Exponential Model for Survival Analysis

Faiz. A. M. Elfaki
Department of Mathematics, Statistics and Physics, College of Arts and Science, Qatar University, P.O. Box 2713, Doha, Qatar

Abstract:
This paper discusses the parametric model based on partly interval censored data, which is occur in many fields including engineering, medical, economic and other studies. By Partly Interval Censored (PIC) we mean that the exact failure time for some subject is observed but for the others are only known to fall within an interval [1]. In medical and reliability studies the most important function is the survival function. However, the survival function will be estimated using a parametric model based on imputation techniques in the present of PIC data and simulation data. Our proposal model is useful and easily implemented using R software.

Key words:
PIC, Exponential model

1. Introduction

The Cox's regression model is flexible model that can be used as a semi-parametric model or parametric methods and therefore it is the most practical and well-known statistical model to investigate the relationship between predictors and the time-to-event through the hazard function [2, 3]. In this model, there was no need for the researcher to assume a particular survival distribution for the data [4]. The only assumption made in the model is about the proportional hazards and this is why it is also called Cox proportional hazards regression [5].
Unlike the Cox's regression model that does not specify the distribution function of hazard function, there are several parametric models such as Weibull, exponential, log-normal, and log-logistic models where hazard function has to be specified [6]. Studies have indicated that under certain circumstances, such as strong effect or strong time trend in covariates or follow-up depending on covariates, parametric models are good alternatives to the Cox's regression model [4, 5, 6]. If the parametric models better fit the data, a more precise estimation of parameters would be achieved [7].
Parametric model based on Cox model have been proposed by many authors such as; [8]; [9]; [10]; [11]; [12]; ([13], 14]); [15]; etc.
Maximum likelihood (ML) is used for estimation of parameters in survival parametric models, while Cox's regression model is used for partial likelihood. However, the model that presented in this paper is the PHE based on the EM algorithm.

2. The Model

The Cox model is the most widely used model in survival analysis area such as filed of clinical trials, engineering, economic, etc. This model was introduced by Cox in 1972 for analysis of survival data with and without events.
Let T be continuous random variable, $\theta = (\theta_1,...,\theta_p)'$ be a vector regression parameters, $z = (z_1,...,z_p)$ be a exploratory variable associated with the individual or covariates and $h_0(t)$ is a baseline hazard, $h_0(t)$. Then the model can be written as;

$$h(t/z) = h_0(t)\exp(z\theta) \quad (1)$$

2.1The Proportional Hazards Exponential model (PHE)

Other than Cox model in survival analysis we can used model such as exponential and Weibull, both of which are parametric. In additional to that, the Cox PH model, the Weibull model allows more flexibility because the associated hazard rate is not constant with respect to time. In other words, if we replace the baseline hazard in equation (1) by given the exponential distribution. In this case, the *pdf* of the exponential incorporating the regressor variables (θ') is given as;

$$f(t \mid z) = e^{\theta' z}\exp(-te^{\theta' z}) \quad (2)$$

and reliability function as

$$R(t,z) = e^{-t.e^{\theta' z}}. \quad (3)$$

We can imply that the PH failure rate $h(t \mid z)$ of the exponential distribution as

$$h(t \mid z) = h_0(t)e^{\theta' z} \quad (4)$$

and thus the baseline hazard is given by

$$h_0(t) = 1. \quad (5)$$

[16] showed the log-likelihood function as follows:

$$l = \ln(L) = \sum_{i=1}^{F}\ln\left[f(T_{F,i})\right] + \sum_{j=1}^{S}\ln\left[R(T_{S,j})\right], \quad (6)$$

Manuscript received December 5, 2017
Manuscript revised December 20, 2017

If we replace equation (2) and (3) into equation (6). Equation (6) become;

$$l = \sum_{i=1}^{F} \ln\left(\exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right) \exp\left(-T_{F,i} e^{\sum_{j=0}^{m} \theta_j z_{ij}} \right) \right) - \sum_{i=1}^{S} T_{S,i} \exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right)$$

$$= \sum_{i=1}^{F} \left[\sum_{j=0}^{m} \theta_j z_{ij} - T_{F,i} \exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right) \right] - \sum_{i=1}^{S} T_{S,i} \exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right). \quad (7)$$

