Confidence Intervals Estimation for Survival Function in Log-Logistic Distribution and Proportional Odds Regression Based on Censored Survival Time Data

Kamil ALAKUŞ, Necati Alp ERILLİ

Ondokuz Mayıs University, Faculty of science and Arts, Department of Statistics, 55139 Turkey

Abstract

Log-logistic and Weibull distributions have both accelerated survival time property. The log-logistic distribution has also proportional odds property. Log-logistic distribution has unimodal hazard curve which changes direction. Link [6,7] presented a confidence interval estimate of survival function using Cox’s proportional hazard model with covariates. Her idea more recently extended by [1] to the exponential distribution and [2] to exponential proportional hazard model, respectively. The same idea has been extended to the Weibull proportional hazard regression model by [3]. In this study, it is formed on confidence interval for log-logistic distribution survival function for any values of the time provided that the survival times have a log-logistic distributed random variable. It is also extended the same results to the proportional odds regression. A real time data and a simulation data examples are also considered in the study for illustration the discussed confidence interval.

Keywords: Confidence interval; Hazard function; Point estimation; Survival analysis; Survival function; Log-logistic distribution; Proportional odds regression

Introduction

There are two types of estimation for any identity. One is point estimation and the other is confidence interval estimation. In 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 with many authors. For example, for Kaplan-Meier survival function confidence interval estimate is studied using Greenwood formulae by [4,5] and many others. In Cox’s proportional hazard model, [6,7] formed log transformed confidence interval for survival function with covariates. Her idea more recently extended by [1] to the exponential distribution and [2] to exponential proportional hazard model, respectively. Weibull proportional hazard function is also investigated by [3]. Interval estimate for survival function is generally useful in the analysis of survival or life time data. In this study, symmetric type and proportional odds transformed confidence interval approach is developed for log-logistic survival function without covariates and proportional odds regression survival function with covariates.

The plan of this study is given as follows. In the next section, the log-logistic distribution and proportional odds regression and their important functions are presented. In section 3, it is formed confidence interval estimate for the survival function of log-logistic distribution and proportional odds regression model too. In section 4; it has given a real data example and as an extension of the real data, simulation study for illustrating the proposed method in this study. The study completed with a discussion section.

Log-logistic distribution and proportional odds regression model

Log logistic distribution: The log-logistic distribution has the proportional odds property and the distribution is the natural one to use in conjunction with the proportional odds model. Cox and Oakes [8] demonstrated that the log-logistic distribution is the only one that shares the accelerated survival time property and proportional odds property. Situations in which the hazard function changes, direction of hazard function can arise. For example, a patient faces an increasing hazard of death in heart transplantation over the first few days or after the transplant, while the body adapts to the new recovers. The hazard then decreases with time as the patient recovers. In situations such as this, a unimodal hazard function may be appropriate.

A particular form of unimodal hazard is the function

\[ \lambda(t) = \left( \alpha t^{-1} \right) \left( 1 + (\lambda t)^{\alpha} \right)^{-1} \]  

This hazard function decreases monotonically if \( \alpha \leq 1 \). If \( \alpha > 1 \), than the hazard has a stable mode. The survival function corresponding to the hazard function in equation (1) is given below

\[ S(t) = \exp \left( \int_0^t \left( \alpha s^{-1} \right) \left( 1 + (\lambda s)^{\alpha} \right)^{-1} ds \right) \left( 1 + (\lambda t)^{\alpha} \right)^{-1} \]  

Probability density function is also given below

\[ f(t) = \lambda(t) S(t) = \left( \alpha t^{-1} \right) \left( 1 + (\lambda t)^{\alpha} \right)^{-2} \]  

This is the density function of a random variable \( T \) which has a log-logistic distribution with parameters \( \alpha \) and \( \lambda \). The distribution is so called because the variable \( \log T \) has log-logistic distribution and a symmetric distribution whose probability density function is very similar to the normal distribution [9].

