MATEMATIKA, 2013, Volume 29, Number 1c, 161-167

A Comparative Study of the Estimation Methods on Frailty Mixture Survival Model

[1]Oh Yit Leng and [2]Zarina Mohd Khalid
[1]Faculty of Business and Law, Multimedia University, Jalan Ayer Keroh Lama
75450 Melaka, Malaysia.
[2]Department of Mathematical Sciences, Faculty of Sciences, Universiti Teknologi Malaysia
81300 UTM Skudai, Johor Malaysia.
e-mail: [1]yloh@mmu.edu.my

Abstract Frailty mixture survival models are statistical models which allow for a cured fraction and frailty. The cured fraction refers to a proportion of individuals who are expected not to experience the event of interest, while frailty refers to unobserved information amongst the individuals who experience the event of interest. In this study, we extend the frailty mixture survival model by including covariates into the frailty part of the model. We also employed both semiparametric and parametric methods in Gamma frailty mixture model. Using parametric method, the baseline survival function is assumed to follow Weibull distribution, while using semiparametric method, the cummulative baseline hazard function is assumed to be unknown since in some cases a parametric assumtionis diffuclt to justify. Estimation methods based on EM-algorithm and newton-raphson are utilized to obtain the maximum likelihood estimates of the unknown model parameters involved in both the semiparametric and parametric model. The study aims to compare the performance of estimators using these two methods in terms of their accuracy and efficiency measures.

Keywords Cured fraction; Frailty; EM-algorithm; Maximum Likelihood; Semiparametric, Para- metric.

2010 Mathematics Subject Classification 62N02

1 Introduction

In studies of survival analysis it is common to assume that all individual in the target population will experience the event of interest or failure if the research period is sufficiently long. However, in some cases a fraction of individuals called as long-term survivor, cured, or non-susceptible will not experience the event of interest. Survival model which had been used to analyse survival data with cured fraction are known as cured model. Cured model divides the target population into two subpopulations, which are long-term survivors and short-term survivors. The long-term survivors refer to individuals who are not expected to failure, while short-term survivors refer to individuals who are expected to failure with a proper survival function.

Most of the survival models assume that all individuals experience failure with equal risk. Yet, in practice all individuals are expected to failure with varying risk, because there are many other factors which will influence the failure time, for example genetic, and life style. These factors are typically unknown, and hence cannot be explicitly included in the analysis. Models, which take into account the unknown factor between individuals are known as frailty models. Frailty is defined as unknown, unobservable, multiplicative factor acting on the survival function. Models that allow cured fraction and take into account the frailty between individuals are named as frailty mixture survival model, a combination of cured model and frailty model, where the short-term survivors follow the frailty survival model. In this study gamma frailty mixture model is focussed.

2 Frailty Mixture Model

2.1 Frailty Mixture Model

Assume W denote a non-negative frailty random variable with distribution function *F (w),* then the hazard function for an individual with frailty W at time *t* is given as

$$h(t|W = w) = wh_0(t)e^{\beta^T X} \tag{1}$$

where $h_0(t)$ is a constant function to all individuals, X is a vector of covariates, while β is a vector of coefficients. Therefore, as stated in Price and Manatunga [1] the survival function is given as

$$\begin{aligned} S(t) &= \int \{B(t)\}^w dF(w) \\ &= L(s) \end{aligned} \tag{2}$$

where $L(s) = E(e^{-sw})$ denote the Laplace transform of the frailty distribution and $s = -logB(t) = H_0(t)$ where $H_0(t)$ is the cumulative baseline function and *B(t)* is the baseline survival function. Distribution of W has support strictly greater than zero indicating positive risk for all individuals. Hence the individual who have higher risk will have higher frailty. In standard frailty model the frailty variable W is assumed to follow a parametric distribution or positive stable distribution such as gamma distribution, and these frailty distributions do not allow any individuals to have zero risk, in other words the general frailty model do not allow for a cured fraction. To include a cured fraction into frailty the studied population was divided into long-term survivors with probability *1-p* and sub-group of short-term survivors with probability *p*, who are subject to varying risk that measured through a frailty term. The survival model is named as Frailty mixture survival model.