The first and second derivative of the equation (7) with respect to θ_i *(i= 1, 2, …, m)* are given by:

$$\frac{\partial l}{\partial \theta_i} = \sum_{i=1}^{F} \left[\sum_{j=0}^{m} z_{ij} - T_{F,i} z_{ij} \exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right) \right] - \sum_{i=1}^{S} T_{S,i} z_{ij} \exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right) \quad (8)$$

and

$$\frac{\partial^2 l}{\partial \theta_i^2} = \sum_{i=1}^{F} \left[\sum_{i=0}^{m} - T_{F,i} z_{ij}^2 \exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right) \right] - \sum_{i=1}^{S} T_{S,i} z_{ij}^2 \exp\left(\sum_{j=0}^{m} \theta_j z_{ij} \right). \quad (9)$$

The values of estimated parameters $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$ are obtained by letting equation (9) equal to zero $\theta_i (i = 1, ..., m)$. An iterative process is used to solve this system of equations for θ. However, with Newton-Raphson method we may encounter problems such as overflow or degeneration especially when the initial values chosen are not appropriately close to parameter estimates. Convergence may not be achieved. In this paper we shall adopt the EM algorithm method to our model since the EM algorithm uses the likelihood estimate based on the expected values given for initial condition, which then can be maximized in the standard way. It's also preferable to the Newton-Raphson method, because the likelihood for incomplete data is estimated through the expected values.

3. The EM algorithm

The Expectation-Maximization (EM) algorithm is used to find the model parameters when the data is incomplete. It is an iterative procedure to approximate the maximum likelihood function. The EM algorithm for parametric model is a useful tool in situations when the observed data likelihood function is complicated, while the complete data likelihood function is straightforward ([17]; [18]). The algorithm contains a sequence of iterations between the E-step and the M-step.

3.1. E-step

The conditional expectation of l based on failure time (T_F) given the observation (y, w, z), and the parameter estimates at the previous step $\theta^{(r)}$ (is the estimate of the parameter at the *r*-th iteration) is computed:

$$Q(\theta \mid \theta^{(r)}) = E[l(\theta; T_F) \mid y, w, z; \theta^{(r)}] \quad (10)$$

3.2. M-step

$Q(\theta \mid \theta^{(r)})$ obtained from the E-step is maximized using the complete data procedure. In this section, we review some basic results of the EM algorithm.

The function $Q(\theta \mid \theta')$ is finite dimensional and a differentiable function of θ and θ'. If θ is functional, then differentiability condition may not be satisfied. This is the reason why we include a parametric step in our proposed method.

THEOREM 1.3: [17] Suppose $\theta^{(r)}, r = 0,1,2,...,$ is an instant of a generalized EM algorithm (GEM) such that $\frac{\partial}{\partial \theta} Q(\theta \mid \theta^{(r)}) = 0$.

THEOREM 1.4: [17] Suppose $\theta^{(r)}, r = 0,1,2,...,$ is an instant of a generalized EM algorithm (GEM) such that

(1) $\frac{\partial}{\partial \theta} Q(\theta \mid \theta^{(r)}) = 0$

(2) $\theta^{(r)}$ converges to θ'

(3) $\frac{\partial^2}{\partial \theta \partial \theta'} Q(\theta \mid \theta^{(r)})$ is negative definite with eigenvalues bounded away from zero. Then

$\frac{\partial}{\partial \theta} l_W(\theta') = 0,$ and $\frac{\partial^2}{\partial \theta \partial \theta'} Q(\theta' \mid \theta')$ is negative definite.

These two theorems justify the EM for finite dimensional parameter space. Moreover, the log-likelihood for an uncensored exponential sample is given by:

$$l(\theta; T_F) = n \log \theta - \theta \sum_{i=1}^{n} T_{F,i}. \quad (11)$$

From equation (11) and equation (10). Then, we can write equation (10) as:

$$Q(\theta \mid \theta^{(r)}) = n\log\theta - \theta\sum_{i=1}^{n} E[l(\theta;T_{F,i}) \mid y_i, w_i;\theta] \quad .$$

(12)

Two steps was used to fit regression parametric models to the data (Clayton,1985). First is to calculate the E-step as follows:

$$E[l(\theta;T_{F,i}) \mid y_i, w_i;\theta)] = y_i + (1-w_i)/\theta \; . \quad (13)$$

From equation (13) and (12). Equation (12) can be written as:

$$Q(\theta \mid \theta^{(r)}) = n\log\theta - \theta\sum_{i=1}^{n} y_i + (n-d)/\theta \, , \quad (14)$$

where d is number of failures observed at time t_i .

