Scientific Research and Essays Vol. 5(13), pp. 1589-1594, 4 July, 2010
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 ©2010 Academic Journals

Full Length Research Paper

Confidence intervals estimation for survival function in Weibull proportional hazards regression based on censored survival time data

Kamil Alakuş

Department of Statistics, Faculty of science and Arts, Ondokuz Mayıs University, 55139 Kurupelit- Samsun, Turkey. E-mail: kamilalakus@gmail.com. Tel: +90 362 312 19 19/ 5231.

Accepted 4 June, 2010

Weibull distribution plays a central role in the analysis of survival or life time data. Link (1984, 1986) presented a confidence interval estimate of survival function using Cox's proportional hazard model with covariates. Her idea more recently extended by Alakuş et al. (2007) to the exponential distribution and Alakuş et al. (2007) to exponential proportional hazard model, respectively. Alakuş (2010) studied confidence intervals for survival function from Weibull distribution. The same idea may be extended to the Weibull proportional hazard model provided that the survival times have a Weibull distributed random variable. In this study, we formed confidence interval for Weibull survival function for any values of the time and the covariates. Real data examples are also considered for illustrating the discussed confidence interval.

Key words: Confidence interval, hazard function, point estimation, survival analysis, survival function, Weibull distribution, Weibull proportional hazard model.

INTRODUCTION

There are two types of estimation for any identity. One is point and the other confidence interval. Survival analysis literature confidence interval estimate for the survival function is not new. Especially confidence interval estimate for the baseline survival function is extensively studied by many authors. For example, for Kaplan-Meier survival function confidence interval estimate is studied using Greenwood formulae by Kaplan and Meier (1958), Thomas and Grunkemeier (1975) and many others. In Cox's proportional hazard model, Link (1984, 1986) formed log transformed confidence interval for survival function with covariates. Her idea is more recently extended by Alakuş et al. (2007) to the exponential distribution and Alakuş et al (2007) exponential proportional hazard model, respectively.

For Weibull distribution baseline survival function, the confidence interval estimate is studied by many authors (For example; Alakuş, 2010). For the Weibull proportional hazard model, the confidence interval estimate for survival function is a new idea. Interval estimate for survival function is often useful in the analysis of survival or life time data. In this study, symmetric type confidence interval method is developed for Weibull survival function with covariates. The plan of this study is given as follows: Weibull proportional hazard regression model and its important functions are presented; next is the confidence interval estimate for the survival function from the Weibull proportional hazard model; and lastly, a real data example for illustrating the proposed method in this study is give. We completed the study with discussion.

WEIBULL PROPORTIONAL HAZARD MODEL

Suppose that the values $x_1,\cdots,x_k$ of k covariates $X_1,\cdots,X_k$, are recorded for each of n observations. Under the proportional hazard model, the hazard of death at time t for the i*th* observation is:

$$\lambda(t_i;\mathbf{x}_i)=\lambda_0(t_i)\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i), \quad (1)$$

for $i=1,\cdots,n$. Although this model has a similar appearance to Cox hazard model (Cox, 1984; Collet, 1994), there is one fundamental difference which

concerns the signification of the baseline hazard function $\lambda_0(t)$. In the Cox regression model, the form of $\lambda_0(t)$ is unspecified, and the share of the function is essentially determined by the actual data. In the model being considered in this study, the survival times are assumed to have a Weibull distribution, and this imposes a constant parametric form on $\lambda_0(t)$ when $\alpha = 1$, degreasing parametric form when $\alpha < 1$ and increasing parametric form when $\alpha > 1$.

