MATEMATIKA, 2008, Volume 24, Number 2, pp. 231–242

Cross-sectional and Longitudinal Approaches in a Survival Mixture Model

[1]**Zarina Mohd Khalid & [2]Byron J.T. Morgan**
[1]Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia.
[2]Institute of Mathematics, Statistics and Actuarial Science, University of Kent, England.
[1]e-mail: zarina@fs.utm.my

Abstract In this article, we explore the performance of estimators in mixture survival model using simulated data recorded cross-sectionally and longitudinally. We use the maximum likelihood estimation approach in estimating the unknown model parameters. We found, in particular, that the maximum likelihood estimator for the proportion of long-term survivors in longitudinal setting gain better efficiency and precision for a certain distance of recording time.

Keywords Cross-sectional; longitudinal; mixture; survival; maximum likelihood estimate.

1 Introduction

Standard survival function is typically used to describe the survival experience of individuals related to a particular event of interest, such as death, occurrence of complications, relapse, and so on. Here, the population is assumed to consist of the individuals who are all at risk of failure. A mixture survival function, however, extends the standard function by allowing the possibility of long term survivorship of an individual to be incorporated at the outset. In this case, the population is assumed to consist of two subgroups of individuals: susceptible or at-risk individuals and long-term survivors who will never fail. This is practically true in the study of occurrence of complications, for example, where a proportion of patients may never get the complications in long term.

A study by Maller & Zhou [1] has explored survival analysis with long-term survivors. They were particularly interested in estimating the proportion of long-term survivors in the survival mixture model.

Likewise, Yu et al. [2] also explored the same modelling approach and concentrated on estimating the proportion of long-term survivors. Yu et al. [2] further concluded that the estimate of the proportion of long-term survivors is sensitive to the choice of distributions of the survival times for susceptible patients and the length of follow-up time. Price & Manatunga [3] did also explored and applied a number of survival models which include the survival mixture models that incorporate frailty or random effect in the susceptible survivor function and a compound Poisson model in their research. However, these studies did not incorporate any covariates in the selected models.

Another work by Young et al. [4] also adopted a mixture survival model to analyze data related to the occurrence of retinopathy complications in Type I diabetic patients. In their study however, they have incorporated several covariates in both survival functions for long-term and susceptible patients in gaining the knowledge about the prognostic factors that accelerate the progression of the complication.

The analyses in all above studies, nonetheless, were carried out cross-sectionally, that is the data used were recorded at a single time point, for example at the baseline visit in Young et al. [4]. The estimators resulting from cross-sectional modeling may be inaccurate and imprecise compared to the case where data collected longitudinally are utilized. The rational is that the longitudinally recorded data may provide better information about the state of every individual, and thus will result in more accurate estimation of model parameters for better decision making.

The aim of our study is thus to compare the performance of estimators in a mixture survival model that is analyzed cross-sectionally and longitudinally. In addition, we also incorporate a vector of covariates in the model through the mixture component and the survivor function for the susceptible individuals. For analysis, we use simulated data and the simulation procedure will be described.

2 Mixture Survival Model

Suppose the non-negative response variable is T which represents the time from a defined origin to an event of interest. The mixture survival function is given as

$$S(t) = \rho + (1-\rho)S^*(t), \qquad t > 0$$

where $S^*(t)$ is the standard proper survival function and ρ denotes the proportion of long-term survivors and $\rho \in [0,1]$. In our study, we assume T to follow a two-parameter Weibull [5] distribution having real positive spread and shape parameters denoted as μ and δ respectively. The proper survival function is then $S^*(t) = \exp(-[\mu t]^\delta)$ and hence,

$$S(t) = \rho + (1-\rho)\exp(-[\mu t]^\delta) \tag{1}$$

As mentioned in Section 1.1, we introduce covariates in the model above through the spread parameter μ and the long-term survival parameter ρ. Here,

