Erciyes University Journal of the Institute of Science and Technology
Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi
ISSN 1012-2354
Cilt (Volume): 31 , Sayı (Issue): 2, Mayıs/May -2015
http://fbe.erciyes.edu.tr/

Weibull dağılımında sansürlü sağ kalım verileri için güven aralığı tahmini

***Kamil ALAKUŞ[1], Necati Alp ERİLLİ[2], Yüksel ÖNER[1]**
[1]*Department of Statistics, University of Ondokuz Mayis, Samsun, Turkey*
[2]*Department of Econometrics, University of Cumhuriyet, Sivas, Turkey*

Anahtar Kelimeler:
Güven Aralığı, Hazard fonksiyonu, Sağ kalım analizi, Sağ kalım fonksiyonu, Weibull dağılımı.

ÖZET

Weibull Dağılımı, sağ kalım analizlerinde veya gerçek hayat verilerinde önemli bir rol oynamaktadır. Daha önce yapılan çalışmalarda, ortak değişkenli orantılı Cox hazard modeli kullanarak güven aralığı tanıtılmıştır. Bu fikir daha çok üstel orantılı hazard fonksiyonu ile üstel dağılımın birlikte kullanımlarının genişletilmiş hali olmaktadır. Verilerin sağ kalım zamanlarının Weibull dağılımı göstermesi sonucu ile yukarıdaki fikre benzer şekilde verilere Weibull sağ kalım fonksiyonu uygulanabilmektedir. Bu çalışmada, verideki sağ kalım zamanlarındaki herhangi bir değere ait Weibull sağ kalım fonksiyonu için güven aralığı oluşturulmuştur. Önerilen güven aralığı için gerçek zaman verisi üzerinde uygulama yapılarak tartışılmıştır.

Confidence intervals estimation for survival function in weibull distribution based on censored survival time data

Key Words:
Confidence interval, Hazard function, Survival analysis, Survival function, Weibull distribution.

ABSTRACT

Weibull distribution plays an important role in the analysis of survival or life time data. Previous articles presented a confidence interval estimate of survival function using Cox's proportional hazard model with covariates. The idea is more recently extended to the exponential distribution and exponential proportional hazard model, respectively. The same idea may be extended to the Weibull distribution which provides that the survival times have a Weibull distributed with random variable. In this study, we formed confidence interval for Weibull survival function in any values of the survival time in the data. Real time data examples are also considered for the discussed confidence intervals.

*Sorumlu Yazar (Corresponding author) e-posta: kamilal@omu.edu.tr

1. Introduction

Survival analysis is a discipline of statistics which deals with death in biological organisms and failure in mechanical systems. This topic is called reliability theory or reliability analysis in engineering and duration analysis or duration modeling in economics or event history analysis in sociology. Survival analysis attempts to answer questions such as; fraction of a population which will survive in a certain time, or what rate they will die or fail, etc.

To answer such as questions, it is necessary to define "lifetime". In the case of biological survival, death is unambiguous but for mechanical reliability, failure may not be well-defined. There may be mechanical systems in which failure is partial, a matter of degree, or not localized in time. Even in biological problems, some events (like, heart attacks other organ failure) may have the same ambiguity. The theory outlined below assumes well-defined events at specific times; other cases may be better treated by models which explicitly account for ambiguous events.

More generally, survival analysis involves the modeling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature – traditionally only a single event occurs for each subject, after which the organism or mechanism is dead or broken. Recurring event or repeated event models relax that assumption. The study of recurring events is relevant in systems reliability and in many areas of social sciences and medical research. A general question can be asked at this position: “Why don’t we compare mean time-to-event between our groups using a t-test or linear regression”. Because they ignore censoring. Subjects are said to be censored if they are lost to follow up or drop out of the study or if the study ends before they die or have an outcome of interest. They are counted as alive or disease-free for the time they were enrolled in the study. Second question can be asked, too: “Why don’t we compare proportion of events in our groups using risk/odds ratios or logistic regression”. Because it ignores time. For these cases, survival analysis is mandatory for studies which includes time-to-time or censored situations.

