Malaysian Journal of Mathematical Sciences 10(S) August: 159–173 (2016)
Special Issue: The 7^{th} International Conference on Research and Education in Mathematics (ICREM7)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Simulation of Left-truncated and Case-k Interval Censored Survival Data with Time-Varying Covariates

Thirunanthini Manoharan *[1], Jayanthi Arasan [1,2], Habshah Midi[1,2], and Mohd Bakri Adam[1,2]

[1]*Department of Mathematics, Faculty of Science, 43400 UPM Serdang, Selangor, Malaysia*
[2]*Laboratory of Computational Statistics and Operations Research, Institute for Mathematical Research, 43400 UPM Serdang, Selangor, Malaysia*

E-mail: mthirunanthini@gmail.com
**Corresponding author*

ABSTRACT

This research focuses on simulation of left-truncated and case-k interval censored survival data from the log-normal model with a time-varying covariate. Left-truncated data usually arises in prevalence cohort study where randomly selected individuals from medical records may have contracted certain disease for some duration of time but are free from event of interest at time of entry into a survival study. In this research, we proposed a simulation methodology by fixing the percentage of truncation at 20% and 60% with the width of 4 months of inspection interval. The procedure was computationally demanding due to the presence of left-truncation and time-varying covariates. The suitability of the proposed method was assessed based on the bias, standard error and root mean square of the parameter estimates for the log-normal survival model.

Keywords: simulation, left-truncation, case-k interval censored, time-varying covariates and log-normal distribution.

1. Introduction

Left-truncation occurs in a clinical survival study when it is not feasible to observe a patient from the time of contraction of a certain disease but at some time point later which may be due to the study design, cost or time constraint. In other words, these individuals are not observed from the beginning of the study, but at some time point later, u. Also, the lifetimes of these individuals, $t \geq u$ as they have to survive long enough in order to be selected into the study. Subsequently, their lifetimes, t are said to be left-truncated at u. These individuals are then followed prospectively with fixed k inspection times where the exact event time is unknown except that it falls within an interval of (t_l, t_r) where $t \in (t_l, t_r)$ and $t_l \leq t_r$ with probability of 1. This type of data is known as left-truncated and case-k interval censored (LTIC) survival data , where left-truncated observations are existing cases (prevalence cohort) usually sampled from medical registry records.

Therefore, individuals may enter the study at random age or time points as the date of diagnosis may differ from one individual to the other, however, in the presence of left-truncation only those who are free from failure are observed by the researcher, refer Guo (1992). Additionally, other factors that affect t known as covariates, x are only considered from the time of entry into the study, refer Guo (1992), Klein and Moeschberger (2003), Lawless (1982). The lifetime after selection forms the response variable, hence the term prospective, refer Lawless (1982). Hence, the truncation time u contains no information on the lifetime t or t is independent of u.

On the other hand, time-dependent covariates vary over time or equally measured on regular basis for an individual in a study. By accommodating the record of a time-dependent covariate up to a specified time, say t enables a researcher to study the complete effect of these variables on the survival time T. As an example, accounting for the change in the level of covariates such as age, glucose level, blood pressure or tumour sizes provides the up to date affect of these variables on the hazard and survival rate of the individuals in the study, subsequently providing more reliable prognosis of the future life expectancy comparatively when these covariates are measured only at the time origin Nardi and Schemper (2003). Two types time-dependent covariates that are usually encountered in a survival study are the internal and external covariates. The former time-dependent covariate can be measured repeatedly for an individual over a specified time period of t for as long as the patient is still alive. Examples of such covariate includes a patient's blood pressure level, glucose level, red and white blood count, refer Collett (2003), Kalbfleisch and Prentice (2011).

Subsequently, external time-dependent covariates can be predicted independently which does not require a patient to be alive at the time of measurement. Following Kalbfleisch and Prentice (2011), an evident example for this type of covariate is the age factor which is measured at the time origin for an individual where the change in the value can be determined for any particular time interval without requiring the presence of this individual in the study. Also, some external covariates may influence the survival time of an individual at time t which however exist independently in the sense that the covariate's level at time t and after t is independent. Example of such covariate may exist in respiratory survival studies, where the presence of air pollutant may affect the life span of individuals with heart disease or lung cancer where the change in the air quality is independent of any particular individual in the study, refer Collett (2003), Kalbfleisch and Prentice (2011), Kiani and Arasan (2012).