$$\mu = \exp(-[\alpha_0 + \alpha_1 \mathbf{X}]) \tag{2}$$

and

$$\rho = \frac{1}{1+\exp(\beta_0 + \beta_1 \mathbf{X})} \tag{3}$$

where $\mathbf{X}$ is a vector of explanatory variables introduced into ρ and μ with their corresponding coefficients α_1 and β_1 respectively. We further assume that the explanatory variables $\mathbf{X}$ are distributed with standardized multivariate normal distributions. Furthermore, for an unconstrained estimation procedure, the shape parameter δ is reparameterised to its natural logarithm, $\delta_0 = \log_e(\delta)$ where $\delta_0 \in \mathbb{R}$. The vector of unknown parameters that needs to be estimated, say θ, is then $(\alpha_0,\ \alpha_1,\ \beta_0,\ \beta_1,\ \delta_0) \in \mathbb{R}^5$.

3 Simulation

We use simulated data for analysis. In order to simulate the data of interest, we assume that the vector of unknown parameters has a set of real values in the model of interest. In this study, we choose $\theta = [2, -0.25, 1.5, 0.5, 0.25]$ where the parameter values resemble the MLEs obtained in the real data analysis in [6] where any increase in any of the covariates at a specific time t, will lead to a reduction in the chance of surviving the event of interest.

Firstly, we assume that there are N individual individuals in the study which consists of $\rho 100\% (= N_0)$ long-term survivors and $(1-\rho)100\% (= N_1)$ susceptible individuals at the beginning of study period, when $t = 0$. Needless to say, N_0 remains the same throughout the study period.

We use a table-look-up method to simulate N_0, whereas its complement is simply given by $N_1 = N - N_0$. The failure time for susceptible individuals, on the other hand, is simulated by using the inversion method from its cumulative density function (*cdf*), subject to the condition that the *cdf* is a proper function. We know that $F(t) = 1 - S(t)$ where $S(t)$ is as given in Equation (1). Therefore,

$$\begin{aligned} F(t) &= 1 - S(t) \\ &= 1 - \Big[\rho + (1-\rho)S^*(t)\Big] \\ &= (1-\rho)\Big[1 - S^*(t)\Big] \end{aligned}$$

where $S^*(t)$ is the survivor function for the susceptible individuals. However, as $t \to \infty$, $S^*(t) \to 0$ and hence $S(t) \to \rho$, so that $F(t) \to (1-\rho)$. This shows that $F(t)$ is not a proper function. However, it can clearly be seen that $F(t)/(1-\rho) \to 1$ as $t \to \infty$, where

$$\begin{aligned} \frac{F(t)}{1-\rho} &= 1 - S^*(t) \\ &= F_0(t) \end{aligned}$$

which is the *cdf* of the failure times for the susceptible individuals. Consequently, we will instead simulate t from this *cdf* which is a proper function, i.e. t can be simulated directly from the susceptible *cdf* given by $F_0(t) = 1 - S^*(t)$. Now, suppose that U is a uniform random variable with its values $\{u_i\}$ ranging in $(0,1)$. By the inversion method we have

$$\begin{aligned} t_i &= F_0^{-1}(u_i) \\ &= \frac{1}{\mu_i}\Big[-\log_e(1-u_i)\Big]^{\frac{1}{\delta}} \end{aligned}$$

or more efficiently,

$$t_i = \frac{1}{\mu_i}\Big[-\log_e(u_i)\Big]^{\frac{1}{\delta}}$$

for $i = 1, 2, \ldots, N_1$.

The i-th susceptible individual will have failed or still be surviving the event of interest during a time of examination, say at τ_j for $j = 1, 2, \ldots, k$ visits. Therefore, by definition, at τ_j, the failure time $\{t_i\}$ for a susceptible individual can either be $t_i \leqslant \tau_j$ or $t_i > \tau_j$.