The suitability of the log-logistic distribution for the analysis of a data set can be empirically checked using a linear relationship derived

*Corresponding author: Kamil ALAKUŞ, Ondokuz Mayis University, Faculty of science and Arts, Department of Statistics, 55139 Turkey. E-mail: kamilal@omu.edu.tr

Received December 21, 2010; Accepted June 06, 2011; Published September 25, 2011


Copyright: © 2011 ALAKUŞ K, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
from expression of \( S(t) \) and \( F(t) = 1 - S(t) \). The odds of surviving beyond time \( t \) are \( S(t)/F(t) = (\lambda t)^{\alpha} \) and consequently log odds of survival beyond \( t \) can be expressed as \( \log \left( \frac{S(t)}{F(t)} \right) = -\alpha \log \lambda - \alpha \log t \) or equivalently
\[
\text{log} \left( \frac{F(t)}{S(t)} \right) = \alpha \log \lambda + \alpha \log t \quad \text{(4)}
\]

That is, the log-logistic distribution corresponds to a linear model for the log odds of failure over the logarithm of time, with slope \( \alpha \). The Kaplan-Meier sample estimate \( \hat{S}_m(t) \) can be used to calculate the log odds and a plot of them versus \( \log t \) should follow approximately a straight line for the log-logistic model in order to be suitable.

**Proportional odds regression model**

The application of accelerated survival time and proportional odds models to the analysis of reliability data has recently been described by [12]. The general proportional odds model for survival data was introduced by [13,14] describes the proportional odds model.

The log-logistic distribution is commonly extended to include a vector of covariates \( x \) by reformatting the survival function is given as follows
\[
S(t; x) = \left[ 1 + (\lambda t)^{\alpha} \exp(\alpha x^T \theta) \right]^{-1}
\quad \text{(5)}
\]

Under the accelerated failure time model, the hazard of death at the time of \( t \) is given below
\[
h(t; x) = \left( \alpha \exp(\alpha x^T \theta) \right) \left[ 1 + (\lambda t)^{\alpha} \exp(\alpha x^T \theta) \right]^{-1}
\quad \text{(6)}
\]

Thus the survival time for the \( i \)-th observation also has a log-logistic distribution and therefore it has both accelerated failure time and proportional odds property.

**Confidence intervals for survival function**

**Log-logistic distribution**: First multiply, both numerator and denominator, by \( \alpha \) with equation (2). The equation value will not change. Therefore we can re-write equation (2) as follows
\[
S(t; x) = \alpha \left[ 1 + (\lambda t)^{\alpha} \exp(\alpha x^T \theta) \right]^{-1}
\quad \text{(7)}
\]

If we take natural logarithm of \( (\lambda t)^{\alpha} \) identity, we can get the result as \( \log \alpha \) + \( \log \lambda \). Now let \( R = \log \alpha + \log \lambda \) be the score (risk score) function. Then we denote \( \log \lambda \) with \( \beta_0 \) and \( \log \alpha \) with \( \beta_1 \). Therefore the survival function can be written as \( R = \beta_0 + \beta_1 x \). Therefore the survival function calculated using by
\[
\hat{S}(t; \beta_0 + \beta_1 x) = \alpha \left[ 1 + e^{\beta_1 x_t} \right]^{-1}
\quad \text{(8)}
\]

Estimated survival function is also given by
\[
\hat{S}(t) = \left[ 1 + e^ {\beta_1 x_t} \right]^{-1}
\quad \text{(9)}
\]

As a result we can prepare for confidence intervals of survival function for the log-logistic distribution. To do this, we first can form confidence intervals for \( \hat{R} \), then extended it to the survival function. For the score function \( R \), 100(1-\( \alpha \))% confidence interval is given by
\[
\Pr \left( \hat{R}_l \leq R \leq \hat{R}_u \right) = 1 - \alpha
\quad \text{(10)}
\]

or
\[
\text{Pr} \left( \hat{R}_l \leq R_i \leq \hat{R}_u \right) = 1 - \alpha
\quad \text{(11)}
\]

Here \( \lambda \) denotes coordinate value of standard normal distribution at the significance level of \( \alpha/2 \) and \( \text{se}(\hat{R}) \) is also denotes standard error of estimated score function. The estimated standard error of the score function calculated using by
\[
\text{se}(\hat{R}) = \left[ x_T \text{Var}(\hat{\alpha}) x_i \right]^{1/2}
\quad \text{(12)}
\]

where, \( x_T = [1 \ 1] \) is a unit column vector in simple model and \( \text{Var}(\hat{\alpha}) \) is also variance-covariance matrix of estimated parameters.

