ISSN 2347-1921

WEIBULL SEMIPARAMETRIC REGRESSION MODELS UNDER RANDOM CENSORSHIP

E.A. Rady, M.M.E. Abd El-Monsef and A.M. Sobhy
ISSR, Cairo University
elhoussainy@yahoo.com
Faculty of Science, Tanta University
mmezzat@science.tanta.edu.eg
Faculty of Science, Tanta University
ayat.mohamed@science.tanta.edu.eg

ABSTRACT

Semiparametric regression is concerned with the flexible combination of non-linear functional relationships in regression analysis. The main advantage of the semiparametric regression models is that any application benefits from regression analysis can also benefit from the semiparametric regression.In this paper, we derived a consistent estimator of parametric portion and nonparametric portion in Weibull semi-parametric regression models under random censorship.

KEYWORDS:

Semi-parametric regression model; weighted maximum likelihood; censoring.

ISSN 2347-1921

1. INTRODUCTION

Regression models can be either linear or nonlinear. A linear model assumes the relationships between variables are straight-line relationships, while a nonlinear model assumes the relationships between variables are represented by curved lines. The typically used models are parametric or nonparametric models, when a model includes both a parametric and nonparametric components, a semiparametric model is needed. Variable selection for semiparametric regression models consists of two components: model selection for nonparametric components and select significant variables for parametric portion. Thus, it is much more challenging than that for parametric models such as linear models and generalized linear models because traditional variable selection procedures including stepwise regression and the best subset selection require model selection to nonparametric components for each submodel.Thus,semiparametric models are intermediate between parametric and nonparametric models: they are larger than parametric models, but smaller than nonparametric models. Because if only partial information concerning the functional form of the responsecovariate 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[6]used a quasi–likelihood function to estimate the parameters in a semiparametric model. This method of estimation only requires specification of the secondmoment properties of the data, rather than specification of the entire distribution. Hunsberger[2] used a weighted likelihood (Staniswalis, [8]), sometimes termed a local likelihood (Hastie [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. A popular regression model for the analysis of survival data is the Cox proportional hazards regression model. It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest. The Cox regression model is a semiparametric model, making fewer assumptions than typical parametric methods but more assumptions than those nonparametric methods described above. Observations are called censored when the information about their survival time is incomplete; the most commonly encountered form is right censoring. Lawless [4] discussed the parametric regression models for lifetime distribution in detail, but if the relationship between a lifetime and a set of concomitant variables cannot 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 Weibullsemiparametric regression models, further study on other models will be developed in our future work.

2. SEMI-PARAMETRIC WEIBULL REGRESSION MODELS

When individuals have constant hazard functions that may depend on concomitant variables, a Weibullregression model is appropriate. Parametric Weibull regression models have been studied by Lawless [4]and Khokan, Bari and Khan [3]. Suppose the life times are assumed to be Weibull distributed with p.d.f. ofZ, given x, is

$$\mathrm{f(z/x)} = \frac{\gamma}{\theta(\mathrm{x})}\left(\frac{\mathrm{z}}{\theta(\mathrm{x})}\right)^{\gamma-1} \exp\left[-\left(\frac{\mathrm{z}}{\theta(\mathrm{x})}\right)^{\gamma}\right], z \geq 0$$

is in parametric regression models, the most useful functional form for$\theta(\mathrm{x}) = \exp(\mathrm{x}\beta)$, where x and β can be vectors, respectively. The ordinary maximum likelihood method can be relied on to estimate the parameters β.