Consider an observation for which the values of the k covariates in the model of equation (1) are all equal to zero. The hazard function for such an individual is $\lambda_0(t)$. If the survival time of observation has a Weibull distribution with scale parameter θ and shape parameter α, then its hazard function is such that $\lambda_0(t) = \alpha\theta^{\alpha} t^{(\alpha-1)}$. Using equation (1), the hazard function for i*th* observation in the study is then given by:

$$\lambda(t_i;\mathbf{x}_i) = \alpha\theta^{\alpha} t_i^{(\alpha-1)} \exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i). \quad (2)$$

From the form of this hazard function we can see that the survival time of the i*th* observation in the study has a Weibull distribution with scaled parameter $\theta^{\alpha}\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i)$. This again is a manifestation of the proportional hazards property of the Weibull distribution. The survival function corresponding to the hazard function in equation (2) is found to be:

$$S(t_i;\mathbf{x}_i) = \exp\left\{-\theta^{\alpha}\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i)t_i^{\alpha}\right\}. \quad (3)$$

The probability density function can be found by differentiating the survival function with respect to *t* and multiplying by (-1), or from the result that $f(t;\mathbf{x}) = \lambda(t;\mathbf{x})S(t;\mathbf{x})$, hence;

$$f(t_i;\mathbf{x}_i) = \alpha\theta^{\alpha} t^{(\alpha-1)}\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i)\times \exp\left\{-\theta^{\alpha}\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i)t_i^{\alpha}\right\}. \quad (4)$$

The Weibull proportional hazards model is fitted by constructing the likelihood function of the total n observations using the following equation:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left\{f(t_i;\mathbf{x}_i)\right\}^{w_i}\left\{S(t_i;\mathbf{x}_i)\right\}^{1-w_i}, \quad (5)$$

and maximizing this function with respect to the unknown parameters, $\theta, \theta_1, \cdots, \theta_k$ and α. Where w_i denotes an indicator variable which is zero if the survival time of the i*th* observation is censored and unit otherwise. In practice, this done using computer software for survival analysis.

Specifically, suppose that the estimates of the parameters in the model of equation (5) are $\hat{\theta}, \hat{\theta}_1, \cdots, \hat{\theta}_k$ and $\hat{\alpha}$. The estimated survival function for the i*th* observation in the study, for which the values of the covariates in the model are $x_{1i}, x_{2i}, \cdots, x_{ki}$, is then:

$$S(t_i;\mathbf{x}_i) = \exp\left\{-\hat{\theta}^{\hat{\alpha}}\exp(\hat{\alpha}\hat{\boldsymbol{\theta}}^T\mathbf{x}_i)t_i^{\hat{\alpha}}\right\}. \quad (6)$$

CONFIDENCE INTERVALS FOR SURVİVAL FUNCTION IN WEIBULL PROPORTIONAL HAZARD MODEL

Since, for the Weibull proportional hazard model is given by $\lambda(t_i;\mathbf{x}_i) = \lambda_0(t_i)\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i)$ where $\lambda_0(t_i)$ represent baseline hazard function from the Weibull distribution which is given by $\lambda_0(t_i) = \alpha\theta^{\alpha}t_i^{(\alpha-1)}$. Thus the Weibull proportional hazard model can be written as $\lambda(t_i;\mathbf{x}_i) = \alpha\theta^{\alpha}t_i^{(\alpha-1)}\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i)$ and accordingly survival function is $S(t_i;\mathbf{x}_i) = \exp\left\{-e^{\alpha\boldsymbol{\theta}^T\mathbf{x}_i}(\theta t_i)^{\alpha}\right\}$. Cumulative hazard function is given by $H(t_i;\mathbf{x}_i) = (\theta t_i)^{\alpha}\exp(\alpha\boldsymbol{\theta}^T\mathbf{x}_i)$. Relationship between the cumulative hazard function and survival function is that the survival function can be written as $S(t_i;\mathbf{x}_i) = \exp\{-H(t_i;\mathbf{x}_i)\}$. Taking logarithm of the hazard function we get $\log\{\lambda(t_i;\mathbf{x}_i)\} = \log\alpha + \alpha\log\theta + \alpha\boldsymbol{\theta}^T\mathbf{x}_i + (\alpha-1)\log t_i$. Defining $\alpha\log\theta$ with β_0, $\alpha\theta_1$ with $\beta_1, \cdots, \alpha\theta_k$ with β_k and $\log\alpha$ with β_{k+1}, then it might be rewritten as $\log\{\lambda(t_i;\mathbf{x}_i)\} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \beta_{k+1} + (\alpha-1)\log t_i$. Let $R_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \beta_{k+1} = \boldsymbol{\beta}^T\dot{\mathbf{x}}_i$ be score function of i*th* observation. Thus the survival function is written as $S_0(t_i;\mathbf{x}_i) = \exp\left(-e^{R_i}t_i^{\alpha}/\alpha\right) = \exp\left(-e^{\boldsymbol{\beta}^T\mathbf{x}_i}t_i^{\alpha}/\alpha\right)$. We can form confidence intervals using the relationship between the score function and survival function. That is, $100(1-\alpha)\%$ confidence intervals for R_i is given by:

$$\Pr\left\{\hat{R}_i - z_{\alpha/2}se(\hat{R}_i) \le R_i \le \hat{R}_i + z_{\alpha/2}se(\hat{R}_i)\right\} = 1-\alpha \quad (7)$$

or

$$\Pr\left(\hat{R}_{\text{low}} \le R_i \le \hat{R}_{\text{upp}}\right) = 1-\alpha. \quad (8)$$

Here $z_{\alpha/2}$ denotes coordinate value of standard normal

distribution at the significance level of $\alpha/2$ and $se(\hat{R}_i)$ also denotes as standard error of the score function and calculated using $se(\hat{R}_i) = \left\{\dot{\mathbf{x}}_i^T Var(\hat{\boldsymbol{\beta}})\dot{\mathbf{x}}_i\right\}^{1/2}$. In the last equation, $\dot{\mathbf{x}}_i^T = [1 \quad x_{1i} \quad \cdots \quad x_{ki} \quad 1]$ is an information column vector of i*th* observation and $Var(\hat{\boldsymbol{\beta}})$ is also estimated variance-covariance matrix of estimated model parameters. So the estimated variance-covariance matrix might be given by:

$$V(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_0,\hat{\beta}_k) & Cov(\hat{\beta}_0,\hat{\beta}_{k+1}) \\ Cov(\hat{\beta}_0,\hat{\beta}_1) & Var(\hat{\beta}_1) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_k) & Cov(\hat{\beta}_1,\hat{\beta}_{k+1}) \\ \vdots & \vdots & & \vdots & \vdots \\ Cov(\hat{\beta}_0,\hat{\beta}_k) & Cov(\hat{\beta}_1,\hat{\beta}_k) & \cdots & Var(\hat{\beta}_k) & Cov(\hat{\beta}_k,\hat{\beta}_{k+1}) \\ Cov(\hat{\beta}_0,\hat{\beta}_{k+1}) & Cov(\hat{\beta}_1,\hat{\beta}_{k+1}) & \cdots & Cov(\hat{\beta}_k,\hat{\beta}_{k+1}) & Var(\hat{\beta}_{k+1}) \end{bmatrix} \quad (9)$$

Therefore estimated survival function is given by $\hat{S}(t_i;\mathbf{x}_i) = \exp\left\{-e^{\hat{R}_i} t_i^{\hat{\alpha}}/\hat{\alpha}\right\} = \exp\left\{-e^{\boldsymbol{\beta}^T\mathbf{x}_i} t_i^{\hat{\alpha}}/\hat{\alpha}\right\}$. Now, we can conducted easily a $100(1-\alpha)\%$ confidence intervals of survival function using score function confidence intervals. Namely, the confidence intervals for survival function of Weibull proportional hazard model are given by $\hat{S}_{\text{low}}(t_i;\mathbf{x}_i) = \exp\left(-e^{\hat{R}_{\text{upp}}} t_i^{\hat{\alpha}}/\hat{\alpha}\right)$ for lower limit and $\hat{S}_{\text{upp}}(t_i;\mathbf{x}_i) = \exp\left(-e^{\hat{R}_{\text{low}}} t_i^{\hat{\alpha}}/\hat{\alpha}\right)$ for upper limit respectively. Shortly, $100(1-\alpha)\%$ confidence intervals for survival function of the Weibull proportional hazard model is as following:

$$\Pr\left(\hat{S}_{low}(t_i;\mathbf{x}_i) \le S(t_i;\mathbf{x}_i) \le \hat{S}_{\text{upp}}(t_i;\mathbf{x}_i)\right) = 1-\alpha \quad (10)$$

ILLUSTRATIVE EXAMPLES

Here, we consider real data illustration to confidence intervals for survival function as mentioned above. For this reason, First some information about the data is given. Second, we use the data for illustrating confidence intervals estimation for the survival function Weibull proportional hazard regression model.