3.1 Cross-sectional Setting

In a cross-sectional analysis, we only have one reference time point, say τ_1 which is the length of time between the time origin and the reference point, at which there are N_{1f} susceptible individuals who fail before or at τ_1 and N_{1s} who still survive the event of interest at τ_1 so that their failure times are greater than τ_1. Notice that $N_{1f} + N_{1s} = N_1$ which is the total number of susceptible individuals in the study. The simulated values for N_{1f} and N_{1s} are given below:

$$N_{1f} = \sum_{i=1}^{N_1} I(t_i \leqslant \tau_1)$$

and

$$N_{1s} = \sum_{i=1}^{N_1} I(t_i > \tau_1)$$

respectively, where $I(.)$ is an indicator variable having values of 1 if true and 0 otherwise.

3.2 Longitudinal Setting

Extending the cross-sectional work to a longitudinal analysis, we now have more than one reference time point. Here, for illustration, we consider two reference points, say τ_1 and τ_2, where τ_2 is the length of time between the time origin and the second reference time point, and $\tau_2 > \tau_1$. Subsequently, in addition to the information known at τ_1, we will have further information about failures and survivors at the next reference point, τ_2. In other words, some of the survivors at τ_1 may develop the event of interest by τ_2 and the others would still survive at the second time. As a result, N_{1s} can be further divided into N_{2f} and N_{2s} where N_{2f} is the number of individuals who fail in $(\tau_1, \tau_2]$ and N_{2s} is the number of individuals who still survive past τ_2. In short, the simulated values of N_{2f} and N_{2s} can be expressed as follows:

$$\begin{aligned} N_{2f} &= \sum_{i=1}^{N_{1_s}} I(t_i \leqslant \tau_2 |\ t_i > \tau_1) \\ &= \sum_{i=1}^{N_{1_s}} I(\tau_1 < t_i \leqslant \tau_2) \end{aligned}$$

and

$$\begin{aligned} N_{2s} &= \sum_{i=1}^{N_{1_s}} I(t_i > \tau_2 |\ t_i > \tau_1) \\ &= \sum_{i=1}^{N_{1_s}} I(t_i > \tau_2) \end{aligned}$$

with $I(.)$ as an indicator variable and $I(t_i > \tau_1) = 1$.

In this simulation study, the choices of τ_1 and τ_2 are arbitrary, with the constraint that τ_2 is always greater than τ_1. We can expect that the earlier the time of the first visit, that is the smaller the value of τ_1 is, the smaller the number of failures at τ_1 will be. Similarly, a smaller gap, denoted by $\Delta = \tau_2 - \tau_1$, between τ_1 and τ_2 will generally result in a smaller number of failures in the interval $(\tau_1, \tau_2]$. In the latter situation, if Δ is too small, it may lead to a coincidence of cross-sectional and longitudinal approaches. This can also arise in another circumstance, where if τ_1 is large enough, all susceptible individuals may have failed by τ_1 and the survivors at τ_1 consist only of those long-term survivors who never develop the event of interest. Subsequently, there will be no failures in $(\tau_1, \tau_2]$, and hence the longitudinal likelihood function is reduced to the cross-sectional expression.

For the case with more than two visits, that is $k > 2$, we can generalise the above procedure for longitudinal analysis as follows: recall that at the start of the study, we have a total of N individuals of which N_0 and N_1 are the numbers of long-term survivors and susceptible individuals respectively. We index the visits by $j = 1, 2, \ldots, k$. When $j = 1$, N_1 consists of N_{1f} failures and N_{1s} survivors at τ_1. When $j = 2$, N_{1s} further divides into N_{2f} failures who fail in $(\tau_1, \tau_2]$ and N_{2s} survivors at τ_2. Similarly, when $j = 3$, N_{2s} consists of N_{3f} susceptible individuals who fail in $(\tau_2, \tau_3]$ and the other N_{3s} still survive at τ_3, and so on. Generally then, at $j = k$, $N_{(k-1)s}$ divides into N_{kf} individuals who fail in $(\tau_{k-1}, \tau_k]$ and N_{ks} individuals who still survive at τ_k, where

$$N_{kf} = \sum_{i=1}^{N_{(k-1)s}} I(\tau_{k-1} < t_i \leqslant \tau_k)$$

and

$$N_{ks} = \sum_{i=1}^{N_{(k-1)s}} I(t_i \geqslant \tau_k)$$

where $I(.)$ is 1 if true, or 0 otherwise. When k is large enough,

$$N_{1f} + N_{2f} + \ldots + N_{kf} = N_1$$

which is the number of susceptible individuals in the study, and $N_{ks} = 0$. Note that the number of long-term survivors, N_0, stays the same throughout the study period. Therefore, we can expect that when k is large enough, all susceptible individuals will fail and the survivors at τ_k are only long-term survivors.

