Confidence Interval for Survival Model based on Partly Interval-Censored data

Norazelah Zainudin, F. A. M. Elfaki, and M. Yeakub Ali

Abstract—Cox Proportional Hazard model and its application on Empirical Likelihood (EL) based on Partly Interval-Censored data is proposed in this paper. Also, this paper will provide the information on the new version of Empirical Likelihood namely Adjusted Empirical Likelihood (AEL) that serve the purpose of improvising EL to solve the EL under coverage problem. A major advantage of the approach is its simplicity and it can be easily implemented by using R software. Simulation studies are conducted which indicate that the approach performs well comparable to the existing methods.

Keywords—Partly Interval-Censored data; Empirical likelihood; Adjusted empirical likelihood

I. INTRODUCTION

Cox Proportional Hazard Model has been widely used for over century for many applications in the well-known industries such as medical, economical and statistical analysis. The purpose of this model is to help ones to use any of methods such as Maximum Likelihood Estimator, Expectation Maximization and many more to estimate the baseline hazard function towards solving the problem that based on mathematical analysis study.

Numerous of study has been done regarding this model and some of the research involving engineering field such as the application of tool replacement using Weibull’s Proportional Hazard function and modelling in reliability. With the establishment of those studies, it proves that this model could be used and applied in engineering arena and of course, with some improvisation, this model could help to enhance the progressiveness and productivity of production area. This paper shall discuss the inference of these model using Empirical Likelihood and Adjusted Empirical Likelihood method based on partly interval-censored data by constructing the confidence region for EL and AEL. A simulation study is held and the result is compared between both methods to see the successfullness of the new method in overcoming the EL method problem.

As the research done within this study of Proportional Hazard Model, a similar pattern of study were detected which is vast information has been provided by previous researcher based on right, left and interval-censored mechanism. The most research done were right-censored analysis while only a few study made for left-censored. From time to time, the information on interval-censored research has been keep increasing and considered as on going. Same goes to partly interval-censored field which is also find to be ongoing study since there are lack of information could be provided about this type of censored. Partly interval-censored data consist of exact and interval-censored data. As mention by Elfaki et al. (2012), the final conclusion made in their paper is that the higher number of exact data exist in the simulation studies, the estimation yield would be much better compared to those with higher number of interval-censored data. Therefore, early theory could be made that instead of using only interval-censored data, adding some exact data in our study would actually help to improvise the yielding result, making partly interval-censored greater than interval-censored reaction within this model.

II. EMPIRICAL LIKELIHOOD AND ADJUSTED EMPIRICAL LIKELIHOOD METHOD

Empirical likelihood (EL) is the approach that being used to obtain an assumption using the random sample of Independent and Identically Distributed (IID) variables. The confidence interval used by EL could be determined by the data whereby the range of the intervals included is upheld. According to Qin and Zhang (2008), they have implemented EL with confidence intervals (CI) to interpret the uncertain data of structural differences within the populations which is specially made for mean and distribution function differences. Meanwhile, Wong et al. (2009) stated that diagnostic technique based on EL approach was being developed using partial linear model. EL method written by Zhang and Zhao (2013) is being used as a function to estimate mean functional missing response that were missing without notice (non-ignorable mechanism). As discussed by Jinnah (2007), EL has many qualities over Normal Approximation distribution (NA) such as the capability of providing better coverage probability for small sample sizes which is agreed by Zhang and Zhao (2013) as they were referring it as the under coverage problem. If the regression parameter in NA method should first be estimated to construct the confidence region, this step could be
eliminated in EL and able to yield the confidence region that is
dconformed to dataset without the need of being symmetric.

As suggested by Jinnah (2007), the score function of Cox’s
model is given by:

\[ U(\theta_0) = \sum_{i=1}^{n} \left( Z_i - \frac{\hat{a}_i(t, \theta_0)}{\hat{a}_0(t, \theta_0)} \right) dM_i(t) \]  

(1)

in such a way that

\[ M_i(t) = N_i(t) - \int_0^t \exp(\theta_0' Z_i) I(x_i \geq s) \lambda_0(s) ds \]

(i = 1, ..., n) are IID, martingales.

Therefore, the estimation of baseline hazard function \( \lambda_0(s) \) need to be done since it is not specified earlier. Hence, we will consider the \( \lambda_0(s) \) as \( h_0(s) \) to satisfy the partly interval-censored need and replace it with \( \hat{\lambda}_0(s) \).

Then equation (1) can be written as:

\[ U(\theta_0) = \sum_{i=1}^{n} \left( Z_i - \frac{\hat{a}_i(t, \theta_0)}{\hat{a}_0(t, \theta_0)} \right) d\hat{M}_i(t) \]

(2)

where

\[ \hat{M}_i(t) = N_i(t) - \int_0^t \exp(\theta_0' Z_i) I(x_i \geq s) \hat{\lambda}_0(s) ds \]

and

\[ \hat{\lambda}_0(s) = \sum_{i=1}^{n} \int_0^T \frac{dN_i(s)}{\sum_{j=1}^{n} I(x_j \geq s) e^{\theta_0' Z_j}} \]

(3)

The next step is to consider \( p = (p_1, ..., p_n) \) as the probability vector. With, \( \sum_{i=1}^{n} p_i = 1 \) and \( p_i \geq 0 \) for all \( i \).