2.2 Gamma Frailty and Gamma Frailty Mixture Survival Model

Assume the frailty W follow gamma distribution where

$$f(w) = w^{\alpha-1}\theta^{\alpha}e^{-w\theta}/\Gamma(\alpha) \tag{3}$$

with $\alpha,\ \beta > 0$. To ensure *E(Y)=1* we set $\alpha = \beta.$ The Laplace transform of *Y* for unit mean is

$$L(S) = \left(1 + \frac{s}{\theta}\right)^{-\theta} \tag{4}$$

where $s = -logB(t) = H_0(t)$. Thus, from (2) the survival function for gamma frailty survival model is given as

$$S(t) = \left(1 + \frac{H_0(t)}{\theta}\right)^{-\theta} \tag{5}$$

while the gamma frailty mixture survival model is given as

$$S(t) = 1 - p + p\left(1 + \frac{H_0(t)}{\theta}\right)^{-\theta} \tag{6}$$

Price and Manatunga [2] assumed a parametric distribution for the cumulative baseline hazard function $H_0(t)$ to obtain estimates in model (6) but they do not take into account the effect of covariates in both cured and frailty. Therefore, in this study we propose to include covariates in the frailty part and model (6) will be revised as follows:

$$S(t) = 1 - p + p\left(1 + \frac{H_0(t)\exp(\boldsymbol{bx})}{\theta}\right)^{-\theta} \tag{7}$$

where $\boldsymbol{b}$=$\{b_0, b_1, \dots, b_m\}$ is a vector of unknown parameters. Similar to cox regression model we assume that covariates influence the survival time with *exp(**bx**).* If we assume $H_0(t)$ to be a parametric distribution say weilbull distribution, model (7) can be estimated using the maximum likelihood method, similar to the approach applied in Price and Manatunga [1]. However, it is more convenient to use a semiparametric method if a parametric assumption is difficult to justify. It will also make the model applicable without knowing any information about the baseline survival function.

In this study, we focused in both semiparametric and parametric methods for model (7). For parametric, we utilized newton raphson approach to obtain the maximum likelihood estimator while semiparametric we utilized EM algorithm. A simulation study was carried out to compare and evaluate the performance of both semiparametric and parametric methods.

3 Result and Discussion

3.1 Parametric Method

The baseline survivor function is assumed to follow a weibull distribution, $H_0(t) = -\lambda t^{\alpha}$, thus model (7) can be expressesd as

$$S(t) = 1 - p + p\left(1 + \frac{-\lambda t^{\alpha}\exp(\boldsymbol{bx})}{\theta}\right)^{-\theta} \tag{8}$$

and the likelihood function can be expressed as

$$L\big((T|\eta)\big) = \prod_{j=1}^{n}\left\{p\alpha\lambda t_j{}^{\alpha-1}e^{\boldsymbol{bx}}\left(1 + \frac{\lambda t_j{}^{\alpha}e^{\boldsymbol{bx}_j}}{\theta}\right)^{-(\theta+1)}\right\}^{\delta_j}\left\{1 - p + p\left(1 + \frac{\lambda t_j{}^{\alpha}e^{\boldsymbol{bx}_j}}{\theta}\right)^{-\theta}\right\}^{1-\delta_j} \tag{9}$$

where $\eta = \{\alpha, \lambda, \theta, \rho, b\}$ is the vector of unknown parameter, *T* is the vector of survival time, *t* for *j=1,...,n*, δ_j denote the censor indicator, $\delta_j = 1$ for noncensored data and $\delta_j = 0$ for censored data and x_j are covariates. Newton-Raphson approach had been utilized in this study to estimate the unknown parameters. The variance-covariance matrix can be obtained by using observed fisher information matrix which are the minus hessian matrix.