For all cases, we choose an arbitrary $N = 500$ patients in the study. For each pair of values (τ_1, τ_2), we simulate the data set 250 times where at each time, the simulated data are used for estimation. Therefore, we will have 250 sets of the MLEs for the unknown parameters for each pair of values (τ_1, τ_2). The average of these MLEs and their standard errors are then used for summary and comparison.

4 Maximum Likelihood Estimation

We adopt the method of maximum likelihood estimation to estimate our model parameters. As such, the likelihood functions for both cross-sectional and longitudinal settings need to be explicitly stated. In this study, we assume all individuals have the same length of time between the initial visit and the first follow-up visit, so that for all i, $t_i = \tau_1$. Furthermore, we define $m_{1_{CS}} = N_{1f}$ and $m_{2_{CS}} = N_{1s} + N_0$. Without loss of generality, the likelihood function in cross-sectional setting can be expressed by the following expression:

$$L_{CS} \quad = \quad \prod_{i=1}^{m_{1_{CS}}} Pr(T_i \leqslant \tau_1) \prod_{i=m_{1_{CS}}+1}^{m_{1_{CS}}+m_{2_{CS}}} Pr(T_i > \tau_1) \tag{4}$$

where CS is short for “cross-sectional” and $m_{1_{CS}} + m_{2_{CS}} = N$.

For longitudinal setting, on the other hand, the likelihood function is defined as follows. Let us first define N_0 as the number of long-term survivors, N_{1f} as the number of failures at τ_1, N_{jf} as the number of failures in $(\tau_{j-1}, \tau_j]$ for $j = 2, 3, \ldots, k$ and N_{ks} as the number of susceptible individuals who still survive at τ_k. The longitudinal likelihood function can then be expressed as:

$$L_L \quad = \quad \prod_{i=1}^{N_{1f}} Pr(T_i \leqslant \tau_1) \times \prod_{j=2}^{k} \left[\prod_{i=1}^{N_{jf}} Pr(\tau_{j-1} < T_i \leqslant \tau_j) \right] \times \prod_{i=1}^{N_{ks}+N_0} Pr(T_i > \tau_k) \tag{5}$$

Here, our investigation focuses on the case of two follow-up visits. Similar to the cross-sectional approach, we assume that all individuals have the same length of time between the initial visit and the first follow-up visit denoted by τ_1. We also assume the same length of time between the initial visit and the second follow-up visit denoted by τ_2. Furthermore, we define $m_{1_L} = N_{1f}$, $m_{2_L} = N_{2f}$ and $m_{3_L} = N_{2s} + N_0$. The likelihood function therefore has the following form:

$$L_L \quad = \quad \prod_{i=1}^{m_{1_L}} Pr(T_i \leqslant \tau_1) \prod_{i=m_{1_L}+1}^{m_{1_L}+m_{2_L}} Pr(\tau_1 < T_i \leqslant \tau_2) \prod_{i=m_{1_L}+m_{2_L}+1}^{m_{1_L}+m_{2_L}+m_{3_L}} Pr(T_i > \tau_2).$$

where L stands for “longitudinal” and $m_{1_L} + m_{2_L} + m_{3_L} = N$. It can clearly be seen that there is an extra (middle) term in the longitudinal likelihood function which in part indicates the extra information employed in the estimation procedure from the observation at τ_2. Moreover, if $\tau_2 \to \tau_1$, then $Pr(\tau_1 < T_i \leqslant \tau_2) \to 0$, and hence $L_L \to L_{CS}$.