Data: Lung cancer study

The data are taken from Statistical Sciences (1995). The lung cancer data were conducted by the North Central Cancer treatment Group. The lung cancer data frame includes the survival times (in days) and indicator variable (status) of death or censoring plus the following 8 additional variables on each patient. These are institution, patient's age, sex, physician's estimate of the ECOG performance score, physician's estimate of the Karnofsky score, patient's assessment of his/her Karnofsky score, calories consumed at meals including beverages and snacks and weight loss in the last 6 months. A total of 228 patients was used in the study. 138 patients were male and 90 patients were female. Total censoring ratio is 27.63%. Male patient's censoring ratio is 18.84% and female patient's censoring ratio is 41.11%.

Confidence intervals for survival function in Weibull proportional hazard model

Firstly, the survival times come from a Weibull distribution. This analysis has been taken by Alakuş (2009). Second, using a Weibull proportional hazard regression model with a single covariate x_i equal to 1 if the patient is female and 0 if the patient is a male, we have the results given in Table 1. Here, it was observed that there is evidence of difference in survival between the sex groups. Cox's regression model results were also given in the Table 2. As seen in Table 1 $\hat{\beta}_1$ value is very close to Cox's estimate in Table 2. Similar results are also said to standard error of $\hat{\beta}_1$. Likelihood ratio test results are very close in both models, too.

For calculating the confidence intervals, the variance-covariance matrix of the Weibull proportional hazard model is needed. Therefore, estimated variance-covariance matrix for Weibull proportional hazards regression model is given by:

$$V(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.0009083458 & -0.0092766110 & -0.0007705399 \\ -0.0092766110 & 0.0285786250 & 0.0017315867 \\ -0.0007705399 & 0.0017315867 & 0.0038338805 \end{bmatrix}$$

Once an estimate of the survival curve and the corresponding confidence interval for a given covariate pattern has been obtained. For instance, from Figure 1, when survival time 306 days with a male patient's point estimate of survival probability is 0.4362 for separated model and for Weibull proportional hazard model is 0.4456. The same patient's approximate 95% confidence interval of survival probability rang from 0.3692 - 0.5190 in Weibull proportional model. Therefore, the same patient's confidence interval range changes from 0.3546 -0.5148 in the separated model. For a second example, when survival

Table 1. Results of Weibull proportional hazard model to the lung cancer data.

Parameter	Value	Std. Err.	z-Test	p-value
$\hat{\beta}_0 = \hat{\alpha}\log\hat{\theta}$	-7.793	0.095	-81.764	0.000
$\hat{\beta}_1 = \hat{\alpha}\hat{\beta}$	-0.524	0.169	-3.099	0.002
$\hat{\beta}_2 = \log\hat{\alpha}$	0.281	0.062	4.537	0.000
Likelihood ratio test	10.4	1 degrees of freedom		0.001

Table 2. Results of Cox Proportional hazard model to the lung cancer data.

Parameter	Value	Std. Err.	z-Test	p-value
$\hat{\beta}_1$	-0.531	0.167	-3.176	0.002
Likelihood Ratio Test	10.6	1 degrees of freedom		0.001
Wald Test	10.1	1 degrees of freedom		0.002
Score Test	10.3	1 degrees of freedom		0.001