Taking advantage of equation (3), for \( 1 \leq i \leq n \), we let:

\[ W_{n,i} = \int_0^T \left( Z_i - \frac{\hat{a}_i(t, \theta_0)}{\hat{a}_0(t, \theta_0)} \right) d\hat{M}_i(t) \]

(4)

Hence, the profile empirical likelihood ratio is defined as:

\[ R(\theta_0) = \sup \left\{ \prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i W_i(\theta), p_i \geq 0 \right\} \]

(5)

By using the Lagrangian multiplier approach, now we have:

\[ p_i = \frac{1}{n} \left( 1 + \lambda^T W_i(\theta) \right)^{-1}, \quad i = 1, ..., n \]

(6)

where \( \lambda = (\lambda_1, ..., \lambda_n)^T \) is the solution of

\[ \frac{1}{n} \sum_{i=1}^{n} W_i(\theta) = 0 \]

(7)

With regard to equation (7), thus these two properties of \( W_i \) is hold:

\[ n^{-1} \sum_{i=1}^{n} W_i \to N(0, I) \]

(8)

\[ \hat{I}_1(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} W_i W_i^T, \]

where, \( \hat{I}_1(\theta_0) \rightarrow I_1(\theta_0) = I(\theta_0) \) in probability

However, in general if \( I_1(\theta_0) \neq I(\theta_0) \), the assumption of
regularity conditions strongly holds that \(-2 \log R(\theta)\)
converges in the distribution of independent Chi-square
random variables with one degree of freedom as
aforementioned. Therefore, if we take \( \tilde{I}(\theta) = -2 \log R(\theta) \)
then we could get the confidence interval with asymptotic
100(1 - \alpha)% since we knew that it is converges in the
distribution to \( \chi^2_{\alpha} \). Thus, the confidence region constructs to be:

\[ R_1 = \{ \theta : \tilde{I}(\theta) \leq \chi^2_{\alpha}(\alpha) \} \]

(8)

Other than its advantages over NA, it also exhibit some
disadvantages when it comes to larger sample sizes, n in which
it will experiences a difficulty called under coverage problem.
It happened occasionally when the number of n keep
increasing and would cause the result provided distant from
our nominal level. Hence, it is important for us to develop new
type of Empirical Likelihood to overcome these particular
problem and it is called Adjusted Empirical Likelihood (AEL)
method. The purpose of developing AEL is to analyze whether
or not this new method of solving the partly interval-censored
data would yield better coverage probability as compared to
EL.

To take advantage of Jinnah (2007) proposed method of
AEL where the implementation of trace vector has taken place,
with \( tr(\cdot) \) denoted as trace vector, and we let the distribution
of \( \rho(\theta) \) be like \( \rho(\theta) = p + tr[I_1^{-1}(\theta) I(\theta)] \), therefore the
AEL method for this research could be written as:

\[ \hat{\rho}(\theta) = \frac{tr[I_1^{-1}(\theta) \tilde{I}(\theta)]}{tr[I_1^{-1}(\theta) I(\theta)]} \]

(9)

This is since the distribution of \( \hat{\rho}(\theta_0) \) could be
approximated by \( \chi^2_{\alpha} \) and as the asymptotic distribution of
AEL ratio could be written as \( \tilde{I}_{ad}(\theta) = \hat{\rho}(\theta_0) \tilde{I}(\theta_0) \),
therefore the adjustment factor for this equation is \( \hat{\rho}(\theta) \) with
\( I_1(\theta) \) and \( I(\theta) \) replaced by \( \tilde{I}(\theta) \) and \( \hat{I}(\theta) \).

Now, let us define \( \tilde{\rho}(\theta) \) to be \( \hat{\rho}(\theta) \) and \( \tilde{I}(\theta) \) be
replaced by \( \tilde{S}(\theta) \) in equation (9), then we obtain:

\[ \tilde{S}(\theta) = \left( \frac{\sum_{i=1}^{n} W_i(\theta)}{n} \right) \times \left( \frac{\sum_{i=1}^{n} W_i(\theta)}{n} \right)^T \]

(10)
Therefore, \( \hat{r}(\theta) \) can be written as:
\[
\hat{r}(\theta) = \frac{\text{tr}[\mathbf{I}^{-1}(\theta) \hat{\mathbf{S}}(\theta)]}{\text{tr}[\mathbf{I}^{-1}(\theta) \mathbf{S}(\theta)]}
\] (11)

Which later produce new AEL ratio of \( \hat{I}_{ad}(\theta) = \hat{r}(\theta) \hat{I}(\theta) \). Under the regularity condition, the EL statistic do converges in the distribution of \( \chi^2_p \). Therefore, the asymptotic 100(1 − \( \alpha \))% confidence interval for AEL should be similar as before that is:

\[
R_2 = \{ \theta : \hat{I}_{ad}(\theta) \leq \chi^2_p(\alpha) \}
\] (12)

### III. Simulation Data

A simple simulation studies was held in order to test the effectiveness between EL and AEL methods in which the set-up for the study is similar to the one held by Jinnah (2007).