In estimation theory, there are two types of estimation for any identity. One is point estimation and the other is confidence interval estimation. Survival analysis literature on the confidence interval estimate for the survival function is studying for a while. 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 by using Greenwood formulae by (Kaplan and Meier, 1958), and many others. In Cox’s proportional hazard model formed log transformed confidence interval for survival function with covariates studied by Link (1984). Her idea is more recently extended to the exponential distribution and exponential proportional hazard model (Alakus et all., 2007). For Weibull distribution baseline, survival function for the confidence interval estimate is a new idea. Interval estimate for survival function is often useful in the analysis of survival or lifetime data. In this study, a symmetric type of confidence interval method is developed for Weibull survival function without covariates.

2. Weibull distribution

In practice, the assumption of a constant hazard function or equivalently exponentially distributed survival times, is rarely tenable. A more general form of hazard function is given in equation (1):

$$\lambda(t) = \alpha\theta^{\alpha} t^{(\alpha-1)} \qquad t > 0 \tag{1}$$

It is a function which depends on two parameters θ and α, which are both greater than zero. In particular case where $\alpha = 1$ the hazard function takes a constant value θ and the survival times have an exponential distribution. For other values of α, the hazard function increases or decreases monotonically. But that does not change its direction. The shape of hazard function depends on critically for the value of α, while the parameter θ is scale parameter so α is known as the shape parameter. For this particular choice of hazard function, the survival function is given in equation (2).

$$S(t) = \exp\left(-\int_0^t \alpha\theta^{\alpha} s^{\alpha-1} ds\right) = \exp\left\{-(\theta t)^{\alpha}\right\} \tag{2}$$

The corresponding probability density function is then given by equation (3).

$$f(t) = \lambda(t)S(t) = \alpha\theta^{\alpha} t^{(\alpha-1)} \exp\left\{-(\theta t)^{\alpha}\right\}, \quad t > 0 \tag{3}$$

Equation (3) has the density of a random variable that has a Weibull distribution with scale parameter θ and shape parameter α.

2.1. Confidence intervals for survival function

For a Weibull distribution, hazard function is given by $\lambda_0(t_i) = \alpha\theta^{\alpha} t_i^{(\alpha-1)}$ and accordingly survival function is $S_0(t_i) = \exp\left\{-(\theta t_i)^{\alpha}\right\}$. Cumulative hazard function is also given by $H_0(t_i) = (\theta t_i)^{\alpha}$. The relationship between the cumulative hazard function and survival function can be written as $S_0(t_i) = \exp\{-H_0(t_i)\}$. Taking logarithm of the hazard function we can get;

$\log\{\lambda_0(t_i)\} = \log\alpha + \alpha\log\theta + (\alpha-1)\log t_i$ .Here defining $\alpha\log\theta$ with β_0 and $\log\alpha$ with β_1 then it might be rewritten as $\log\{\lambda_0(t_i)\} = \beta_0 + \beta_1 + (\alpha-1)\log t_i$.

Let $R_i = \beta_0 + \beta_1 = \beta^T I$ be the score value of i.*th* observation. Thus the survival function is written as $S_0(t_i) = \exp\left\{-e^{R_i} t_i^{\alpha}/\alpha\right\} = \exp\left\{-e^{\beta^T \mathbf{1}} t_i^{\alpha}/\alpha\right\}$. We can form confidence intervals by using the relationship between the score function and the survival function. So, $100(1-\alpha)\%$ confidence intervals for R_i is given by equation (4) or equation (5):

$$\Pr\left\{\hat{R}_i - z_{\alpha/2} se(\hat{R}_i) \le R_i \le \hat{R}_i + z_{\alpha/2} se(\hat{R}_i)\right\} = 1-\alpha \tag{4}$$

or

$$\Pr\left(\hat{R}_{\text{low}} \le R_i \le \hat{R}_{\text{upp}}\right) = 1-\alpha \tag{5}$$