We can easily form a 100(1-\( \alpha \))% confidence intervals of survival function using score function confidence intervals. So that, the confidence intervals for survival function of a log-logistic distribution are given by
\[
\hat{S}_R \left( t_i \right) = \left[ \hat{\alpha} + e^{\hat{\beta}_0 + \hat{\beta}_1 x_t} \hat{\theta} \right]^{-1}
\quad \text{for lower limit and}
\]
\[
\hat{S}_R \left( t_i \right) = \left[ \hat{\alpha} + e^{\hat{\beta}_0 + \hat{\beta}_1 x_t} \hat{\theta} \right]^{-1}
\quad \text{for upper limit, respectively. So,}
\]

100(1-\( \alpha \))% confidence intervals for survival function of the log-logistic distribution can be given as following
\[
\Pr \left( \hat{S}_R \left( t_i \right) \leq S(t_i) \leq \hat{S}_R \left( t_i \right) \right) = 1 - \alpha
\quad \text{(13)}
\]

**Proportional odds regression model**

Let \( R = \log \alpha + \log \lambda + \alpha x \), be score function for \( i \)-th observation. If we denote \( \log \alpha \) with \( \beta_0 \) and \( \log \lambda \) with \( \beta_1 \), then the score function for \( i \)-th observation can be written as \( R = \beta_0 + \beta_1 x_i \). Here \( \beta_0 \) is (p\times1) column vector and equals to \( \alpha \theta \).

Let \( \beta_0 \) + \( \beta_1 \times \beta_0 \) denote as \( \beta \) a \( (k \times 1) \) size column vector and \( \gamma_0 \) also a same size column vector too. Then the score function for \( i \)-th observation can be written as \( R = \beta x_i \gamma_0 \). So that, the survival function calculated using in (2) can be written as in equation (14)
\[
S(t; x) = \alpha \left[ 1 + e^{\beta x_i \gamma_0} \right]^{-1}
\quad \text{(14)}
\]

Estimated survival function is also given by
\[
\hat{S}(t; \beta_0 + \beta_1 x_i \gamma_0) = \alpha \left[ 1 + e^{\beta x_i \gamma_0} \right]^{-1}
\quad \text{(15)}
\]

Therefore we can prepare confidence intervals of survival function for the proportional odds regression. As we made for log-logistic distribution, firstly we can form confidence intervals for \( R \) then extended it to the survival function. For the score function \( R \), 100(1-\( \alpha \))% confidence interval is given by equation (10) or equation (11). The estimated standard error of the score function for \( i \)-th observation is calculated using equation (16)
\[
\text{se}(\hat{R}) = \left[ y_i^T \text{Var}(\hat{\alpha}) y_i \right]^{1/2}
\quad \text{(16)}
\]

where, \( y_i^T = [1 \ 1 \ x_{i1} \ \cdots \ x_{ik}] \) a column vector in this model and \( \text{Var}(\hat{\alpha}) \) is also variance-covariance matrix of estimated parameters. In this model, the estimated variance-covariance matrix might be given as follows
\[
\text{Var}(\hat{\alpha}) = \begin{pmatrix}
\text{Var}(\hat{\beta}_0) & \cdots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_k) \\
\vdots & \ddots & \vdots \\
\text{Cov}(\hat{\beta}_0, \hat{\beta}_k) & \cdots & \text{Var}(\hat{\beta}_k)
\end{pmatrix}
\quad \text{(17)}
\]

A 100(1-\( \alpha \))% confidence intervals of survival function uses
confidence intervals of score function. Namely, the confidence intervals for survival function of a proportional odds model are given by 

\[ \hat{S}_{\text{low}}(t; y) = e^{\hat{a} - 2 \hat{\beta}_{\text{POR}} f(t)^2} \] for lower limit and 

\[ \hat{S}_{\text{upp}}(t; y) = e^{\hat{a} + 2 \hat{\beta}_{\text{POR}} f(t)^2} \] for upper limit, respectively. So, 

\[ 100(1-\alpha)\% \text{ confidence intervals for survival function of the proportional odds regression can be shown as follows} \]

\[ \Pr\left( \hat{S}_{\text{low}}(t; y) \leq S(t; y) \leq \hat{S}_{\text{upp}}(t; y) \right) = 1 - \alpha \] (18)

Model selection criterion

In this study we tested performance of proposed model with log-likelihood value and besides with AIC and BIC model selection criterions. The Akaike information criterion (AIC) is a measure of the relative goodness of fit of a statistical model. It was developed by [15]. Bayesian Information Criterion (BIC) is a criterion for model selection among a class of parametric models with different numbers of parameters. It has been introduced by [16]. The AIC and BIC is given as follows, respectively

\[ \text{AIC} = -2 \times \log \text{lik} + 2 \times p \] (19)

\[ \text{BIC} = -2 \times \log \text{lik} + p \ln (n) \] (20)

When data has small samples size, corrected AIC can be given

\[ \text{AICc} = -2 \times \log \text{lik} + \frac{2 \times (p + 1)}{n - p - 1} \tag{21} \]

where \( p \) is defined as number of free parameters.

