GAVIN PUBLISHERS

OPEN ACCESS

Journal of Biostatistics & Biometrics

Research Article

Liu Y and Li H. J Biostat Biom: JBSB-109.
DOI: 10.29011/JBSB-109.100009

Semiparametric Likelihood Estimation with Clayton-Oakes Model for Multivariate Current Status Data

Yeqian Liu* **and Hanyi Li**

Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA

***Corresponding Author:** Yeqian Liu, Assistant Professor, Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA

Citation: Liu Y, Li H (2020) Semiparametric Likelihood Estimation with Clayton-Oakes Model for Multivariate Current Status Data. J Biostat Biom: JBSB-109. DOI: 10.29011/JBSB-109.100009

Received Date: 16 April, 2020; **Accepted Date:** 24 April, 2020; **Published Date:** 30 April, 2020

Abstract

In biomedical and public health studies, current status data arises when each subject is observed only once and the failure time is only known to be either smaller or larger than the observation time. This paper discusses regression analysis of multivariate current status failure time data which usually arise in epidemiological studies and animal carcinogenicity experiments. We propose an EM-algorithm to estimate the regression coefficients under Clayton-Oakes model for multivariate current status data. The Clayton-Oakes model is a very promising model for analysis of multivariate failure time data, it characterize the dependence of multiple failure time by using a gamma frailty. The simulation study indicated that the method works well for practical situations. An application from a tumorigenicity experiment is provided.

Keywords: Multivariate current status data; Clayton-Oaks model; Frailty model; EM algorithm; Maximum likelihood estimation

Introduction

Current status data arise when each subject in a survival study is observed only once and the failure time of interest is known only to be smaller or larger than the observation time. In other words, the failure time of interest is either left- or right-censored instead of being observed exactly. Current status data occur quite often in many applications, including tumorigenicity experiments and epidemiological investigations of the natural history of a disease.

Current status data is a special case of interval-censored data and is often referred to as case 1 interval-censored data [1]. Interval-censored failure time data occur when the failure time of interest is only observed to be bracketed by two adjacent examination times [2]. The interval-censored data reduce to current status data if all intervals include either zero of infinity. In this paper, we will deal with multivariate current status data, arising when there are multiple failure times of interest in a survival study and only current status data are available for each failure time.

We will focus on the regression analysis of multivariate current status data. In order to model the covariate effects on the failure times of interest as well as accounting for the potential dependence of the failure times, we propose to use the Clayton-Oakes (CO) model jointly and the proportional hazards model marginally. The CO model is one of parametric copula models which express the joint distribution of failure times as a function of the marginal distributions and a dependence parameter [3]. The proportional hazards model is one of the most commonly used semiparametric models in survival analysis. In this paper, we will consider the estimation of the CO model with marginal proportional hazards model based on multivariate current status data.

Glidden DV, et al. (1999) [4], considered the estimation of the CO model based on multivariate right-censored failure time data and they developed an approximate EM algorithm for the determination of maximum likelihood estimates. Unlike the right-censored data, the failure time of interest is never observed exactly under current status data. Therefore, the well-developed tools used for right-censored data cannot be readily generalized to fitting current status data. For example, the partial likelihood approach and the elegant martingale theory used in the estimation of proportional hazards model under right-censored data do not carry over to the case of current status data [5]. Thus, the regression analysis of current status data is not an easy task. Here we are trying to develop an EM algorithm for the estimation of a semiparametric regression model with multivariate current status data. This is even more challenging that the univariate case.

Sun T, et al. (2019) [3], pointed out that a reparametrization of the CO model is equivalent to one of gamma frailty models discussed in *Wang N, et al. (2015)* [6]. Therefore, after a reparametrization, we are essentially dealing with a gamma frailty model but the marginal hazard functions would become much more complicated that proportional hazards functions [6]. The CO model and the equivalent gamma frailty model will be given in section 2. The frailty model approach has been commonly used in the analysis of multivariate failure time data. This approach assumes that there exists a common but unobserved latent variable, called frailty that characterizes the dependence of the possibly correlated failure times. *Chen MH, et al. (2009*) [7], considered a frailty model approach for regression analysis of multivariate current status data.