Suppose the life times are assumed to be Weibull distributed with c.d.f. ofZ, given x, is

$$F_z(z/x) = 1 - \exp\left(-\left(\frac{z}{\theta(x)}\right)^{\gamma}\right)$$

andthe survival function

$$S(z/x) = 1 - F_z(z/x) = 1 - \left[1 - \exp\left(-\left(\frac{z}{\theta(x)}\right)^{\gamma}\right)\right] = \exp\left(-\left(\frac{z}{\theta(x)}\right)^{\gamma}\right)$$

We will discuss semiparametricWeibull regression models. The survival function and p.d.f. of Z, given x and t, are assumed to be

$$S(z/x,t) = \exp\left(-\left(\frac{z}{\theta(x,t)}\right)^{\gamma}\right), \quad t \geq 0 \qquad (1)$$

and

$$f(z/x,t) = \frac{\gamma}{\theta(x,t)}\left(\frac{z}{\theta(x,t)}\right)^{\gamma-1} exp\left[-\left(\frac{z}{\theta(x,t)}\right)^{\gamma}\right], \quad t \geq 0 \qquad (2)$$

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 $t \in \Re, \beta$ is a $scale$.

ISSN 2347-1921

The model (2) is 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

$$f(y/x,t) = \frac{\gamma}{\theta(x,t)}\left(\frac{e^y}{\theta(x,t)}\right)^{\gamma-1}\exp\left[-\left(\frac{e^y}{\theta(x,t)}\right)^{\gamma}\right]$$

$$f(y/x,t) = \gamma\exp\left[\gamma[y-(x\beta+g(t)]-e^{\gamma[y-(x\beta+g(t)]}\right], -\infty<y<\infty \quad (3)$$

and survival function

$$S(y/x,t) = exp\left(-\frac{e^y}{e^{x\beta+g(t)}}\right)^{\gamma} = exp\left(-e^{\gamma\left(y-(x\beta+g(t))\right)}\right)$$

Alternately, we can write

$$y = x\beta + g(t) + \varepsilon \quad (4)$$

whereε has a standard extreme value distribution with p.d.f. to

$$\gamma\exp(\gamma s - \exp(\gamma s)), -\infty < y < \infty$$

3. THE WEIGHTED MAXIMUM LIKELIHOOD METHOD

Yang[9]discussed this method on exponential regression model but this paper will discuss it on Weibull regression models. Suppose that associated with each individual is 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 lifetimeY has p.d.f. and survival functions

$$f(y/x,t) = \gamma exp\left[\gamma[y-(x\beta_0+g(t)] - e^{\gamma[y-(x\beta_0+g(t)]}\right] \quad (5)$$

and

$$S(y/x,t) = \exp\left(-e^{\gamma\left(y-(x\beta_0+g(t))\right)}\right), \quad (6)$$

respectively, where $\beta_0$is the true parameter value.

The likelihood function for a censored sample based on nindividuals 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[2], let the parameter$\lambda_i = x_i\beta_0 + g(t_i)$, then $x\beta_0$is the parametric portion,with $\beta_0$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$\upsilon \geq 2$ continuous derivatives. Several assumptions are made that allow an association between *x* and *t* (Rice[5] andSpeckman[7]).

Assume the regression model $x_i = r(t_i) + \eta_i$where $r(t)$ is a smooth function with υ continuous derivatives and $\eta_i$are independent random error terms with $E(\eta_i) = 0$and $E({\eta_i}^2) = \sigma^2$. Now $\lambda_i$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 $\beta_0$and $h_i = h(t_i), (i = 1,\ldots, n)$ in the Weibullsemiparametric regression model by maximizing the weighted likelihood function

$$WL(\beta,\theta) = \frac{1}{n}\sum_{i=1}\sum_{j=1}\left\{w\left(\frac{t_i-t_j}{b}\right)/\sum_{j=1}^{n}w\left(\frac{t_i-t_j}{b}\right)\right\}\left\{\log[f(y_i/x_i,t_i)]^{\delta_i}[1-F(y_i/x_i,t_i)]^{1-\delta_i}\right\}$$