Now we can updates the parameters by fitting the simple model to the transformed observations based on M-step. However, in the paper we update estimate for θ by compute the value of $\hat{\theta}_{i+1}$ which maximizes equation (14):

$$\hat{\theta}_{i+1} = \max^{-1} Q(\theta \mid \theta^{(r)}) . \quad (15)$$

The two steps may be combined to give the following formula ([12]):

$$\hat{\theta}_{i+1} = n\left(\sum_{i=1}^{n} y_i + \frac{n-d}{\hat{\theta}_i}\right)^{-1} . \quad (16)$$

However, M step is equivalent to finding the solution to

$$E[\frac{\partial}{\partial\theta}[l(\theta;T_{F,i}) \mid y_i, w_i;\theta]] = 0, \quad (17)$$

where the maximum likelihood estimate of θ for which solution is $\frac{\partial}{\partial\theta}[l(\theta;T_{F,i})] = 0,$ can be solved iteratively by solving equation (17).

4. Numerical Examples

We applied the PHE model mentioned early in this paper to the breast cancer data that have been modified by [19]. There was 46 patient treated by Radiation (R) and 48 patients treated by Radiation + adjuvant Chemotherapy (R+C). Reader refer to [19] for more detail to this data set. Figure 1 show the results obtained based on parametric exponential which is look almost similar to the one obtained by Turnbull. The hazard rate found to be 0.0292 and lower and upper bound are 0.0239 0.0356. Also, the likelihood ratio test is 14.2 with their P-value 0.000163. In additional to that, the result obtained by Turnbull from interval data set is found to be similar to the one obtained by our model from PIC data sets. However, based on PIC data, the midpoint show better results in term of smallest P-value compared to the one obtained by Turnbull.

5. Simulation Data

The simulated data were generated 1000 times (with 819 uncensored, 181 censored and the total time at risk is 50819.8) from breast cancer data with two failure times (R, R+C) that is mentioned early in last section. The generation of data was carried out by using R software. To generate the data we used the mean and standard deviation as 3.5711595, 0.6394705 for R, and 4.0205093, 0.3773525 for R+C, respectively.

Figure 2 showed the result of the estimation of survival function obtained by exact observation-Cox compared with parametric exponential model based on midpoint imputation technique. The estimated of the survival function from the two type of failures is very similar to the one obtained by our model compared with one obtained by Cox with exact data. The estimated hazard rate found to be 0.0161 with lower and upper bound are 0.0150 0.0173, respectively. However, this result indicated that our method is better in term of likelihood ratio test 42.54 and P-value 0.0024.

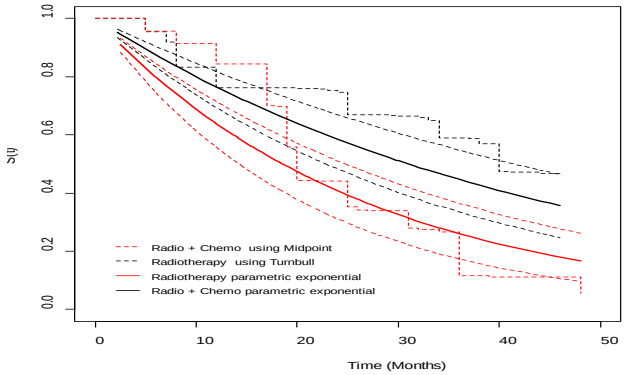

Figure 1: Survival function estimated by midpoint imputation compared with Turnbull based On exponential model from cancer PIC data