Given a set of candidate models for the data, the preferred model is the one with the smallest value of Log Likelihood, AIC and BIC. 

Illustrative example

In this section a real data illustration considered to confidence intervals for survival function which we discussed it earlier sections in this study. For this reason, we first give some information about the data just in the next subsection. Second, we use the data for illustrating confidence intervals estimation for the survival function of the proportional odds regression model.

Real time data: Ovarian cancer study

Data are from [17] on ovarian cancer. The data are taken from [18]. The ovarian cancer frame includes the survival times (in days) and indicator variable (status) of death or censoring plus the following 4 additional variables on each patient. These are patient’s age (age), an indicator of the extent of the residual disease (residual.dz), treatment given (rx) and measure of performance score or functional status using the Eastern Cooperative Oncology Group’s scale (ecog.ps). The survival analysis chapter in the S-Plus documentation describes these data sets further and illustrates survival analysis methods with them. There were 26 patients in the study. Total censoring ratio is 53.85%.

Confidence intervals for survival function in Proportional odds regression model

We first have taken the goodness of fit for the survival times which comes from a log-logistic distribution. To do this, a common and useful technique for checking the validity of a parametric model is to embed it in larger parametric model and use, e.g., the likelihood ratio test to check whether the reduction to the actual model is valid; for applications in survival analysis [8,19].

Secondly, the tests of the survival times come from a log-logistic distribution. We use the Kolmogorov-Smirnov type test. The test statistic result is \( D_n = 0.1251 \). This further indicates that a log-logistic distribution is a reasonable one.

Thirdly, one use of the Nelson-Aalen or Kaplan-Meier estimators for survival data is to check graphically whether the survival time will appear to follow a certain parametric distribution; in fact, this was the rationale behind the estimator in [20] original paper. For the log-logistic distribution with log odds function is

\[ \log \left( \frac{F(t)}{1-F(t)} \right) = \alpha \log \lambda + \alpha \log t \]

so that log odds plotted against log \( t \) should yield an approximately straight line for the log-logistic distribution. This result is given in (Figure 1a) for the ovarian cancer data. The curve is roughly linear, suggesting that a model may be appropriate. In the same Figure, the corresponding log odds estimates (straight lines) based on log-logistic distribution are added the graph and it can seen for approximate the Kaplan-Meier estimates quite good too. In (Figure 1b), QQ plot support same results in (Figure 1a). In (Figure 1c) Baseline Hazard function shows unimodal shape. Also in (Figure 1d), proportional odds regression hazard is a increasing function for risk score function of age.

As a result, (Figure 1d) summarizes that when ages and survival times increases hazard function increases in same direction.

We have also fit the Exponential Hazard Regression (HER), Proportional odds Hazard Regression (POR), Log-Normal Hazard Regression (LNR) and Weibull Hazard Regression (WHR) models to the data in ovarian cancer study. The log likelihood, AIC, AICc and BIC values for each model are given in (Table 1). From the table, we can see that the proportional odds regression model provides the best fits for the data.

As we can see from (Table 1), Proportional odds regression values are smallest for all criterions. Thus we can easily say, Proportional odds regression model is better than the other models.

Therefore we can fit the Proportional odds regression model and the results are given in (Table 2). From the (Table 2), we can see that all three parameters are very significant.

For calculating the confidence intervals of survival functions estimated variance-covariance matrix is given below

\[ V(\hat{\beta}) = \begin{bmatrix} 9.3920 & 0.2210 & -0.1519 \\ 0.2210 & 0.0589 & -0.0031 \\ -0.1519 & -0.0031 & 0.0025 \end{bmatrix} \]

Some calculations of confidence intervals for survival function in the log odds regression model is given in (Table 3). Survival probabilities estimates and 95% confidence intervals are also given (Figure 2).

Approximate 95% confidence limits are obtained using the risk score function and log odds transformation approach.

Figure 2 shows the log odds regression survival function estimate for ovarian cancer with approximate 95% confidence intervals using (18). Because of small sample data the confidence interval results are little bigger than expected. However we can easily say, the result is quite good.