Following Cox and Oakes (1984), Klein and Moeschberger (2003), Nardi and Schemper (2003), parametric models often remain a useful tool as they are fitted much faster and offers more efficient estimates under conditions such as dependency of the survival times on the covariates (either fixed or time-varying) and when parameter values are far from zero. Subsequently, simulation procedures enables a researcher to asses the performance of a proposed parametric estimator concurrently in determining suitable inferential procedures for the parameters in a specified model. This methodology is crucial in order to draw reliable, precise and important information from the sample data in hand.

In this research, we proposed simulation methodology for LTIC survival data with a time-varying covariate which mimics a lung cancer survival data. The following section discusses the simulation algorithm involved in simulating LTIC observations with time-varying covariates where the survival times, T are known to follow the log-normal lifetime distribution.

2. Survival and hazard function of left-truncated observations with a time-varying covariate

Let x_i be a covariate value for the i^{th} individual which is updated at the sequence of update time $\{\tau_{imj}\}$ with $j = 0, 1, ..., r_i$. The complete history of x_i covariate after each update time can be defined as $x_i = (x_{i0}, x_{i1}, ..., x_{ir_i})$ with covariate update times $\tau_{ij} = (\tau_{i0}, \tau_{i1}, ..., \tau_{ir_i})$. In other words, x_{i0} is the covariate baseline value at τ_{i0}, x_{i1} is the covariate value after the first update time τ_{i1} and x_{ir_i} is the covariate value after the r_i^{th} update time. Following

Petersen (1986) and Arasan and Lunn (2008), a time-varying discrete or continuous covariate x which is updated using a step-function stays constant at $x(\tau_j)$ within the interval $\left[\tau_j, \tau_{(j+1)}\right)$ and suffers a jump to $x(\tau_{j+1})$ at $\tau_{(j+1)}$.

Examples of such change occurs in covariate that changes from one level to another either dependently or independently, see Kiani and Arasan (2012). For instance, change in blood pressure before or after the update time could either increase or decrease independently. In contrary, staging of a disease will either continuously decrease or increase depending on the level before the update time. On a similar note, the covariate x_i stays constant within $\left[\tau_{ij}, \tau_{i(j+1)}\right)$ and changes to $x_{i(j+1)}$ in the next subsequent interval. For the i^{th} observation, let $\boldsymbol{x}_{i[t_i]}$ denote the complete history of covariate values up to time t_i. Following that, the survival function for the i^{th} left-truncated observation with lifetime $t_i \geq u_i$ and conditional on $\boldsymbol{x}_{i[t_i]}$ is given as follows:

$$\begin{aligned} S_{\boldsymbol{\theta}}\left[t_i | t_i \geq u_i, \boldsymbol{x}_{i[t_i]}\right] = \exp\Big[-\Big\{ & \int_{u_i}^{\tau_{i1}} h_{\boldsymbol{\theta}}(s_i|x_{i0})ds \\ & + \int_{\tau_{i1}}^{\tau_{i2}} h_{\boldsymbol{\theta}}(s_i|x_{i1})ds \\ & + ... + \int_{\tau_{ir_i}}^{t_i} h_{\boldsymbol{\theta}}(s_i|x_{ir_i})ds \Big\}\Big], \end{aligned} \tag{1}$$

with $h_{\boldsymbol{\theta}}(s_i|x_{ij})$ is the hazard rate evaluated at the sequence of update times τ_{ij}, $j = 0, 1, ..., r_i$. Note that the expression in (1) shows that the value of the covariate for the i^{th} left-truncated observation need to be observed from the time observations are recruited in the study; $t_i \geq u_i$ and $\boldsymbol{\theta}$ is the vector of parameters in a specified model.

