{\partial^2 WL(\beta,\theta)}{\partial\beta^2} = \frac{1}{n}\sum_i\sum_j w_{i,j}\{-\eta_j^2\exp[y_j - (\eta_j\beta + h(t_i))]\}, \tag{11}$$

$$\frac{\partial^2 WL(\beta,\theta)}{\partial\theta_i^2} = \frac{1}{n}\sum_i\sum_j w_{i,j}\{-\exp[y_j - (\eta_j\beta + h(t_i))]\}, \tag{12}$$

$$\frac{\partial^2 WL(\beta,\theta)}{\partial\beta\theta_i} = \frac{1}{n}\sum_i\sum_j w_{i,j}\{-\eta_j\exp[y_j - (\eta_j\beta + h(t_i))]\}. \tag{13}$$

The maximum likelihood equations

$$\frac{\partial WL(\beta,\theta)}{\partial\beta} = 0, \tag{14}$$

and

$$\frac{\partial WL(\beta,\theta)}{\partial\theta_i} = 0, \qquad (i = 1,\ldots,n) \tag{15}$$

are readily solved by the Newton–Raphson method.

To estimate β and h, another approach can be used here. First, for fixed β we estimate h as a function of β to obtain $\hat{h}_\beta$; note that this is a nonparametric estimation problem. Then, setting $h = \hat{h}_\beta$, we estimate the parametric component β; note that this is a parametric problem. Hence this approach to estimation in the semiparametric model effectively separates the estimation problem into parametric and nonparametric component (see Severini and Wong, 1992 and Severini and Staniswalis, 1994). For each fixed t and β, $\hat{h}(t_i)$, the estimator of $h(t_i)$, is obtained by solving (10).

$$\hat{h}(t_i) = -\log\left[\frac{\sum_j w_{i,j}\delta_j}{\sum_j w_{i,j}\exp(y_j - \eta_j\beta)}\right]. \tag{16}$$

Hence

$$\frac{\partial\hat{h}(t_i)}{\partial\beta} = -\frac{\sum_j w_{i,j}\eta_j\exp(y_j - \eta_j\beta)}{\sum_j w_{i,j}\exp(y_j - \eta_j\beta)}. \tag{17}$$

In (14), let θ is replaced by $\hat{\theta}(\beta)$, the estimator of θ, then we can obtain the estimator of β by solving

$$\frac{\partial WL(\beta,\hat{\theta}(\beta))}{\partial\beta} = 0. \tag{18}$$

Since

$$\frac{\partial}{\partial\beta}\left(\frac{\partial WL(\beta,\hat{\theta}(\beta))}{\partial\beta}\right) = \frac{1}{n}\sum_i\sum_j w_{i,j}\{\eta_j\exp[y_j - (\eta_j\beta + \hat{h}(t_i))](-\eta_j + \frac{\partial\hat{h}(t_i)}{\partial\beta})\}, \tag{19}$$

(18) can be solved by the Newton–Raphson procedure to get the estimator of β:

$$\beta = \beta_g - \left[\frac{\partial}{\partial\beta}\left(\frac{\partial WL(\beta,\hat{\theta}(\beta))}{\partial\beta}\right)\right]^{-1}\left[\frac{\partial WL(\beta,\hat{\theta}(\beta))}{\partial\beta}\right]\Big|_{\beta=\beta_g}. \tag{20}$$

4 Simulation

A small simulation study was conducted to study the finite sample properties of the $\hat{\beta}$ and $\hat{h}$ in the semiparametric model defined in Section 2. The standard extreme random number is generated by using the uniform random number generator RANUNI in SAS and the following transformation:

$$\epsilon = \log[-\log(1-u)],$$

according to the model (4): $Y = x\beta + g(t) + \epsilon$. If $u \sim U[0,1]$, then $\epsilon \sim$ standard extreme distribution with p.d.f equal to $\exp(s - \exp(s))$, $-\infty < s < \infty$.

In this simulation, the t_i's are equally spaced as $t_i = i/60$, for $i = 1, \ldots, 60$. A single Monte Carlo realization consists of $n = 60$ observations. For the generated data set, $x = r(t) + \eta$, $r(t) = 1$, $\eta \sim N(0, 0.1^2)$, $\beta = 10$ and $g(t) = 6(1-3t)^2$, hence, $h(t) = r(t)\beta + g(t)$, the model (4) becomes

$$Y = \eta\beta + h(t) + \epsilon.$$

The simulated censoring random variable U was uniform on [0, 80], resulting in about 22% censoring of the generated data.

A kernel and a bandwidth must be chosen to use in the weighted likelihood. The quadratic kernel of Müller's (1984) with $\nu = 2$ was used through:

$$\begin{aligned} w(v) &= \frac{15}{16}(1-v)^2, \quad -1 \le v \le 1 \\ &= 0, \quad \text{elsewhere} \end{aligned}$$

and $b = 0.6$ for estimating $r(t)$ and $b = 0.05$ for estimating $h(t)$ were used, respectively.

The simulation shows that the MWLE method is estimating $\beta_0 = 10$ and $h(t)$ well. We obtain $\hat{\beta} = 10.028$ using the SAS procedure PROC IML. Fig. 1 plots the estimates of the function $h(t)$ versus t, it indicates the estimated curve captures the true curve closely.

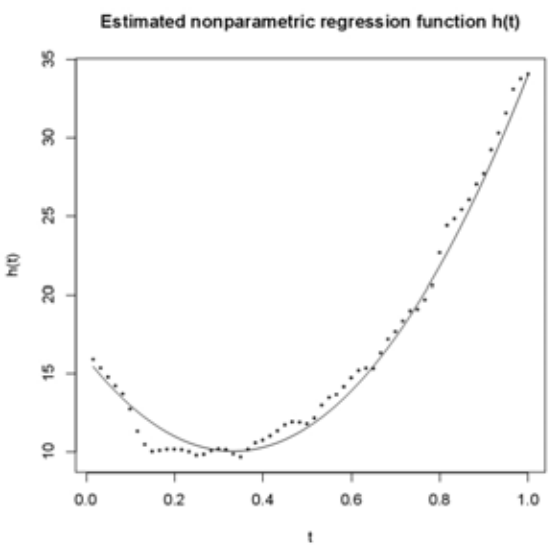

Fig. 1. Plot for the estimated nonparametric function $h(t)$ from the simulation study. The dotted line is the estimated function, the solid line is the true function.

References

[1] T. Hastie, R. Tibshirani. Generalized additive models. *Statistical Science*, 1986, **1**: 297–318.

[2] S. Hunsberger. Semiparametric Regression. *Journal of the American Statistical Association*, 1354–1365.
[3] J. Lawless. *Statistical Models and Methods for Lifetime Data*. John Wiley, 1982.
[4] H. Muller. Smooth optimum kernel estimators of densities, regression curves, and modes. *The Annals of Statistics*, 1984, **12**: 766–774.
[5] J. Rice. Convergence rates for partially splined models. *Statistics and Probability Letters*, 1986, **4**: 203–209.
[6] T. Severini, J. Staniswalis. Quasi–Likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association*, 1994, **89**: 501–511.
[7] T. Severini, W. Wong. Generalized profile likelihood and conditionally parametric models. *The Annals of Statistics*, 1992, **20**: 1768–1802.
[8] P. Speckman. Kernel, Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society*, 1988, **50**: 413–436.
[9] J. Staniswalis. The kernel estimate of a regression function in Likelihood–Based models. *Journal of the American Statistical Association*, 1989, **84**: 276–283.