The rest of this paper is organized as follows. In section 2, we present models, assumptions and the likelihood function. In section 3, we develop a sieve maximum likelihood approach for the estimation of the unknown parameters. The sieves are constructed based on piecewise linear functions. By using the sieve method, we avoid the estimation of infinite dimensional parameters. To determine the sieve maximum likelihood estimate, we propose an EM algorithm by treating the unobservable latent variable as missing values. In section 4, we conduct simulation studies to examine the finite sample performance of our proposed method. Section 5 applies the approach to NCTR's tumorigenicity experiment. Section 6 contains some discussion and concluding remarks.

Assumptions, Models and the Likelihood Function

Suppose we are interested in K failure times in a survival study, denoted by $T_1, \ldots, T_K$. Let Z_k be the covariates that may affect T_k, k=1, …, k, respectively. We assume the proportional hazards model for each failure time, i.e. the marginal hazard function for T_k is given by

$$\lambda_k(t) = \lambda_{0k}(t)\, exp\,(\beta' z_k),\ k = 1, \ldots, K,$$

Where $\lambda_{0k}(t)$ is the baseline hazard function for T_k and β is the vector of regression parameters. Note that we assume a common vector of covariate effects. Also, we assume that the joint distribution of the failure times follows the Clayton-Oakes (CO) model, i.e., the joint survival function is given by

$$S(t_1, \ldots, t_K; \alpha) = \left[\sum_{k=1}^{K} \exp\,(\alpha \Lambda_k(t_k)) - K + 1\right]^{-\alpha^{-1}},\ \alpha \in (0, \infty),$$

Where the parameter α specifies the CO model and $\Lambda_k(t) = \int_0^t \lambda_k(u)du$. When α = 0, the joint survival function becomes the product of the marginal survival functions and failures are independent. Increasing values of $\alpha\alpha$imply increasing positive association between failures [8].

The CO model is equivalent to the following gamma frailty model [4]. Suppose ξ is a gamma random variable with mean 1 and unknown variance α. Conditional on the frailty ξ, it is assumed that the failure times are independent with hazard functions:

$$\xi \lambda_{0k}(t)\, exp\,(\beta' z_k + \alpha \Lambda_{0k}(t)\, exp\,(\beta' z_k)),\ k = 1, \ldots, K, \quad (2.1)$$

Where $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(u)du$.

Suppose that only current status data are available for each T_k, and the observed data for a single subject are $O^* = \{c_k, z_k, \delta_k = I(T_k \le c_k \wedge \tau),\ k = 1, \ldots, K\}$, where c_k denotes the observation time for kth failure and τ denotes the length of study. Without loss of generality, we will assume that the $c_k's$ are the same, and the common observation time is denoted by c. Also, we assume that c is independent of T_k given Z_k and the distribution of c does not depend on the unknown parameters $(\beta, \alpha, \Lambda_{0k}(\cdot))$, i.e., the censoring mechanism is noninformative. Then, given ξ, the likelihood of observing δ_k is proportional to

$$[1 - exp(-\xi\, exp\,(\alpha \Lambda_{0k}(t)\, exp\,(\beta' z_k))/\alpha)]^{\delta_k}$$

$$exp(-(1-\delta_k)\xi\, exp\,(\alpha \Lambda_{0k}(t)\, exp\,(\beta' z_k))/\alpha).$$

This likelihood involves infinite-dimensional parameters $\Lambda_{0k}(t)$ which is difficult to deal with. We propose a sieve approach to estimate $\Lambda_{0k}(t)$, that is, to approximate it by a piecewise linear function

$$\Lambda_{0k}(t) = \sum_{j=1}^{J} I_j(t) \sum_{a=1}^{j} e^{\gamma_{ka}},$$

Where $0 = s_0 < s_1 < \cdots < s_J = \tau$ are constants, $I_j = (s_{j-1}, s_j]$, and $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kJ})'$ are unknown parameters, j = 1, … ,J, k = 1, … ,K.