Let us consider a model with at most one covariate update time, $j = 0, 1$ and two covariate levels, x_{i0} and x_{i1}. Therefore, the survival function in (1) can be further simplified to accommodate before and after covariate update time as follows:

$$S_{\boldsymbol{\theta}}\left[\tau_{i1} | \tau_{i1} \geq u_i, \boldsymbol{x}_{i[t_i]}\right] = \exp\left[-\left\{\int_{u_i}^{\tau_{i1}} h_{\boldsymbol{\theta}}(s_i|x_{i0})ds\right\}\right], \tag{2}$$

$$\begin{aligned} S_{\boldsymbol{\theta}}\left[t_i | t_i \geq \tau_{i1}, \boldsymbol{x}_{i[t_i]}\right] = \exp\Big[-\Big\{ & \int_{u_i}^{\tau_{i1}} h_{\boldsymbol{\theta}}(s_i|x_{i0})ds \\ & + \int_{\tau_{i1}}^{t_i} h_{\boldsymbol{\theta}}(s_i|x_{i1})ds \Big\}\Big]. \end{aligned} \tag{3}$$

Suppose that the lifetime of the i^{th} left-truncated observation follows a log-normal distribution. Following that, the hazard function before and after the covariate update time is derived from (2) and (3). These are given in (4) and (5) respectively:

$$h_{\boldsymbol{\theta}}\left[\tau_{i1}|\tau_{i1} \geq u_i, \boldsymbol{x}_{i[t_i]}\right] = \frac{\phi\left(\frac{\log(\tau_{i1})-(\beta_0+\beta_1 x_{i0})}{\sigma}\right)}{\tau_{i1}\sigma\left(1-\Phi\left(\frac{\log(\tau_{i1})-(\beta_0+\beta_1 x_{i0})}{\sigma}\right)\right)}, \tag{4}$$

$$h_{\boldsymbol{\theta}}\left[t_i|t_i \geq \tau_{i1}, \boldsymbol{x}_{i[t_i]}\right] = \frac{\phi\left(\frac{\log(t_i)-(\beta_0+\beta_1 x_{i1})}{\sigma}\right)}{t_i\sigma\left(1-\Phi\left(\frac{\log(t_i)-(\beta_0+\beta_1 x_{i1})}{\sigma}\right)\right)}, \tag{5}$$

with $\phi(\cdot)$ is the density function of the standard normal distribution.

3. Simulation of LTIC observations with time-varying covariates

The simulation study proposed by Kiani and Arasan, Kiani and Arasan (2012) and Balakrishnan and Mitra, Balakrishnan and Mitra (2014) is adopted and modified to mimic the small cell lung cancer survival data studied by Tai et.al, Tai et al. (2007) which provided a satisfactory fit with the log-normal distribution. We have considered cases where covariate levels are independent and dependent. The following assumptions hold with the simulation study:

1. The date of diagnosis are available for all the left-truncated observations.
2. All the observations are monitored continuously at fixed k inspection times and the baseline/initial value of a covariate are measured at first inspection times.
3. The lifetimes, t and left-truncation times, u are non-informative and independent of each other.
4. All the individuals were available for observations at all inspection times.
5. All the individuals were event free at the beginning time point of the study.

3.1 Simulation of the month of diagnosis

1. Fix the month of truncation or the beginning time point of the study, namely y.

2. Simulate a set of random number of months which basically represents the month of diagnosis of the lung cancer with unequal probabilities with replacement; before y_{b_l} and after the month of truncation y_{c_m} where $l = 1, 2, ..., n_1$ and $m = 1, 2, ..., n_2$. Note that y_{b_l} represents all prevalence cohort with $y > y_{b_l}$ and is fixed to yield a desired proportion of left-truncated observations. The remaining observations are incidence cohort, y_{c_m} observed from the beginning time point of the study with $y = 0$ and $y < y_{c_m}$. Note that this simulation study the total observation is $n = n_1 + n_2$.

3. Calculate the left-truncation time, $u_l = y - y_{b_l}$.

3.2 Simulation of the covariate update time

1. Simulate the update time, $\tau_{l1} \sim \exp(\lambda)$ and $\tau_{m1} \sim \exp(\lambda)$. In fact τ_{l1} and τ_{m1} can be simulated from any continuous random distribution. Adjust the value of λ to yield larger or smaller values of τ_{l1} and τ_{m1}.