Here $z_{\alpha/2}$ denotes the coordinate value of standard normal distribution at the significance level of $\alpha/2$ and $se(\hat{R}_i)$ also denotes the standard error of the score function. It can be calculate using by $se(\hat{R}_i)=\{I^T Var(\hat{\beta}) I\}^{1/2}$. In the last equation, $I^T=[1 \quad 1]$ is an unit column vector and $Var(\hat{\beta})$ is also variance-covariance matrix of the estimated parameters. In this simple form, the estimated variance-covariance matrix might be given by equation (6):

$$Var(\hat{\beta})=\begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0,\hat{\beta}_1) \\ Cov(\hat{\beta}_0,\hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix} \quad (6)$$

Therefore estimated survival function is given in equation.7:

$$\hat{S}_0(t_i)=\exp\left(-e^{\hat{R}_i} t_i^{\hat{\alpha}}/\hat{\alpha}\right)=\exp\left(-e^{\boldsymbol{\beta}^T \mathbf{1}} t_i^{\hat{\alpha}}/\hat{\alpha}\right) \quad (7)$$

Now, we can form easily a $100(1-\alpha)\%$ confidence interval of survival function using the score function of confidence intervals. Namely, the confidence intervals for survival function of a Weibull distribution are given by $\hat{S}_{\text{low}}(t_i)=\exp\left(-e^{\hat{R}_{\text{upp}}} t_i^{\hat{\alpha}}/\hat{\alpha}\right)$ for lower limit and $\hat{S}_{\text{upp}}(t_i)=\exp\left(-e^{\hat{R}_{\text{low}}} t_i^{\hat{\alpha}}/\hat{\alpha}\right)$ for upper limit, respectively. Shortly, $100(1-\alpha)\%$ confidence interval for survival function of the Weibull distribution is given in equation (8):

$$\Pr\left\{\hat{S}_{\text{low}}(t_i)\le S_0(t_i)\le \hat{S}_{\text{upp}}(t_i)\right\}=1-\alpha \quad (8)$$

3. Application

In this subsection, we consider real data illustration to confide intervals for survival function which we discussed in the earlier sections of the study. For this reason, firstly we will give some information about the data in the next subsection. Secondly, we use the data for illustrating the confidence intervals estimation for the survival function of the Weibull distribution.

3.1. Data: lung cancer study

Data are taken from *Statistical Sciences (1995).* The lung cancer data was conducted by the North Central Cancer Treatment Group. The lung cancer data frame includes the survival times (in days) and the indicator variable (status) of death or censoring plus the following 8 additional variables on each patient. These are institutions, 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. There were total 228 patients in the study. 138 patients are male and 90 patients are female. Total censoring ratio is 27.63%. Male patient's censoring ratio is 18.84% and female patient's censoring ratio is 41.11%.

3.2. Confidence interval for survival function in weibull distribution

We firstly have taken the goodness of fit test for the survival times coming from a Weibull distribution. To do this, a common and useful technique for checking the validity of a parametric model is to be embedded in the larger parametric model and use, e.g., the likelihood ratio test to check whether the reduction of the actual model is valid; for applications in the survival analysis used in Cox and Oakes (1984).

There are numbers of special techniques which are valuable for model checking. One of the known methods for checking Nelson-Aalen estimator is looking graphic of scatter. Thus we can have an idea whether the survival time appear to follow a certain parametric distribution or not. Actually this rationale statement can be find in Nelson (1969) original paper. For example, consider the Weibull distribution with hazard rate function $\lambda_0(t)=\alpha\theta^\alpha t^{(\alpha-1)}$ and cumulative hazard function $H_0(t)=(\theta t)^\alpha$. Here $\log H_0(t)=\alpha\log\theta+\alpha\log t$, so that $\log\hat{H}_0(t)$ plotted against $\log t$ should yield an approximately straight line for the Weibull distribution.

Figure 1 shows the log Nelson-Aalen estimates for the integrated hazard for male and female lung cancer patients. Both curves roughly linear, suggesting that a model with increasing hazards may be appropriate. In the same Figure 1, the corresponding log integrated hazard estimates (straight lines) based on Weibull distribution are added and they are seen to approximate the Nelson-Aalen estimates quite well.