$$= \frac{1}{n}\sum_{i} WL(\beta,\theta_i) \quad (7)$$

with respect$\beta and\theta$, where $\theta = [\theta_1,\ldots, \theta_n]'$. Here $\theta_i$is used to indicate thefunction of parameters of h_i. Expressions are given in terms of log lifetime $y$and its p.d.f. (5). that independent observations $(y_i, x_i, t_i), i = 1,\ldots, n$ areavailable, 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 logcensoring time, respectively.Throughout this article the sum is assumed to be from 1 to n. $WL(\beta,\theta)$ depends on the unobserved_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$\hat{h_i}$tosimultaneously maximize $WL(\beta,\theta)$with respect to β andθ.To understand the motivation for the weighted likelihood function, one can refer to the approaches of Staniswalis [3] and Hunsberger[5]. 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 estimateβ_0, because all of the observations contain information aboutβ_0. The weighted likelihood function can be writtenas

$$WL(\beta,\theta)=\frac{1}{n}\sum_i\sum_j w_{i,j}\{\delta_i\log f(y_i/\eta_i,t_i)+(1-\delta_i)\log S(y_i/\eta_i,t_i)\}$$

$$=\frac{1}{n}\sum_i\sum_j w_{i,j}\{\delta_i[\log\gamma+(y_i-(\eta_i\beta+h(t_i))-exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}]$$

$$-(1-\delta_i)exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}\} \quad (8)$$

$$=\frac{1}{n}\sum_i\sum_j w_{i,j}\{\delta_i[\log\gamma+(y_i-(\eta_i\beta+h(t_i))]-exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}\}$$

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 $\eta$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

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

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

$$\frac{\partial WL(\beta,\theta)}{\partial\beta}=\frac{1}{n}\sum_i\sum_j w_{i,j}\left\{-\delta_i\eta_i+\gamma\eta_i exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}\right\} \quad (9)$$

$$\frac{\partial WL(\beta,\theta)}{\partial\theta_i}=\frac{1}{n}\sum_i\sum_j w_{i,j}\left\{-\delta_i+\gamma exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}\right\} \quad (10)$$

$$\frac{\partial^2 WL(\beta,\theta)}{\partial\beta^2}=\frac{1}{n}\sum_i\sum_j w_{i,j}\left\{-\gamma^2\eta_i{}^2 exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}\right\} \quad (11)$$

$$\frac{\partial^2 WL(\beta,\theta)}{\partial\theta_i{}^2}=\frac{1}{n}\sum_i\sum_j w_{i,j}\left\{-\gamma^2 exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}\right\} \quad (12)$$

$$\frac{\partial^2 WL(\beta,\theta)}{\partial\beta\theta_i}=\frac{1}{n}\sum_i\sum_j w_{i,j}\left\{-\gamma^2\eta_i exp\{\gamma(y_i-(\eta_i\beta+h(t_i))\}\right\} \quad (13)$$

The maximum likelihood equations

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

and

$$\frac{\partial WL(\beta,\theta)}{\partial\theta_i}=0 \quad (15)$$

For each fixed $t\ and\ \beta,\hat{h}(t_i)$, the estimator of $\hat{h}(t_i)$, is obtained by solving (10).

$$\hat{h}(t_i)=-\frac{1}{\gamma}\log\left[\frac{\sum_j w_{i,j}\delta_i}{\gamma\sum_j w_{i,j}\{exp[\gamma(y_i-\beta\eta_i)]\}}\right] \quad (16)$$

ISSN 2347-1921

Hence

$$\frac{\partial \hat{h}(t_i)}{\partial \beta} = -\frac{\sum_j w_{i,j}\eta_i\{exp[\gamma(y_i - \beta\eta_i)]\}}{\sum_j w_{i,j}\{exp[\gamma(y_i - \beta\eta_i)]\}} \quad (17)$$

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

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

Since

$$\frac{\partial}{\partial \beta}\left(\frac{\partial WL(\beta,\hat{\theta}(\beta))}{\partial \beta}\right) = \frac{1}{n}\sum_i \sum_j w_{i,j}\left\{\gamma\eta_i exp\{\gamma(y_i - (\eta_i\beta + h(t_i))\}\left(-\gamma\eta_i - \gamma\frac{\partial \hat{h}(t_i)}{\partial \beta}\right)\right\} \quad (19)$$