Define $\gamma = (\gamma_1', \ldots, \gamma_K')$ and denote all unknown parameters as θ = (β,α,γ), then the likelihood function for a single subject is given by

$$L^*(\theta; O^*) = \int \prod_{k=1}^{K} \sum_{j=1}^{J} (I_j(c)[1 - \exp\,(-\xi \exp\,(\alpha \sum_{a=1}^{j} e^{\gamma_{ka}} \exp\,(\beta' z_k))/\alpha)]^{\delta_k} \exp\,(-(1-\delta_k)\xi \exp\,(\alpha \sum_{a=1}^{j} e^{\gamma_{ka}} \exp\,(\beta' z_k))/\alpha)) f(\xi; \alpha) d\xi, \quad (2.2)$$

Where f(ξ;α) is the density function of the gamma distribution with mean 1 and variance α.

From (2.2), it is easy to show that the conditional density function of ξ given Ω* has the form

$$f(\xi|O^*,\theta)=\frac{1}{L^*(\theta;O^*)}\prod_{k=1}^{K}\sum_{j=1}^{j}(I_j(c)[1-\exp(-\xi\exp(\alpha\sum_{a=1}^{j}e^{\gamma_{ka}}\exp(\beta' z_k))/\alpha)]^{\delta_k}$$
$$\exp(-(1-\delta_k)\xi\exp(\alpha\sum_{a=1}^{j}e^{\gamma_{ka}}\exp(\beta' z_k))/\alpha))f(\xi;\alpha) \quad (2.3)$$

We will develop an EM algorithm to determine the sieve maximum likelihood estimate of θ in the next section.

Maximum Likelihood Estimation

Let the observed data be $O=\{O_i=\{\delta_{ik},z_{ik},c_i,\ k=1,\dots,K\},\ i=1,\dots,n\}$, which are n i.i.d. replicates of O*. Then the full likelihood function is given by

$$L(\theta;O)=\prod_{i=1}^{n}L^*(\theta;O_i).$$

Note that for each individual i, there is an unobservable latent variable ξ_i which has the gamma distribution with mean 1 and variance α > 0 with the density function

$$f(\xi;\alpha)=\frac{1}{\alpha^{1/\alpha}\Gamma(\frac{1}{\alpha})}\xi^{frac1\alpha-1}e^{-\xi/\alpha}$$

We want to obtain the value of θ that maximizes L(θ;0), i.e. the maximum likelihood estimate $\hat{\theta}$. Since there is no closed form solution for the maximum likelihood estimate, we try to determine $\hat{\theta}$ using EM algorithm.In order to develop the EM algorithm, we consider the latent variables $\xi=\{\xi_i,\ i=1,\dots,n\}$ to be unobserved values and develop the E-step and M-step iteratively for the EM-algorithm.

In the E-step, we first derive the pseudo-complete data likelihood function. There are two components for the pseudo-complete data: the observed data O and the missing data ξ. The E-step computes the conditional expectation of the log-likelihood of the pseudo-complete data conditioning on ξ given the observed data O. It is very straightforward to write the log-likelihood function as

$$l(\theta;O,\xi)=\sum_{i=1}^{n}l_i(\theta;O_i,\xi_i),$$

Where

$$l_i(\theta;O_i,\xi_i)=\log f(\xi_i;\alpha)+\sum_{k=1}^{K}\sum_{j=1}^{J}I_j(c_i)[\delta_{ik}\log(1-\exp(-\xi_i\exp(\alpha\sum_{a=1}^{j}e^{\gamma_{ka}}\exp(\beta' z_{ik}))/\alpha))$$
$$-(1-\delta_{ik})\xi_i\exp(\alpha\sum_{a=1}^{j}e^{\gamma_{ka}}\exp(\beta' z_{ik}))/\alpha].$$

Then the conditional expectation of the log-likelihood has the form

$$l(\theta;O)=\sum_{i=1}^{n}E[l_i(\theta;O_i,\xi_i)]=\sum_{i=1}^{n}\int l_i(\theta;O_i,\xi_i)f(\xi_i|O_i,\theta)d\xi_i$$

With θ set to be $\theta^{(m)}$ obtained at the mmth iteration. Since the expectation has no closed form, we employ some numerical integral algorithms to compute such expectation

$$E[h(\xi_i)|O_i,\theta^{(m)}] = \int h(\xi_i) f(\xi_i|O_i,\theta^{(m)}) d\xi_i$$

for any function h(ξ_i) of ξ_i.