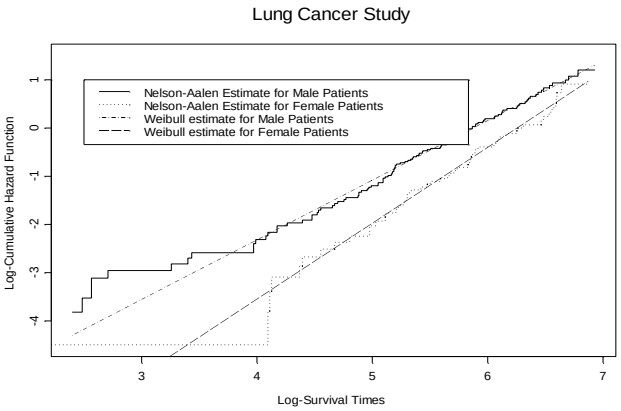

Figure.1. Graphical Test for a Weibull distribution for male and female patients with lung cancer study

The maximum likelihood estimates for α become 1.236967 for male and 1.573363 for female patients. Thus, for both sexes the likelihood ratio tests for the hypothesis $\alpha=1$ give very significant results. The Wald tests also give similar results. Moreover for the test whether the survival times come from a Weibull distribution, we use the Kolmogorov-Smirnov type test. The test statistic results are $D_{138}=0.0347$ for male patients and $D_{90}=0.0767$ for female patients respectively. This indicates that a Weibull distribution for both sexes is a reasonable one.

We have fit the exponential, log logistic, log normal and Weibull models separately to the data on male and female patients from lung cancer study. The log likelihood values for each model are reported in Table 1. From this table, we see that the Weibull model provides the best fit to this data.

Table.1. Results the log likelihood of fitting parametric models to the lung cancer data.

Distribution	Sex Group		General
	Male	Female	Male+Female
	Log Likelihood Value	Log Likelihood Value	Log Likelihood Value
Exponential	-767.7623	-389.8372	-1157.600
Log logistic	-768.8506	-385.1176	-1154.586
Log normal	-772.8860	-389.3269	-1162.616
Weibull	-764.1697	-382.9108	-1148.652

Therefore we fit the Weibull distribution to separated sex groups and the results are given in Table 2. From the Table 2, we see that both scale and shape parameters are very significant for two sexes.

Table.2. Results of separated Weibull models to the lung Cancer data.

Group	Parameter	Value	Std. Err.	z Test	p value
Male	$\hat{\beta}_{M0}$	-7.267	0.095	-76.291	0.000
	$\hat{\beta}_{M1}$	0.213	0.075	2.828	0.005
Female	$\hat{\beta}_{F0F}$	-9.841	0.138	-71.063	0.000
	$\hat{\beta}_{F1}$	0.453	0.109	4.153	0.000

For calculating the confidence intervals of survival functions, we must give the separated estimated variance-covariance matrices. Estimated variance-covariance matrices for male and female patients are, respectively;

$$V(\hat{\beta}) = \begin{bmatrix} 0.0090723708 & -0.0009019137 \\ -0.0009019137 & 0.0056565087 \end{bmatrix} \text{ and}$$

$$V(\hat{\beta}) = \begin{bmatrix} 0.019177654 & 0.001920457 \\ 0.001920457 & 0.011906970 \end{bmatrix}$$

A Weibull survival curve with approximate 95% confidence interval plot for male patients is given in Figure 2. We shall here illustrate the calculation of confidence interval using the data from the male patients. Figure 2 shows the Weibull survival function estimate for males with approximate 95% confidence limits using (8). The limits can be seen well fitted.

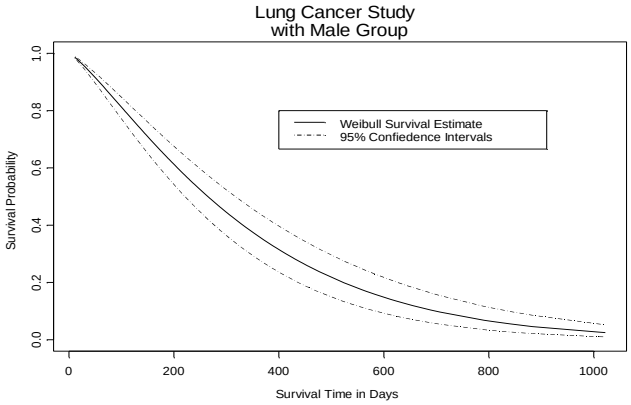

Figure.2. Estimated survival curve for male patients with lung cancer based on a Weibull distribution. Approximate 95% confidence limits are obtained using the score function approach