5 Results

After the simulated data set is achieved, we use these data to estimate the unknown parameters by maximizing the likelihood function with respect to the unknown parameters, and compare the maximum likelihood estimates (MLEs) of the parameters with their assumed true values. We then compare the accuracy and precision of these MLEs between the two modeling approaches by graphical displays, which are omitted in this article.

The accuracy is measured by the distance between the MLEs and the assumed true values for an unknown parameter θ. As such, we will use *bias* to determine the accuracy of an estimator, where *bias* is commonly defined as the difference between the expected value of an estimator and the true value, expressed as

$$bias = \mathrm{E}[\widehat{\theta}] - \theta$$

where $\widehat{\theta}$ is the maximum likelihood estimator for θ. A smaller *bias* indicates that an estimator is closer to the true value on average and hence more accurate.

Although *bias* can be used to measure the accuracy of an estimator, it is the mean square error (MSE) of the estimator that provides a better assessment of the quality of an estimator, particularly in a simulation study where the true parameter values are assumed known at the outset. The MSE of an estimator is known as the expected squared deviation of the estimated parameter value from the true parameter value, and using a standard notation for a scalar parameter it can be decomposed into the following form:

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + (bias)^2.$$

For comparison between single parameter cross-sectional and longitudinal estimators, we will use a relative efficiency measure (e_r) defined as the ratio of the mean-square error of the longitudinal estimator to that of the cross-sectional estimator. The relative efficiency measure e_r is defined as:

$$e_r = \frac{\mathrm{MSE}(\widehat{\theta}_L)}{\mathrm{MSE}(\widehat{\theta}_{CS})}$$

where the numerator is the MSE for the longitudinal estimator, $\widehat{\theta}_L$, and the denominator is the MSE for the cross-sectional estimator, $\widehat{\theta}_{CS}$. When $e_r = 1$, both estimators are equally efficient. However, if $e_r < 1$ then this implies that $\mathrm{MSE}(\widehat{\theta}_L)$ is smaller, and hence $\widehat{\theta}_L$ is more efficient than $\widehat{\theta}_{CS}$. On the other hand, if $e_r > 1$ then this implies that $\mathrm{MSE}(\widehat{\theta}_L)$ is larger and therefore $\widehat{\theta}_{CS}$ is more efficient than $\widehat{\theta}_L$, which we do not anticipate.

5.1 Fixed τ_1 and Varying Follow-up Distance, Δ

Consider the following three cases listed below:

Case 1 : $\tau_1 = 5$ and $\tau_2 = 6, 7, \ldots, 45$, so that $\Delta = 1, 2, \ldots, 40$
Case 2 : $\tau_1 = 10$ and $\tau_2 = 11, 12, \ldots, 45$, so that $\Delta = 1, 2, \ldots, 35$
Case 3 : $\tau_1 = 20$ and $\tau_2 = 21, 22, \ldots, 45$, so that $\Delta = 1, 2, \ldots, 25$.

In each case, τ_1 is fixed but τ_2 increases which leads to a wider distance, Δ, between τ_1 and τ_2. We can expect that $m_{1_{CS}} = m_{1_L}$ for all cases and constant when plotted against τ_2.

The average of simulated $m_{2_{CS}}$ values from 250 iterations is constant in each case, as expected, since τ_1 is unchanged. Between cases however, $m_{2_{CS}}$ at $\tau_1 = 20$ has the highest value. This is simply because more failures will have been identified by then.

The simulated m_{2_L} values however are increasing at a decreasing rate in all cases when Δ increases with the values in Case 1 being higher than those in cases 2 and 3. In the last two cases, τ_1 is longer and hence more failures have been included in m_{1_L}, leaving a smaller fraction of patients who would have failed in $(\tau_1, \tau_2]$.

As for m_{3_L}, its values are decreasing in all cases and approach a value representing the number of long-term survivors in the sample since more susceptible individuals have been identified to have failed by a large value of τ_2 and leaving only the long-term survivors in the sample.