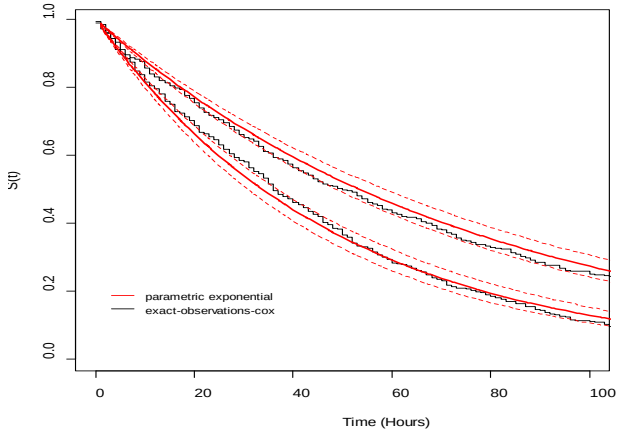

Figure 2: Survival function estimated by exact-observations Cox vs exponential model from simulation data

6. Conclusion

The parametric Cox's PHE has been used successfully to investigate the two causes of failures. The parameters in the model was estimated based EM algorithm. Our approach showed better result compared to the one obtained by Turnbull and exact observation Cox in term of P-value and likelihood ratio test via the simulation studies. Under exponential distribution, it was found that maximum likelihood estimate using the EM algorithm is preferable to others such as the method of Newton-Raphson, because the likelihood for complete data for Cox's model has a much simpler form than the likelihood corresponding to the Cox regression hazard model with censored data. Moreover, the EM algorithm does not always require the inversion of large matrices of large values.

Acknowledgements

The project with code number QUST-CAS-SPR-2017-26 was support this work.

References

[1] Kim. J. S. Maximim Likelihood Estimation for the Proportional Hazards Model with Party Interval-Censored Data. J R. Statist. Soc. 2003, Series B 65: 489-502.
[2] Efron B. The efficiency of Cox's likelihood function for censored data. J Am Stat Associ. 1977;Vol (72):557–65.
[3] Oakes D. The asymptotic information in censored survival data. Biometrika. 1977; 64:441–8.
[4] Paoletti X, Asselain B. Survival analysis in clinical trials: Old tools or new techniques. Surgical Oncology. 2010;(19):55–8.
[5] Vallinayagam V, Prathap S, Venkatesan P. Parametric Regression Models in the Analysis of Breast Cancer Survival Data. Int J Sci Tech. 2014; 3:163–7.
[6] Cox DR, Oakes D. Analysis of survival data. London:CRC Press. 1984.
[7] Klein J, Moeschberger M. Survival analysis: Statistical methods for censored and truncated data. 2nd ed. New York: Springer-Verlag. 2003.
[8] Elfaki, F.A.M., Azram, M. & Usman, M., Parametric Cox ' s Model for Partly Interval Censored Data with Application to AIDS Studies. 2012; 2(5), pp.352–354.
[9] Kundu, D. and Basu, S. Analysis of incomplete data in presence of competing risks. Journal of Statistical Planning and Inference. 2000; 87:221-239
[10] Jr. D. W and Lemesfow. S. Applied Survival Analysis: Regression Modeling of Time to Event Data. New York: John Wiley, 1999.
[11] Andersen, P. K., Borgan, Q., Gill, R. D. and Keiding, N. Statistical Models Based on Counting Processes. New York: Springer-Verlag, Inc. 1993.
[12] Cox, D. R. and Oakes, D. Analysis of Survival Data. London: Chapman and Hall. 1984.
[13] Lawless, J. F. Statistical Model and Methods for Lifetime Data. New York: Wiley. 1982.
[14] Lawless, J. F. Statistical Methods in Reliability. (with discussion). Technometrics, 25: 305-335. 1983.
[15] Basu, A. P. and J. K. Chosh. Asymptotic Properties of A solution to the Likelihood Equation with Life-Testing Application. J. Am. Stat. Assoc., 75: 410-414. 1980.
[16] Kalbfleisch, J. D and Lawless J. F. Estimation of Reliability in Field-Performance Studies. Technometrics. 1988, 30: 365-388.
[17] Dempster, A.P., Laird, N.M. and Rubin, D.B. Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, Vol. 39, 1-22. 1977
[18] Louis, T. A. "Finding the Observed Information Matrix when using the EM Algorithm. J. R. Statist. Soc., 44: 226-33. 1982.
[19] Abdallah Zyoud, F. A. M. Elfaki, and Meftah Hrairi. Nonparametric Estimate Based on Imputations Techniques for Interval and Partly Interval Censored Data. Science International (Lahore) Journal. 2016; Vol 28(2): 879-884.