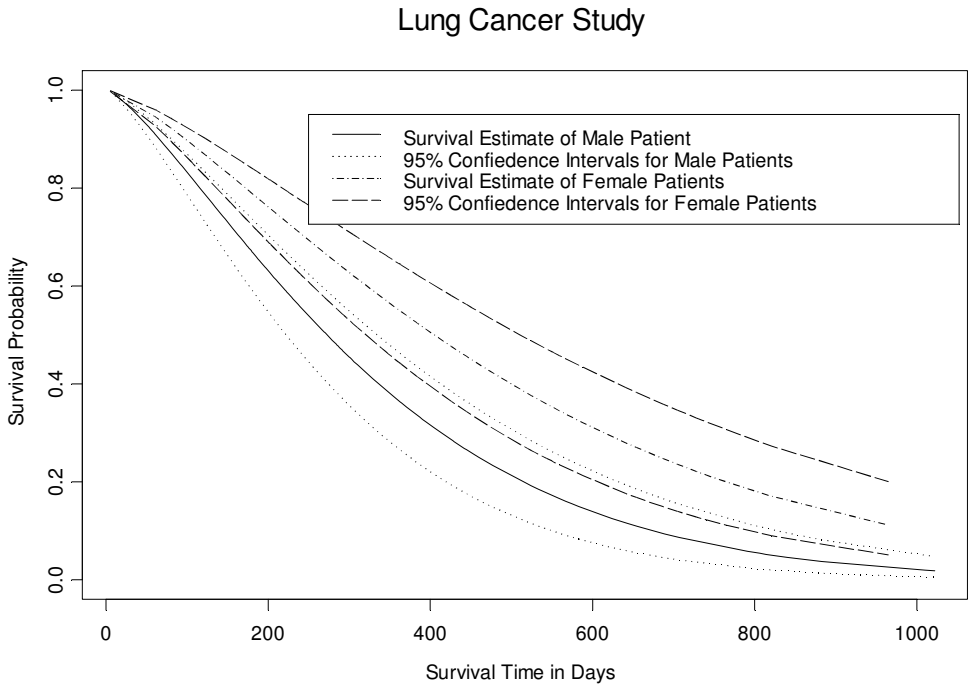

Figure 1. Estimated survival curves for male and female patients with lung cancer based on a Weibull distribution. Approximate 95% confidence limits are obtained using the score function approach.

time 310 days with a female patient's point estimate of survival probability is 0.6424 for separated model and for Weibull proportional hazard model is 0.6144. The same female patient's approximate 95% confidence interval ranges from 0.5228 - 0.6936 for Weibull proportional model. Consequently, same patient's confidence interval range changes from 0.5282 - 0.7358 in the separated model. From these results, the Weibull proportional hazard model confidence intervals are narrower than the separated one.

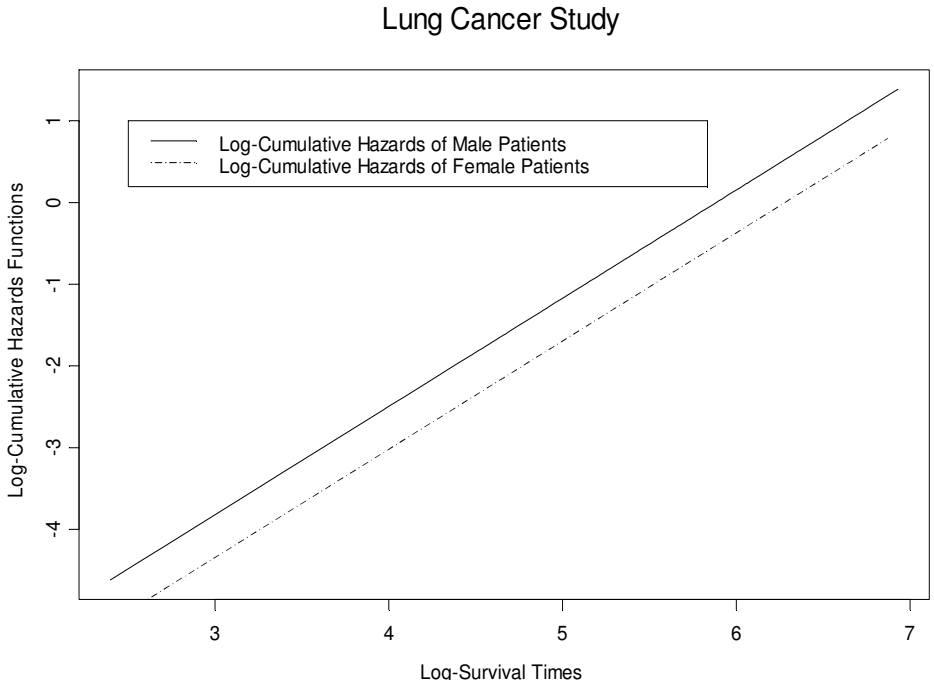

Figure 2. Testing of proportionality for a Weibull Proportional hazards regression model with lung cancer study.