3.2 Semiparametric Method

With the value of the frailty w_i, model (7) can be expressend as

$$S(t_i) = 1 - p + pexp\left(-w_i H_o(t_i)e^{bx_i}\right) \tag{10}$$

Similar to Y.Peng and J.Zhang [2], let $y_i = 0$ if cured and $y_i = 1$ if the individual experience failure. Therefore, the complete likelihood function is given as

$$L = (1-p)^{1-y_i}\left[pexp\left(-w_i H_o(t_i)e^{bx_i}\right)\left(w_i h_o(t_i)e^{bx_i}\right)^{\delta_i}\right]^{y_i} f(w_i) \tag{11}$$

where $f(w_i)$ is the density function of gamma distribution. EM algorithm is used to estimate $U = \{b, \theta, H_o(t)\}$ in model (7). The complete log- likelihood can be expressed as $\log(l) = l1(p) + l2\big(b, H_o(t)\big) + l3(\theta)$, where

$$l1(p) = \sum_{i=1}^{n} [(1 - y_i)\log(1 - p) + y_i \log(p)] \quad (12)$$

$$l2\big(b, H_o(t)\big) = \sum_{i=1}^{n} [-w_i y_i H_0(t_i) e^{bx_i} + \delta_i \log\big(h_0(t_i)\big) + bx_i] \quad (13)$$

$$l3(\theta) = \sum_{i=1}^{n} [\theta log\theta - \log\big(\Gamma(\theta)\big) - w_i\theta + (\,\delta_i + \theta - 1)\log(w_i)] \quad (14)$$

The E-step of EM-algorithm is to compute the conditional expectation of the complete likelihood function with respect to y_i and w_i with $U = U^j$ at j_{th} iteration. From equation (9)

$$\pi_i = E(\big(y_i \big| U^j, O\big) = \delta_i + (1 - \delta_i)\frac{p\left(1+\frac{H_0(t_i)\exp(bx_i)}{\theta}\right)^{-\theta}}{1-p+p\left(1+\frac{H_0(t_i)\exp(bx_i)}{\theta}\right)^{-\theta}} \quad (15)$$

with $U = U^j$ and the observed data *O*. For conditional expectation of $w_i, \log(w_i)$ and $w_i y_i$,

$$w_i \,\Big|y_i = 0, U^j, O \sim Gamma\left(\delta_i + \theta, \tfrac{1}{\theta}\right) \quad (16)$$

$$w_i \,\Big|y_i = 1, U^j, O \sim Gamma\left(\delta_i + \theta, \tfrac{1}{\theta + H_0(t_i)\exp(bx_i)}\right) \quad (17)$$

Therefore,

$$E\big(w_i \big| U^j, O\big) = E(w_i \big| y_i = 0, U^j, O)(1 - \pi_i) + E(w_i \big| y_i = 1, U^j, O)(\pi_i)$$

$$E\big(\log(w_i)\big| U^j, O\big) = E(\log(w_i)\big| y_i = 0, U^j, O)(1 - \pi_i) + E(\log(w_i)\big| y_i = 1, U^j, O)(\pi_i)$$

$$E\big(y_i w_i \big| U^j, O\big) = E(y_i w_i \big| y_i = 0, U^j, O)(1 - \pi_i) + E(y_i w_i \big| y_i = 1, U^j, O)(\pi_i)$$