Table 1: MLEs from Cross-sectional Analysis Approach Using Simulated Data for the Single Reference Time Points Given by $\tau_1 = 5, 10$, and 20

Parameters	True values	MLEs	Time at the FV after diagnosis		
			$\tau_1 = 5$	$\tau_1 = 10$	$\tau_1 = 20$
α_0	2	Mean	1.9202	1.9549	1.8740
		s.e.	3.7196	2.5831	3.5835
		|bias|	0.0798	0.0451	0.1260
		MSE	13.8418	6.6744	12.8573
α_1	-0.25	Mean	-0.4123	-0.4296	-0.3888
		s.e.	2.0424	2.0630	1.8837
		|bias|	0.1623	0.1796	0.1388
		MSE	4.1977	4.2882	3.5676
β_0	1.5	Mean	1.8582	1.9550	1.8253
		s.e.	1.9113	1.4585	0.9473
		|bias|	0.3582	0.4550	0.3253
		MSE	3.7814	2.3342	1.0032
β_1	0.5	Mean	0.8591	0.8211	0.7072
		s.e.	0.7679	0.6131	0.4226
		|bias|	0.3591	0.3211	0.2072
		MSE	0.7186	0.4790	0.2215
δ_0	0.25	Mean	0.1871	0.1291	0.1649
		s.e.	5.5376	5.6073	5.5462
		|bias|	0.0629	0.1209	0.0851
		MSE	30.6690	31.4564	30.7676

5.1.1 Performance of Estimators in Cross-sectional Analysis

Table 1 lists the average, standard error, bias and MSE for the MLEs from the cross-sectional estimation approach. There is a mixture of performance for the estimators of α_0 and α_1 which are contained in the spread parameter μ, as τ_1 increases. For example, at $\tau_1 = 5$, $\widehat{\alpha}_0$ is more biased with a larger standard error when compared to that at $\tau_1 = 10$ and worst at $\tau_1 = 20$. However, $\widehat{\alpha}_1$ is the least biased at the latest time $\tau_1 = 20$ and more biased at $\tau_1 = 10$. The MSEs for $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ also demonstrate different patterns: for $\widehat{\alpha}_0$, its MSE is the smallest at $\tau_1 = 10$, but the MSE for $\widehat{\alpha}_1$ is the largest at the same τ_1.

As for δ_0, its estimator is the most biased when $\tau_1 = 10$ and the least biased at the earliest τ_1. The value of MSE for its MLEs, however, are substantially higher at all values of τ_1 which mainly result from very large standard errors, indicating a very poor quality and precision.

The MSE values for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ contained in the long-term survivor parameter, ρ, create an interesting pattern: as τ_1 gets larger, the values of MSE become smaller, indicating a better quality at a later time. This is consistent with the definition of ρ whereby as τ_1 increases, more failures are correctly classified and the remaining survivors are those who are truly long-term survivors, hence the estimate of the proportion will be more reliable at a much later time. This behaviour can be expected to be a function of the parameters in the survivor function for the susceptible patients.

5.1.2 Performance of Estimators in Longitudinal Analysis

Within the longitudinal analysis approach at different values of τ_1 and τ_2 for $\tau_1 = 5, 10$ and 20, and $\Delta = 1, 2, \ldots$ where $\Delta = \tau_2 - \tau_1$, the maximum likelihood estimators for the parameters involved in defining the susceptible survivor function, which in our illustration include α_0, α_1 and δ_0, acquire good performance in terms of attaining higher efficiency and better precision when τ_1 is not very large. The relative efficiency measure for the longitudinal estimators when $\tau_1 = 20$ is much higher than the corresponding values for the same longitudinal estimators when $\tau_1 = 5$ or 10. This suggests that the values of mean-square error for these longitudinal estimators at a later time have increased more relative to the corresponding value for the same longitudinal estimators at an earlier τ_1.

The longitudinal estimators for β_0 and β_1 are less biased and more precise when τ_1 is larger, and for all values of τ_1, the bias and standard errors reduce as Δ increases. Hence, this leads to smaller values of MSE for these estimators when τ_1 increases and reduces further as Δ becomes wider.