For the lung cancer data, the best parametric model is Weibull proportional hazard model. For this model, check the validity of the assumption of proportionality may be needed. For done this, simple graphical checks of Weibull proportional hazards models are:

One may plot the $\log\hat{H}(t_i;\mathbf{x}_i)$ against $\log t_i$. Under the proportional hazards model, the curves should be approximately parallel.

Alternatively, one may directly plot difference $\log\hat{H}_h(t_i;\mathbf{x}_i)-\log\hat{H}_k(t_i;\mathbf{x}_i)$ against $\log t_i$ and should be approximately linear.

For example, Andersen et al. (1993) emphasized on a simple graphical method for comparing the two models. Hazard plotting was first investigated by Nelson (1969). Testing of proportionality from the Weibull proportional hazard function is given in Figure 2. Figure 2 shows that Weibull proportional hazard model is valid.

DISCUSSION

Many statistical investigations have been made in both estimation and hypothesis testing. There are two type of estimations. One type is point estimation and the other is interval estimation. Both point and intervals estimations can be achieve with an estimator. Interval estimation is generally called confidence interval estimation and naturally the estimators are also called confidence intervals estimators.

In the survival analysis, one of the important functions is the survival function. For this reason, estimation of survival function including with covariates is also very important. Both point estimation and confidence interval estimation of the survival function may be achieved fitting by semi-parametric or parametric proportional hazard models. Semi-parametric proportional hazard model is known as Cox regression model. In Cox regression model confidence interval estimation of survival function was studied by Link (1984; 1986). Her idea was recently extended by Alakuş et al. (2007) to the exponential distribution and Alakuş et al (2007) to exponential proportional hazard model, respectively. Alakuş (2010) studied confidence intervals for survival function from Weibull distribution. However, the problem in Weibull proportional hazard models has not been investigated so far. Therefore, in this study, for analysing of survival time data in which there are many applications found in Weibull proportional hazard model of survival function, point and confidence interval estimation are studied.

Here, results are illustrated using a real data example. The results obtained from the application show that the implication is very satisfactory. However, investigated confidence interval may extend to other parametric proportional hazard models (for example; log-normal, log-logistic etc.). These problems will be investigated in the forthcoming studies.

ACKNOWLEDGEMENTS

The author is highly thankful to the referees and the

editor for their valuable and useful suggestions.

REFERENCES

Alakuş K, Öner Y, Tunç T (2007). Sansürlü ve tamamlanmış örneklemlerde üstel dağılımın sağ kalım fonksiyonu için güven aralığı metotlarının karşılaştırılması. (in Turkish)Antalya: Beşinci İst. Kongresi, pp. 449-59.

Alakuş K, Tunç T, Öner Y (2007). Üstel orantılı hazard regresyon modelinde sağ kalım fonksiyonu için güven aralığı tahmini. (in Turkish) Ankara: İstatistik Araştırma Sempozyumu (İAS'07), pp. 258-65.

Alakuş K (2010). Confidence Intervals Estimation for Survival Function in Weibull Distribution Based on Censored Survival TimData. In appear to G.U.J. Sci.

Andersen PK, Borgan ø, Gill RD, Keiding N (1993). Statistical Models Based on Counting processes. Springer-Verlag: New York.

Collet D (1994). Modelling Survival Data in Medical Research. Chapman and Hall: London.

Cox DR, Oakes D (1984). Analysis of Survival Time Data. Chapman and Hall: London.

Kaplan EL, Meier P (1958). Nonparametric estimation from incomplete observations, J. Am. Statist. Ass., 53: 457-81.

Link CL (1984). Confidence intervals for the survival function using Cox's proportional hazard model with covariates, Biomet., 40: 601-610.

Link CL (1986). Confidence intervals for the survival function in the presence of covariates, *Biomet.,* 42: 219-220.

Nelson W (1969). Hazard plotting for incomplete failure data, J. Qual. Technol., 1: 27-52.

Statistical Sciences (1995). S-PLUS Version 3.3 Supplement. Stat. Sci., Seattle.

Thomas DR, Grunkemeier GL (1975). Confidence interval estimation of survival probabilities for censored data, J. Am. Statist. Ass., 70: 865-871.