The conditional distribution of $w_i|y_i$ is a gamma distribution, hence,

$$E\big(w_i \big| U^j, O\big) = \left(\frac{(\delta_i + \theta)\pi_i}{\theta + H_0(t_i)\exp(bx_i)}\right) + \frac{\delta_i + \theta}{\theta}(1 - \pi_i) \quad (18)$$

$$E\big(\log(w_i)\,\big| U^j, O\big) = \pi_i[\phi_i - \log(\theta + H_0(t_i)\exp(bx_i))] + (1 - \pi_i)[\phi_i - \log(\theta)] \quad (19)$$

$$E\big(y_i w_i \big| U^j, O\big) = \frac{(\delta_i + \theta)\pi_i}{\theta + H_0(t_i)\exp(bx_i)} \quad (20)$$

where $\phi_i = \psi(\delta_i + \theta)$ is the digamma function. Let $A_i = E\big(w_i \big| U^j, O\big)$, $B_i = E\big(\log(w_i)\big| U^j, O\big)$ and $C_i = E\big(y_i w_i \big| U^j, O\big)$ then equation (11), (12), and (13) can be expressed as

$$l1(p) = \sum_{i=1}^{n} [(1 - \pi_i)\log(1 - p) + \pi_i \log(p)] \quad (21)$$

$$l2\big(b, H_o(t)\big) = \sum_{i=1}^{n} [-C_i H_0(t_i) e^{bx_i} + \delta_i \log\big(h_0(t_i)\big) + bx_i] \quad (22)$$

$$l3(\theta) = \sum_{i=1}^{n} [\theta log\theta - \log(\Gamma(\theta)) - A_i\theta + B_i(\, \delta_i + \theta - 1)] \tag{23}$$

The M-step of EM-algorithm is to maximize the expected we computed which are (21)-(23) to update U. Maximizing (21) with respect to p can be done by using maximum likelihood. As mentioned in Peng [3] to maximize (23) with respect to b it can be completed by the cox regression proportional hazard model with additional covarite $\log(C_i)$ with coefficient equal to 1. To maximize (22) with respect to θ we can employ Newton-Raphson. Lastly we can employ Nelson-Aelan approach to estimate $H_0(t)$ which is,

$$\widehat{H}_0(t) = \sum_{t_i<t} \frac{d_i}{\sum_{j\in R_i} C_i \exp(bx_i)} \tag{24}$$

where d_i denotes the number of uncensored times at t_i and R_i is the at risk set at t_i.

3.3 Simulation Study

A simulation study was conducted to compare the performance of both semiparametric and parametric method. In this simulation study, 50 data sets with sample size of 50 were generated from a gamma frailty mixture survival model with the baseline survival function assumed to be Weibull distribution. We used $\theta = 0.5$, $\lambda = 6$, and $\alpha = 2$ while x was generated from standard normal distribution and coefficient $b = \log(2)$. Finally, δ was drawn from binomial distribution with $p = 0.8$. The results obtained from semiparametric and parametric are shown in Table 1. As mentioned in section 3, using parametric method, the baseline survival function is assumed to follow Weibull distribution, while using semiparametric method, the cummulative baseline hazard function is assumed to be unknown since in some cases a parametric assumtionis diffuclt to justify. Therefore, there are 5 unknown parameters for parametric method while semiparametric method has 3 unknown parameters.

As shown in Table 1 overall the estimators of parametric method are closer to the true value as compared to semiparametric estimators. The different between $\hat{b}$, $\hat{\theta}$ and the true value are smaller as compared to semiparametric method while the different between $\hat{p}$ and the true value is bigger as compared to semiparemetric method. However, the mean square error, MSE for $\hat{b}$, $\hat{\theta}$, $\hat{\alpha}$, and $\hat{\lambda}$ parametric estimators are higher than the semiparametric estimators.

Table 2 shown the mean square error of semiparametric and parametric survival function with the survival function obtain with the true value of the unknown parameter. The average of the mean square error of parametric survival function is smaller as compared to semiparametric survival function which means that parametric survival function fit the survival function with true value better than semiparametric model.