The relative efficiency measures for the longitudinal MLEs of β_0 and β_1 is higher than the corresponding values for the estimators of the other unknown parameters. A reasonable explanation for this event is that both longitudinal and cross-sectional estimators for the parameters defining the proportion of long-term survivors are gaining better efficiency and becoming closer to the true values, as well as becoming more precise, as τ_1 increases. Therefore, their values of MSE will not differ very much. In our example, however, the values of MSE for the longitudinal estimators are still more desirable than the MSE values for the cross-sectional estimators in all cases which further reveals that the longitudinal estimators once again achieve better efficiency and precision.

In summary, all longitudinal estimators are more efficient relative to the cross-sectional estimators since the relative efficiency measure is apparently lower than 1 in all cases.

5.2 Varying τ_1 and Fixed Follow-up Distance Δ

Repeating the simulation steps in section 5.1, we obtain the simulated data for use in the estimation procedure for this example. This time, however, τ_1 varies and Δ is fixed at 5, 10 and 20, so that the same distance is generated between the first and second visits, and varying estimates will occur from cross-sectional as well as from longitudinal analysis approaches at all different values of τ_1. In short, the three cases for illustration here are

Case 1 : $\tau_1 = 1, 2, \ldots, 30$ and $\tau_2 = 6, 7, \ldots, 35$, so that $\Delta = 5$
Case 2 : $\tau_1 = 1, 2, \ldots, 30$ and $\tau_2 = 11, 12, \ldots, 40$, so that $\Delta = 10$
Case 3 : $\tau_1 = 1, 2, \ldots, 30$ and $\tau_2 = 21, 22, \ldots, 50$, so that $\Delta = 20$.

This example is analogous to a moving window, instead of a wider window in the previous example. For all cases, the average number of failures, m_1, simulated at the first visit is nearly the same for both analyses, which are increasing at a decreasing rate as τ_1 increases. Next, the average number of patients who fail in $(\tau_1, \tau_2]$, that is m_{2_L}, in the longitudinal analysis is decreasing towards zero as τ_1 increases. This is different from the previous example where m_{2_L} is increasing due to a wider Δ. In the current example however, Δ is constant for all cases and therefore m_{2_L} should be expected to reduce as more failures are included in m_{1_L}. In short, m_{2_L} is inversely related to m_{1_L}. Furthermore, Case 3 has the highest m_{2_L} since Δ is the largest which allows more failures to be identified in the interval. Lastly, the number of survivors at τ_2 can be seen as decreasing and approaching a constant value represented by the number of long-term survivors, with Case 1 with smallest Δ, having the highest value of m_{3_L} for the longitudinal approach indicating that more susceptible patients do not yet fail by τ_2.

5.2.1 Comparisons for Parameters in $S^*(t)$

Now, we look at the MLEs for α_0, α_1 and δ_0 which defines $S^*(t)$. As in the situation in the earlier section, the MLEs for these parameters are more favourable when Δ is smaller at all values of τ_1. Indeed, when τ_1 is large and $\Delta = 20$, the longitudinal MLEs are becoming worse than when $\Delta = 5$ or 10. In all circumstances, however, the standard errors for these MLEs are still better in the longitudinal analysis approach as compared to those obtained from cross-sectional analysis.

Hence, the resulting MSEs for the longitudinal estimators are more desirable than those for the cross-sectional estimators which leads to the relative efficiency measures being less than unity. Note that, for this example as well, the longitudinal estimators perform better when τ_1 is earlier and Δ is smaller.

5.2.2 Comparisons for Parameters in ρ

Parameters in ρ behave in a similar way to before where later time periods provide better estimates and precision. This is indeed true for both analysis approaches. A more convincing set of MLEs are those obtained from the longitudinal approach when Δ is the largest.

6 Discussion

The results from our simulation study have demonstrated the different performances of the MLEs with their corresponding standard errors for different values of the unknown parameters.

Clearly, the MLEs for the parameters defining ρ gain better efficiencies and precisions when the first visit becomes later and the distance between the first and second times of visit becomes wider, provided that the time of the first visit is not so large that no information is gained within the period between the first and second visit.