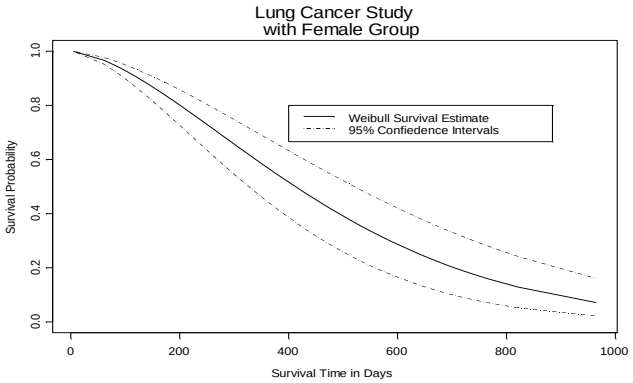

Figure.3. Estimated survival curve for female patients with lung cancer based on a Weibull distribution. Approximate 95% confidence limits are obtained using the score function approach.

Similarly a Weibull survival curve plot with approximate 95% confidence limits for female patients is given in Figure 3. We shall here illustrate the calculation of confidence interval using the data from the female patients. Figure 3 shows the Weibull survival function estimate for females with approximate 95% confidence limits using (8). The intervals are seen satisfactorily.

4. Result and discussion

In the survival analysis, one of the important functions is the survival function. For this reason, the estimation of a survival function is also very important case for survival literature. Both point estimation and confidence interval estimation of the survival function may be achieved by fitting parametric distributions. Semi-parametric proportional hazard model is known as Cox regression model. In the Cox regression model, confidence interval estimation of survival function has been studied by Link (1984) and Link (1986). Her idea was more recently extended to the exponential distribution and exponential proportional hazard model, respectively.

However the problem in Weibull distribution has not been investigated so far. In this study, for the analysis of survival time data, we studied some applications for Weibull distribution of survival function point and confidence interval estimations. The results are illustrated with a real data application. The application results are quite satisfactory for survival studies. The investigated confidence interval may be extended Weibull and other proportional hazard models. These problems will be investigated in forthcoming studies.

References

1. Alakuş, K., Öner, Y. and Tunç, T. Tamamlanmış ve Sansürlü Örneklemlerde Üstel Dağılımın Sağ Kalım Fonksiyonu İçin Güven Aralığı Metotlarının Karşılaştırılması. 5. İstatistik Kongresi Bildiriler Kitabı, Antalya, s.449-459, 2007.
2. Alakuş, K., Tunç, T. and Öner, Y. Üstel Orantılı Hazard Regresyon Modelinde Sağ Kalım Fonksiyonu İçin Güven Aralığı Tahmini. TÜİK, 16. İstatistik Araştırma Sempozyumu Bildiriler Kitabı, s. 258–264., 2007.
3. Cox, D.R. and Oakes, D. Analysis of Survival Time Data. Chapman and Hall, London, 1984.
4. Kalbfleisch, J.D. and Prentice, R.L. The Statistical Analysis of Failure Time Data. Wiley, New York, 1980.
5. Kaplan, E.L. and Meier, P. Nonparametric estimation from incomplete observations. J. Am. Statist. Ass., 53: 457-81., 1958.
6. Link, C.L. Confidence intervals for the survival function using Cox's proportional hazard model with covariates. Biometrics, 40: 601-610., 1984.
7. Link, CL. Confidence intervals for the survival function in the presence of covariates. Biometrics, 42: 219-220., 1986.
8. Nelson, W. Hazard plotting for incomplete failure data. J. Qual. Technol., 1: 27-52., 1969.
9. Statistical Sciences, S-PLUS Version 3.3 Supplement. StatSci: Seattle, USA, 1995.
10. Thomas, D.R. and Grunkemeier, G.L. Confidence interval estimation of survival probabilities for censored data". J. Am. Statist. Ass., 70: 865-871., 1975.