2. Retain the value of τ_{l1} if and only if $y_{b_l} + \tau_{l1} \geq y$, otherwise these observations are removed and new values of τ_{l1} are simulated.

3.3 Simulation of the covariates

1. Independent covariates: Simulate first and second covariate level for the prevalence cohort $(x_{l0},\ x_{l1})$ and incidence cohort $(x_{m0},\ x_{m1})$ independently from the standard normal distribution.

2. Dependent covariates: Simulate at least two covariate levels independently from the standard normal distribution for the prevalence and incidence cohort.

   (a) Divide the z scores of the standard normal distribution into consecutive equal probability intervals. In this study, the z scores are divided into five equal probability intervals, known as quintiles of the standard normal distribution, e.g. $(-\infty, -1.15], (-1.15, -0.32], (0.32, 0.32], (0.32, 1.15]$ and $(1.15, \infty)$.

(b) Select the first covariate level $x_{l0}(x_{m0})$ and second covariate level $x_{l1}(x_{m1})$ for the prevalence (incidence) cohort based on these conditions: i) The second covariate level is higher than the first covariate level, e.g. $x_{l1} > x_{l0}$;$(x_{m1} > x_{m0})$.ii) The first and second level of the covariate should not fall within the same interval. These assumptions appear to be more realistic representing age or staging of a disease.

3.4 Simulation of the lifetimes

1. Simulate random values of $z_l \sim unif(0,1)$ and $z_m \sim unif(0,1)$ for $l = 1, 2, ..., n_1$ and $m = 1, 2, ..., n_2$.
2. Simulate lifetimes t_l based on the following conditions:

$$t_l = \begin{cases} \exp\left[\sigma\Phi^{-1}(Q_l) + (\beta_0 + \beta_1 x_{l1})\right], & z_l < R_l \\ \exp\left[\sigma\Phi^{-1}(1 - z_l) + (\beta_0 + \beta_1 x_{l0})\right], & \text{otherwise,} \end{cases} \tag{6}$$

with,

$$Q_l = 1 - \frac{z_l\left(1 - \Phi\left(\frac{\log\tau_{l1} - (\beta_0 + \beta_1 x_{l1})}{\sigma}\right)\right)}{\left(1 - \Phi\left(\frac{\log\tau_{l1} - (\beta_0 + \beta_1 x_{l0})}{\sigma}\right)\right)},$$

$$R_l = \left(1 - \Phi\left(\frac{\log\tau_{l1} - (\beta_0 + \beta_1 x_{l0})}{\sigma}\right)\right).$$

3. Simulate lifetimes t_m in the similar manner.
4. Retain the value of t_l if and only if $y_{b_l} + t_l \geq y$, otherwise the random variables τ_{l1}, (x_{l0}, x_{l1}), t_l and z_l are removed and new values of these random variables are simulated.

The combination of both observations from the prevalence and incidence cohort form the complete data set of size n with variables t_i, u_i, x_{i0}, x_{i1} and τ_{i1} for $i = 1, 2, ...n$.

4. Likelihood of LTIC survival data with log-normal lifetime distribution.

Consider observations from prevalence and incidence cohort who are monitored periodically at fixed k inspection times with a sequel of inspection times $a_{i1} \leq a_{i2} \leq ... \leq a_{ik}$. The lifetime, t_i of an i^{th} individual could fall within the following intervals with t_{L_i} and t_{R_i} are the left and right endpoints and $Pr(t_{L_i} \leq t_{R_i}) = 1$:

1. $t_{L_i} < t_i \leq t_{R_i}$ and $t_{L_i} < \tau_{i1} \leq t_{R_i}$, t_i is interval censored and covariates are updated.
2. $\tau_{i1} \leq t_{L_i} < t_i \leq t_{R_i}$, t_i is interval censored and covariates are updated.
3. $t_{L_i} < t_i \leq t_{R_i} < \tau_{i1}$, t_i is interval censored and covariates are not updated.
4. $\tau_{i1} \leq t_{L_i} < t_i < \infty$, t_i is right-censored and covariates are updated.
5. $t_{L_i} < t_i < \infty$ and $t_{L_i} < \tau_{i1}$, t_i is right-censored and covariates are not updated.
6. $\tau_{i1} \leq t_{R_i} - \epsilon \leq t_i \leq t_{R_i}$ with $\epsilon \in \Re^+$, t_i is observed exactly and covariates are updated.
7. $t_{R_i} - \epsilon \leq t_i \leq t_{R_i} < \tau_{i1}$ with $\epsilon \in \Re^+$, t_i is observed exactly and covariates are not updated.
8. $0 < t_i \leq a_{i1}$, t_i is left-censored when t_i occurs at an unknown time a_{i1} and after time origin. Left-censored observations are only observed among the incidence cohort and covariates are not updated.

Following that, the likelihood for both prevalence and incidence cohort with log-normal lifetime distribution is derived using (2) to (5) and given in (7) and

(8) respectively.

$$
\begin{aligned}
L(\boldsymbol{\theta}) = \prod_{i=1}^{n} & \left[\frac{(1-A_i)(1-B_i)-(1-C_i)(1-D_i)}{(1-U_i)(1-B_i)}\right]^{\delta I_i \delta G_i (1-\delta P_i)} \\
& \times \left[\frac{(1-D_i)(C_i-M_i)}{(1-U_i)(1-B_i)}\right]^{\delta I_i \delta G_i \delta P_i} \\
& \times \left[\frac{F_i-A_i}{(1-U_i)}\right]^{\delta I_i (1-\delta G_i)(1-\delta P_i)} \\
& \times \left[\frac{(1-D_i)(1-M_i)}{(1-U_i)(B_i)}\right]^{\delta R_i \delta G_i} \\
& \times \left[\frac{(1-A_i)}{(1-U_i)}\right]^{\delta R_i (1-\delta G_i)} \\
& \times \left[\frac{(1-D_i)V_i}{t_{R_i}\sigma(1-U_i)(1-B_i)}\right]^{\delta E_i \delta G_i} \\
& \times \left[\frac{Z_i}{t_{R_i}\sigma(1-U_i)}\right]^{\delta E_i (1-\delta G_i)},
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
L(\boldsymbol{\theta}) = \prod_{i=1}^{n} & \left[\frac{(1-A_i)(1-B_i)-(1-C_i)(1-D_i)}{(1-B_i)}\right]^{\delta I_i \delta G_i (1-\delta P_i)} \\
& \times \left[\frac{(1-D_i)(C_i-M_i)}{(1-B_i)}\right]^{\delta I_i \delta G_i \delta P_i} \\
& \times [F_i-A_i]^{\delta I_i (1-\delta G_i)(1-\delta P_i)} \\
& \times \left[\frac{(1-D_i)(1-M_i)}{(B_i)}\right]^{\delta R_i \delta G_i} \\
& \times [(1-A_i)]^{\delta R_i (1-\delta G_i)} \\
& \times \left[\frac{(1-D_i)V_i}{t_{R_i}\sigma(1-B_i)}\right]^{\delta E_i \delta G_i} \\
& \times \left[\frac{Z_i}{t_{R_i}\sigma}\right]^{\delta E_i (1-\delta G_i)} \times [F_i]^{\delta L_i},
\end{aligned}
\tag{8}
$$

with $\phi(\cdot)$ is the density function of the standard normal distribution and,