As for the parameters defining the susceptible survivor function, which are the spread parameter μ and the shape parameter δ, it is shown that a combination of a reasonably earlier time of the initial visit and a smaller gap between the first and the second visit is preferable in gaining an optimal set of MLEs.

There seems to exist a contradicting phenomenon between these two components of the combined survivor function, whereby the longitudinal MLEs can be better for some parameters and not for the others at different values of (τ_1, τ_2), for example the estimators for the parameters defining the susceptible survivor function perform better at an earlier time and smaller gap between the first and second visit. The reverse however is true for the estimators of the unknown parameters defining the proportion of long term survivors in the data. Nonetheless, the combined results from the longitudinal approach are still desirable since the bias in most cases is small and the level of precision is evidently better in the longitudinal analysis which is further justified by the smaller-than-unity values of the relative efficiency measures.

The results obtained so far for the case with two reference time points for the longitudinal analysis calibrate what we expect about the longitudinal estimators and we gain a better understanding about the estimators' behaviours. These results however inevitably are a function of parameter values used in the simulation.

It is worthwhile noting again that the covariate values used in this simulation study for the longitudinal analysis are assumed fixed, and hence the values are the same at both visits. Since our objective in this study is to investigate and compare the performance of the estimators from the cross-sectional and longitudinal analysis, the longitudinal model assuming the fixed covariates should suffice. Furthermore, the assumption helps making the process of simulating the values of the response variables much simpler.

In practice however, we have at least two difficulties: one concerns the length of time between the initial visit, for example the time of diagnosis, and the time of the first follow-up visit which certainly varies between individuals. Similarly, the duration between the first and the next follow-up visits differ between individuals. This is why we consider a different combination of τ_1 and Δ in our investigation in order to get some ideas on how the estimators perform when such differences in the duration of time occur.

The other difficulty concerns the changing values of the covariates in the longitudinal approach. This is not covered in our simulation, but we need to be aware that changing covariate values can affect the performance of longitudinal parameters, particularly the proportion of long-term survivors, ρ, since analytically this parameter is only a function of x but not t. Therefore, by changing x, increasing the t values may not necessarily reduce ρ, and hence the longitudinal estimator for parameters involved in ρ may perform differently from the performance discussed so far. Furthermore, even in our example with the fixed covariate values, the relative efficiency measure for these estimators, though smaller than unity, are quite close to the unit value.

Our example can be generalised to the case with more than 2 visits for the longitudinal analysis, where at each subsequent visit the number of survivors decreases since more susceptible individuals would be known to fail after some additional time. Due to increasing information gained when the number of visits increases, we should expect that the estimators perform even better than those in the cross-sectional analysis or longitudinal analysis with a lower number of visits. The estimation for the longitudinal model with more than two visits however is more complex, particularly when it involves time-varying covariates in the model. Despite this complexity, it can be done and at this stage it is proposed for future work.

Generally from this investigation, we can see that the longitudinal estimators perform better than do the cross-sectional estimators and the increase in efficiency can be seen to increase substantially. We should anticipate the same performance to persist when we deal with the real data whereby the longitudinal estimators are preferable than the cross-sectional estimators.

References

[1] R.A. Maller and X. Zhou, *Survival Analysis with Long-term Survivors*, Wiley, New York, 1996.

[2] B. Yu, R.C. Tiwari, K.A. Cronin and E.J. Feuer, *Cure Fraction Estimation from the Mixture of Cure Models for Grouped Survival Data*, Statistics in Medicine, 23(2004), 1733–1747.

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

[4] P. Young, B.J.T. Morgan, P. Sonksen, S. Till and C. Williams, *Using a Mixture Model to Predict the Occurrence of Diabetic Retinopathy*, Statistics in Medicine, 14(1995), 2599–2608.

[5] W. Weibull, *A Statistical Distribution of Wide Applicability*, Journal of Applied Mechanics, 18(3)(1951), 293–297.

[6] Zarina M. Khalid, *Survival Modelling for Retinopathy in Type I Diabetes Mellitus*, Unpublished PhD Thesis, University of Kent, UK, 2005.