Simulation study

In order to Ovarian Cancer Data has small sample; to see the
behavior of Proportional odds regression on big samples, a simulation study is studied in this sub-section. Simulation study data has 100 variable value and obtained as follows:

Step 1: Mean, variance and standard deviation for age variable in Ovarian Cancer Data is calculated. ($\sigma = 10.10036$, $\mu = 56.16544$)

Step 2: 100 age variable values has been simulated from normal distribution. Here $\text{Age} = z \sigma + \mu$ where $z \sim N(0,1)$.

Step 3: 100 random variable value has simulated with uniform distribution in $(0,1)$ interval.

Step 4: With the help of parameter estimation values of Ovarian Cancer Data, survival times is simulated. To do this: $R_i = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2^*$ 

Age is evaluated with risk function. Survival time is simulated with the formula $t_i = \left[\frac{\hat{\alpha}(1-u_i)u_i}{\hat{\alpha}}\right]^{1/\alpha}$. Where, $\hat{\alpha} = 2.22611$ and $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2^*$ are given as in (Table 2).

Step 5: 100 random variable value is simulated with Bernoulli distribution which has mean value (successful probability)
Figure 3: (a) Graphical Test For a Log-Logistic Distribution; (b) QQ Plot; (c) Baseline Hazard Function; (d) odds regression Hazard Function for Simulation Study.

Table 3: 95% Confidence Intervals for Survival Probabilities of Proportional odds regression Model to the Ovarian Cancer Data.
p=0.80. Therefore 20% censored survival times is identified using these variables.

The result for simulation data in graphs is given in (Figure 3). In (Figure 3a), the curve is quite linear. In the same Figure, the corresponding log odds estimates (straight lines) based on log-logistic distribution are added the graph and it can be seen for approximate the Kaplan-Meier estimates are very good. In (Figure 3b), QQ plot support same results in (Figure 3a). In (Figure 3) Baseline Hazard function shows unimodal shape. Also in (Figure 3d), proportional odds regression hazard is a increasing for risk score function of age.

Table 4 results shows that Proportional odds regression model has the smallest criterion values than the other models. We can easily say that it is the best model above others. In Table 5, estimation of parameters is given and we can say that all estimations are very significant.

For calculating the confidence intervals of survival functions estimated variance-covariance matrix is given below:

\[
V(\hat{\beta}) = \begin{bmatrix}
1.116442 & 0.010860 & -0.01848 \\
0.010860 & 0.00874 & -0.00016 \\
-0.01848 & -0.00016 & 0.00032
\end{bmatrix}
\]

<table>
<thead>
<tr>
<th>Model</th>
<th>Log Likelihood Value</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHR</td>
<td>-644.8942</td>
<td>1291.788</td>
<td>1294.3936</td>
</tr>
<tr>
<td>POR</td>
<td>-623.6054*</td>
<td>1251.211*</td>
<td>1256.4211*</td>
</tr>
<tr>
<td>LNR</td>
<td>-626.3671</td>
<td>1256.734</td>
<td>1261.9445</td>
</tr>
<tr>
<td>WHR</td>
<td>-643.6997</td>
<td>1291.399</td>
<td>1296.6097</td>
</tr>
</tbody>
</table>

Table 4: Results of Fitting parametric models to the simulated data.

\[
\hat{\beta}_0 = 28.9309, \hat{\beta}_1 = 0.8867, \hat{\beta}_2 = 0.2218
\]

For this reason, in this study we offer a new confidence interval with transformed log odds in symmetric type in proportional odds regression model. Proposed approach is studied with real time data and simulation data. Results were quite good enough. When sample size getting larger, confidence intervals were getting tighter. Model selection criterions Log-Likelihood, AIC and BIC gave the best results for proposed model. Confidence intervals are narrower than other studied models. At the thought real time data has small sample, a larger data set with simulation data also studied. Results are more robust than small sample. Based on the theorem of large numbers law, \(\hat{S}(t)\) will be converges in probability to \(S(t)\). This summarizes us when sample size is quite enough, distribution of risk function approximate to normal distribution and this becomes \(\hat{S}(t)\) closer to \(S(t)\). This is what we are investigating for.

The investigated confidence intervals may be extended to log-normal hazard model. This problem will be investigated by the forthcoming studies.

**Acknowledgements**

The authors are highly thankful to the referees and the editor for their valuable and useful suggestions.

**References**