$$A_i = \Phi\left(\frac{log(t_{L_i}) - (\beta_0 + \beta_1 x_{i0})}{\sigma}\right)$$
$$B_i = \Phi\left(\frac{log(\tau_{i_1}) - (\beta_0 + \beta_1 x_{i1})}{\sigma}\right)$$
$$C_i = \Phi\left(\frac{log(t_{R_i}) - (\beta_0 + \beta_1 x_{i1})}{\sigma}\right)$$
$$D_i = \Phi\left(\frac{log(\tau_{i_1}) - (\beta_0 + \beta_1 x_{i0})}{\sigma}\right)$$
$$M_i = \Phi\left(\frac{log(t_{L_i}) - (\beta_0 + \beta_1 x_{i1})}{\sigma}\right)$$
$$F_i = \Phi\left(\frac{log(t_{R_i}) - (\beta_0 + \beta_1 x_{i0})}{\sigma}\right)$$
$$U_i = \Phi\left(\frac{log(u_i) - (\beta_0 + \beta_1 x_{i0})}{\sigma}\right)$$
$$V_i = \phi\left(\frac{log(t_{R_i}) - (\beta_0 + \beta_1 x_{i1})}{\sigma}\right)$$
$$Z_i = \phi\left(\frac{log(t_{R_i}) - (\beta_0 + \beta_1 x_{i0})}{\sigma}\right).$$

Also the indicator variables is defined in (10) as follows:

$$\delta I_i = \begin{cases} 1, & \text{if individual's survival times are interval censored} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta R_i = \begin{cases} 1, & \text{if individual's survival times are right-censored} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta E_i = \begin{cases} 1, & \text{if individual's survival times are observed exactly} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta L_i = \begin{cases} 1, & \text{if individual's survival times are left-censored} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta G_i = \begin{cases} 1, & \text{if individual's covariates are updated} \\ 0, & \text{otherwise} \end{cases}$$

$$\delta P_i = \begin{cases} 1, & \text{if covariates are updated before } t_{L_i} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

By combining the likelihood in (7) and (8), the log-likelihood for n independent random samples consisting both prevalence and incidence cohort is given in (10):

$$\begin{aligned} \ell(\boldsymbol{\theta}) = & \sum_{i=1}^{n} \delta I_i \delta G_i (1-\delta P_i) \log[(1-A_i)(1-B_i) \\ & - (1-C_i)(1-D_i)] + \sum_{i=1}^{n} \delta I_i \delta G_i \delta P_i \log[1-D_i] \\ & + \sum_{i=1}^{n} \delta I_i \delta G_i \delta P_i \log[C_i - M_i] \\ & + \sum_{i=1}^{n} \delta I_i \delta G_i (1-\delta P_i) \log[F_i - A_i] \\ & + \sum_{i=1}^{n} \delta R_i \delta G_i \log[1-D_i] + \sum_{i=1}^{n} \delta R_i \delta G_i \log[1-M_i] \\ & + \sum_{i=1}^{n} \delta R_i (1-\delta G_i) \log[1-A_i] + \sum_{i=1}^{n} \delta E_i \delta G_i \log[V_i] \\ & - \sum_{i=1}^{n} \delta E_i \log[t_{R_i}\sigma] + \sum_{i=1}^{n} \delta E_i \delta G_i \log[1-D_i] \\ & + \sum_{i=1}^{n} \delta E_i (1-\delta G_i) \log[Z_i] + \sum_{i=1}^{n} \delta L_i \log[F_i] \\ & - \sum_{i=1}^{n} \delta G_i \log[1-B_i] - \sum_{i=1}^{n} (1-\delta v_i) \log[U_i], \end{aligned} \tag{10}$$

where,

$$\delta v_i = \begin{cases} 0, & \text{if individual's survival times are left-truncated} \\ 1, & \text{otherwise.} \end{cases}$$

5. Simulation study

A simulation study is conducted for $N = 2000$ by generating samples of size $n = 30, 60, 100$ and 200. These samples are generated based on the simulation study discussed in Section 3. The percentage of left-truncated observations were fixed at 20% and 60% for a study period of 60 months with width of inspection times of 4 months ($k = 15$). The bias, standard error (SE) and root mean square error (RMSE) are compared for all the parameter estimates under four different settings; 20% truncation with independent covariates (20pt; inc), 60% truncation with independent covariates (60pt; inc), 20% truncation with dependent covariates (20pt; dpc) and 60% truncation with dependent covariates (60pt; dpc). All simulation is done using R statistical programming software. We used the values of RMSE, $\sqrt{\text{SE}^2 + \text{bias}^2}$ to measure the overall performance of the estimator as it measures the average overall error of the parameter estimates compared to both bias and SE which contribute to the average error size of an estimator.