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

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

4. SIMULATION

Asmallsimulationstudywasconductedtostudythefinite sampleproperties of the $\hat{\beta}$ and $\hat{h}$ in the semiparametric model defined in Section 2. The standard extreme random number is generated using the following transformation:

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

According to the model (4): $y = x\beta + g(t) + \varepsilon$. If $u \sim U[0,1]$, then ε follows the standard extreme distribution with p.d.f equal to

$$\gamma \exp(\gamma s - \exp(\gamma s)), \qquad -\infty < s < \infty$$

In this simulation, the t_is are equally spaced as $t_i = i/n$, for $i = 1, \ldots, 100$. A single Monte-Carlo realization consists of $n = 100$ observations. For the generated dataset, $x = r(t) + \eta$, $r(t) = 1$, $\eta \sim N(0, 0.1^2)$, $\beta = 15$ and $g(t) = 6(1-3t)^2$, hence, $h(t) = r(t)\beta + g(t)$, the model (4) becomes

$$y = x\beta + g(t) + \varepsilon$$

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 Müller quadratic kernel with $\nu = 2$ was used through:

and $b = 0.05$ for estimating $h(t)$ were used, respectively. The simulation shows that the WMLE method estimating $\beta_0 = 15$ and $h(t)$ well. We obtain $\hat{\beta} = 15.034$. Fig. 1 plots the estimates of the function $h(t)$ versus t, it indicates the estimated curve captures the true curve closely.

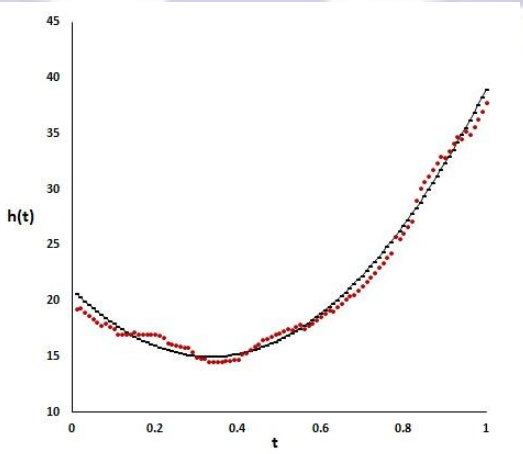

Fig. 1. Plot for the estimated nonparametric function $h(t)$ from the simulation study. The dotted line is the estimated function; the solid line is the true function.

ISSN 2347-1921

REFERENCES

[1] Hastie, T. and Tibshirani, R. (1986).“Generalized additive models”. Statistical science, 1, 297–318.

[2] Hunsberger, S.(1984). “Semiparametricregression”. Journal of the American statistical association, 1354–1365.

[3] Khokan, M., Bari, W. and Khan, J. (2013). “Weighted maximum likelihood approach for robust estimation: Weibull model”. Dhaka university journal of science, 61(2): 153-156.

[4] Lawless, J., (1982).“Statistical models and methods for lifetime data”. John Wiley& Sons. New York.

[5] Rice, J. (1986). “Convergence rates for partially splined models”. Statistics and probability letters, 4 (4): 203–209.

[6] Severini, T. andStaniswalis, J. (1994). “Quasi–likelihood estimation in semiparametric models”. Journal of the American statistical association, 89: 501–511.

[7] Speckman, P. (1988). “Kernel smoothing in partial linear models”. Journal of the royal statistical society, 3 (50):413–436.

[8] Staniswalis, J. (1989). “The kernel estimate of a regression function in likelihood–based models”. Journal of the American Statistical Association, 84(405): 276–283.

[9] Yang, J. (2009). “Exponential semiparametric regression models under random censorship”. World journal of modelling and simulation, 1(5):57-62.