Table 1 Comparison of Semiparametric and Parametric Methods

	$\hat{p}$		$\hat{b}$		$\hat{\theta}$		$\hat{\alpha}$		$\hat{\lambda}$	
	$\bar{\hat{p}}$	MSE	$\bar{\hat{b}}$	MSE	$\bar{\hat{\theta}}$	MSE	$\bar{\hat{\alpha}}$	MSE	$\bar{\hat{\lambda}}$	MSE
Semi Parametric	0.8120	0.005	0.5540	0.0970	0.6500	0.0460	-	-	-	-
Parametric	0.7800	0.003	0.6899	0.1323	0.5347	0.1119	2.070	0.2374	4.3562	0.6869
True Value	0.8000	-	0.6914	-	0.5000	-	2.000		4.0000	-
Different between true value and estimators										
Semi Parametric	0.0120		-0.1374		0.1500		-		-	
Parametric	-0.0200		0.0015		0.0347		0.0700		0.3562	

Table 2 Mean Square error of Survival function

m	Parametric	Semi parametric	m	Parametric	Semi parametric
1	0.002667	0.028556	27	0.00192	0.033171
2	0.00353	0.041243	28	0.002609	0.026831
3	0.000589	0.042353	29	0.000379	0.047723
4	0.005358	0.03155	30	0.000441	0.027118
5	0.001562	0.04061	31	0.000962	0.045304
6	0.003034	0.022069	32	0.005082	0.046949
7	0.002211	0.024704	33	0.000917	0.025755
8	0.004199	0.020282	34	0.002825	0.037519
9	0.002397	0.028435	35	0.001777	0.024612
10	0.001415	0.039546	36	0.00057	0.028056
11	0.001692	0.040139	37	0.007034	0.052099
12	0.009333	0.048095	38	0.002183	0.035853
13	0.00389	0.03883	39	0.001766	0.032611
14	0.003402	0.048016	40	0.001392	0.032805
15	0.004356	0.023503	41	0.005725	0.033488
16	0.003614	0.033825	42	0.001404	0.048747
17	0.005329	0.052807	43	0.007105	0.042347
18	0.00654	0.044654	44	0.000197	0.044122
19	0.013724	0.046868	45	0.000228	0.042682
20	0.002382	0.023275	46	0.001796	0.041549
21	0.000428	0.04391	47	0.001473	0.0308
22	0.005831	0.044699	48	0.002725	0.033739
23	0.004648	0.041837	49	0.001307	0.027147
24	0.003053	0.027209	50	0.003448	0.084063
25	0.002932	0.020523	average	0.002008	0.019349
26	0.00231	0.069895			

5 Conclusion

In this study, we compared semiparametric and parametric model in Gamma frailty mixture survival mode. In parametric method, the baseline survival function was assumed to follow Weibull distribution and newton-raphson approach been employed to estimated $U = \{\alpha, \lambda, \theta, b, p\}$. In semiparametric method the baseline cumulative hazard function was assumed to be unknown and we utilized EM-algorith to estimated θ, b, and p while Nelson-Aalen approach been employed to estimated the baseline cumulative hazard function

The estimators of both parametric and semiparametric methos were closed to the true value, however the mean square error of semiparametric model is smaller than mean square error of parametric methods. Although the mean square error of parametric estimator is higher, the mean square error of parametric survival function is smaller compare to semieparametric survival function. Hence, both semiparametric and parametric model is comparable and parametric method fit the survival function with true value better than semiparametric method. Parametric method is easier to understand and the estimator can be obtained easily with newton-raphson, while semiparametric method is much complicated.

This study is limited to small sample size and small number of data set generated. It is strongly recommend for future study bigger sample size and more data set should be generated and the researcher should consider the covariates effect at the cured fraction.

References

[1] Price, D. L., and Manatunga, A. K. (2001). Modelling Survival Data with a Cured Fraction Using Frailty Models. *Statistics in Medicine.* 20, 1515-1527.

[2] Peng, Y. and Zhang, J. (2008). Estimation Method of the Semiparametric Mixture Cure Gamma Frailty Model. *Statistics in Medicine.* 27, 5177-5194.

[3] Peng, Y. (2003). Fitting Semiparametric Cure Models. *Computational Statistics and Data Analysis.* 41, 481-490.