6. Results and Discussion

Table 1 shows the average percentage of updated covariates, interval censored, right censored, exact lifetimes and left-censored observations generated through the simulation study under the settings of (20pt; inc), (60pt; inc), (20pt; dpc) and (60pt; dpc). By fixing the covariate levels, e.g. compare (20pt; inc) and (60pt; inc), we observe that the percentage of updated covariates are slightly lower at higher percentages of truncation. This may be due to the covariate update time being large for most of the observations under the settings of (60pt; inc) whom may have experience the event of interest prior to the update time. This is also evident with higher percentage of interval censored observations observed when the percentage of truncation is higher compared to when lower percentage of truncation is observed, e.g. compare (60pt; inc) and (20pt;inc) in Table 1.

In contrary, higher percentages of right-censored observations are observed when the percentage of truncation is lower or equally when more observations are recruited from the incidence cohort (new cases), e.g. (20pt;inc). This may be due to the number of inspection times being small for most of these new cases whom may not have experienced the event of interest even until the last inspection times. This subsequently results in the increase of right-censored failure times among the new cases. However, the percentage of exact and left-censored observations are approximately close despite the percentage of truncation.

By fixing the percentage of truncation, e.g. compare (20pt; inc) and (20pt; dpc), the percentage of covariate update times, interval-censored, right-censored, exact observations and left-censored observations are approximately close despite covariate levels, see Table 1.

Table 1: Average percentages of updated covariates (% cov.update), interval censored (IC), right censored (RC), exact observations (EO) and left-censored (LC) observations.

setting	20pt;inc	60pt;inc	20pt;dpc	60pt;dpc
% cov.update	0.8165	0.7771	0.8099	0.7685
% IC	0.7706	0.7741	0.7708	0.7736
% RC	0.0078	0.0061	0.0079	0.0066
% EO	0.2223	0.2198	0.2211	0.2198
%LC	0.0002	0.0001	0.0002	0.0001

Based on Table 2, the absolute value of bias for $\hat{\sigma}$ decreases with the increase in the sample size. However, the trend seems to be unclear for $\hat{\beta}_0$ and $\hat{\beta}_1$. Nevertheless, none of the bias values of these parameter estimates seems to be a concern as these values are insignificant at either 5% or 10% level of significance.

Table 2: Bias, SE and RMSE for parameter estimates with independent and dependent time-varying covariates

parameter		$\hat{\sigma}$			$\hat{\beta}_0$			$\hat{\beta}_1$		
setting	n	bias	SE	rmse	bias	SE	rmse	bias	SE	rmse
20pt;inc	30	-0.0373	0.0666	0.0763	0.0300	0.0901	0.0950	-0.0017	0.0997	0.0997
	60	-0.0265	0.0455	0.0527	0.0310	0.0653	0.0724	-0.0026	0.0644	0.0665
	100	-0.0223	0.0366	0.0429	0.0308	0.0501	0.0588	-0.0013	0.0515	0.0515
	200	-0.0193	0.0248	0.0314	0.0303	0.0346	0.0460	-0.0030	0.0358	0.0359
60pt;inc	30	-0.0350	0.0681	0.0766	0.0344	0.0934	0.0995	-0.0037	0.1000	0.1000
	60	-0.0260	0.0476	0.0543	0.0334	0.0673	0.0752	-0.0033	0.0676	0.0677
	100	-0.0230	0.0373	0.0438	0.0325	0.0501	0.0597	-0.0013	0.0530	0.0531
	200	-0.0192	0.0267	0.0329	0.0324	0.0358	0.0482	-0.0017	0.0373	0.0373
20pt;dpc	30	-0.0307	0.0676	0.0742	0.0285	0.0932	0.0974	-0.0015	0.0966	0.0966
	60	-0.0238	0.0475	0.0531	0.0303	0.0659	0.0726	-0.0082	0.0657	0.0662
	100	-0.0211	0.0368	0.0424	0.0281	0.0507	0.0580	-0.0041	0.0510	0.0511
	200	-0.0189	0.0258	0.0320	0.0299	0.0354	0.0464	-0.0030	0.0347	0.0348
60pt;dpc	30	-0.0358	0.0680	0.0768	0.0378	0.0935	0.1001	-0.0037	0.0968	0.0969
	60	-0.0262	0.0496	0.0561	0.0311	0.0676	0.0744	-0.0041	0.0680	0.0681
	100	-0.0218	0.0372	0.0431	0.0321	0.0513	0.0605	-0.0050	0.0518	0.0520
	200	-0.0176	0.0265	0.0318	0.0324	0.0366	0.0489	-0.0043	0.0362	0.0364