#### TABLE I
RESULT OF SIMULATION STUDY ON 10% CENSORED RATE

<table>
<thead>
<tr>
<th>( n )</th>
<th>90</th>
<th>95</th>
<th>99</th>
<th>90</th>
<th>95</th>
<th>99</th>
<th>90</th>
<th>95</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
</tr>
<tr>
<td>30</td>
<td>89.13683</td>
<td>89.9997</td>
<td>89.32052</td>
<td>89.99930</td>
<td>89.30755</td>
<td>89.9996</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>89.94320</td>
<td>89.9999</td>
<td>89.93731</td>
<td>89.99999</td>
<td>94.65623</td>
<td>94.9999</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>89.98206</td>
<td>89.9999</td>
<td>89.97497</td>
<td>89.99999</td>
<td>89.93513</td>
<td>94.9999</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>89.98790</td>
<td>89.9999</td>
<td>89.98518</td>
<td>89.99999</td>
<td>89.93868</td>
<td>94.9999</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### TABLE II
RESULT OF SIMULATION STUDY ON 40% CENSORED RATE

<table>
<thead>
<tr>
<th>( n )</th>
<th>90</th>
<th>95</th>
<th>99</th>
<th>90</th>
<th>95</th>
<th>99</th>
<th>90</th>
<th>95</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
<td>AEL</td>
<td>EL</td>
</tr>
<tr>
<td>30</td>
<td>89.90272</td>
<td>89.99962</td>
<td>89.73874</td>
<td>89.99980</td>
<td>89.66232</td>
<td>89.99982</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>89.93523</td>
<td>89.99982</td>
<td>94.87290</td>
<td>94.9999</td>
<td>94.83649</td>
<td>89.9991</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>89.98206</td>
<td>89.9999</td>
<td>98.97497</td>
<td>98.99999</td>
<td>89.93513</td>
<td>94.9999</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>89.98790</td>
<td>89.9999</td>
<td>89.98518</td>
<td>89.99999</td>
<td>89.93868</td>
<td>94.9999</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://dx.doi.org/10.15242/IIE.E0315039
The data is generated on 30, 50, 100 and 150 sample sizes to different set of time failure that are T1, T3 and T5 with three different confidence interval level that is; 90%, 95% and 99%. Based on R Software, the simulation is run for 2000 to ensure the results provided parallel to our objectives.

In addition, our simulation also include censored rate percentage of 10%, 40% and 70% to see the effect of light, mild and heavy censored rate on our result. Be noted that this generated data later will be used to calculate the confidence region for both method that is EL and AEL in order to ensure that the calculated region are within the Chi-square distribution.

Table 1, Table 2 and Table 3 (which is not addressed here in this paper) show the results from the simulation studies. The simulation conducted based on the partly interval-censored data in which as discussed earlier, it contains exact and interval-censored information. According to Table 1, with different applications of failure time and from rough observations, it should be agreeable that the AEL has hit our objective of making adjustment of EL since it produced much better result as compared to EL (the value produced are closer to nominal value). Nominal value is the value of 90%, 95% and 99% that we have used during the simulation setup. If we dig more on the result provided in Table 1, we could see as the sample size, \( n \) are getting larger, EL exhibit evident under coverage problem. The percentage of error compared to smaller sample sizes could be assured.

Other than that, the importance of censored rate play a crucial role in enhancing our result. This could be seen clearly in Table 2 and Table 3. As stated in all of the tables above, with mild (40%) and heavy censored rate (70%), the outcome seems to be improve unlike with the application of light censored rate. In addition, the pattern of small to large sample sizes, \( n \) exhibit in Table 1 (light censored rate) seems to be continue for both result in Table 2 and Table 3. This can be assured by the resulting value that we obtain from simulation studies. However, dissimilar idea has been discovered as the failure time keep increasing. With small sample sizes and light censored rate (10%), as the failure time increased from T1 to T5, the result obtained is getting farther from the nominal value. On the other hand, with larger sample sizes (100 and 150) and light censored rate, this trend wear off (the figure gets closer to nominal value) and hold for the rest of simulations. This continuous pattern proves that with the enforcement of mild and heavy censored rate, the accuracy of the yielding outcome could be trusted.

IV. CONCLUSION

From the study that has been held and simulation that has been done, earlier objectives of improving EL by introducing AEL had been successfully met. The idea of improving one method by modifying it to be another method seems to be familiar to mathematicians in order to produce a more promising result that later could be used by others in their field. However, by developing new equation out of EL method would cause more confusion. It is much better to practically improve the available idea so that the effectiveness of those method is unquestionable. In conclusion, the comparison of EL and AEL are obviously discuss in this paper and the improvement of AEL had been achieve using partly interval-censored data.

REFERENCES


http://dx.doi.org/10.1111/1467-9868.00398