Also, the SE and RMSE decreased with the increase in sample size for all the parameter estimates under all settings, see Table 2. Further, as the percentage of truncation increased, e.g. compare (20pt;inc) and (60pt; inc), the SE and RMSE equally increased. This is as expected, as the increase in the proportion of truncation results in the rise of the number of observations excluded from the left-tail of the log-normal distribution. Thus the sampling bias

and the SE of the parameter estimates increases due to the loss of information from the removed observations. Furthermore, the higher percentage of interval censored observations observed in the presence of higher percentage of truncation subsequently results the likelihood function to rely on the survival function with interval censored times rather than the density function with exact failure times.

Overall, the values of SE and RMSE for all the parameter estimates are lower at lower percentage of truncation e.g. (20pt; inc) and (20pt; dpc) compared to when higher proportion of truncation is present, e.g. (60pt; inc) and (60pt; dpc).

7. Conclusions and Recommendations

In conclusion, the proposed estimator performed well under both covariate levels (dependent or independent) despite the percentage of truncation. Based on the values of RMSE, the proposed estimator is optimum at lower percentage of truncation at both covariate levels. In other words the Newton-Raphson iterative algorithm generated more reliable and accurate estimates under the settings of under the settings of (20pt; inc) and (20pt; dpc) compared to (60pt; inc) and (60pt; dpc). This indicates that the increase in the number of observations from an incidence cohort (new cases) will further improve the performance of the proposed estimator.

The results following the simulation study is equally applicable with the parameters of the log-logistic distributions as this distribution shares similar hazard rate properties with the log-normal distribution. Nonetheless, the simulation methodology proposed in this study can be applied to any parametric distributions by specifying the correct survival and hazard function through the general survival function discussed in section 2.

8. Acknowledgements

We would like to extend our gratitude to Fundamental Research Grant Scheme (FRGS), VOT 5524226, University Putra Malaysia and Dr.Patricia Tai, University of Saskatchewan, Saskatoon, Canada.

References

Arasan, J. and Lunn, M. (2008). Alternative interval estimation for parameters of bivariate exponential model with time varying covariate. *Computational Statistics*, 23(4):605–622.

Balakrishnan, N. and Mitra, D. (2014). Some further issues concerning likelihood inference for left truncated and right censored lognormal data. *Communications in Statistics - Simulation and Computation*, 43(2):400–416.

Collett, D. (2003). *Modelling Survical Data in Medical Research*, volume 57. CRC press.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, volume 21. CRC Press.

Guo, G. (1992). Event-history analysis for left-truncated data. *Sociological methodology*, 23:217–243.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

Kiani, K. and Arasan, J. (2012). Gompertz model with time-dependent covariate in the presence of interval-, right-and left-censored data. *Journal of Statistical Computation and Simulation*, (ahead-of-print):1–19.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. Springer.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Nardi, A. and Schemper, M. (2003). Comparing cox and parametric models in clinical studies. *Statistics in medicine*, 22(23):3597–3610.

Petersen, T. (1986). Fitting parametric survival models with time-dependent covariates. *Applied Statistics*, pages 281–288.

Tai, P., Chapman, J.-A. W., Yu, E., Jones, D., Yu, C., Yuan, F., and Sang-Joon, L. (2007). Disease-specific survival for limited-stage small-cell lung cancer affected by statistical method of assessment. *BMC cancer*, 7(1):31–39.