In order to determine $E(h(\xi_i)|O,\theta^{(m)})$, we can rewrite it as

$$E(h(\xi_i)|O_i,\theta^{(m)}) = \frac{E[\psi(\xi_i;\theta^{(m)},O_i)h(\xi_i)]}{E\psi(\xi_i;\theta^{(m)},O_i)}$$

Where the expectations on the right-hand side are taken with respect to the gamma distribution with mean 1 and variance $\alpha^{(m)}$, $\alpha^{(m)}$ denotes the estimate of α obtained at the mth iteration, and

$$\psi(\xi_i;\theta^{(m)},O_i) = \prod_{k=1}^{K}\sum_{j=1}^{J}(I_j(c_i)[1-\exp(-\xi_i\exp(\alpha\sum_{a=1}^{j}e^{\gamma_{ka}}\exp(\beta' z_{ik}))/\alpha)]^{\delta_{ik}}$$

$$\exp(-(1-\delta_{ik})\xi\exp(\alpha\sum_{a=1}^{j}e^{\gamma_{ka}}\exp(\beta' z_{ik}))/\alpha))$$

This suggests that for sufficiently large L, the expectation $E(h(\xi_i)|O,\theta^{(m)})$ can be approximated by

$$E(h(\xi_i)|O_i,\theta^{(m)}) \simeq \hat{E}(h(\xi_i)) = \frac{\sum_{l=1}^{L}\psi(\xi_l;\theta^{(m)},O_i)h(\xi_l)}{\sum_{l=1}^{L}\psi(\xi_l;\theta^{(m)},O_i)}, \quad (3.1)$$

Where $\{\xi_l, l = 1, \ldots, L\}$ are i.i.d. samples from the gamma distribution with mean 1 and variance $\alpha^{(m)}$.

Now, we will describe the M-step, which maximizes the conditional expectation l(θ;0) with replacing all expectations involving functions h(ξ_i) by their approximation $\hat{E}(h(\xi_i))$ given in (3.1) to determine the updated estimate $\theta^{(m+1)}$. We adopt the moment estimate for α in each iteration, that is, the estimate of α at the (m+1)th iteration is given by

$$\alpha^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{E}(\xi_i\xi_i') - 1. \quad (3.2)$$

For the maximum likelihood estimator of parameters β and γ, we have

$$U_\beta(\beta,\gamma) = \frac{\partial l(\theta;O,\xi)}{\partial\beta} = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{a=1}^{j}\exp(\gamma_{ka})I_j(c_i)z_{ik}E\psi_{ikj}^{(1)}(\xi_i;\beta,\gamma_k)$$

$$U_{\gamma_{kj}}(\beta,\gamma) = \frac{\partial l(\theta;O,\xi)}{\partial\gamma_{kj}} = \sum_{i=1}^{n}\sum_{h=j}^{J}\exp(\gamma_{kj})I_h(c_i)E\psi_{ikh}^{(1)}(\xi_i;\beta,\gamma_k)$$

Where

$$\psi_{ikj}^{(1)}(\xi_i;\beta,\gamma_k) = [\frac{\delta_{ik}}{1-\exp[-\xi_i\exp[\alpha e^{\beta z_{ik}}\sum_{a=1}^{j}e^{\gamma_{ka}}+\beta z_{ik}]/\alpha]} - 1]$$

$$\xi_i\exp[\alpha e^{\beta z_{ik}}\sum_{a=1}^{j}e^{\gamma_{ka}} + \beta z_{ik}]$$

Applying the approximation of $\hat{E}$ given in (3.1), we can obtain the working score functions as

$$\hat{U}_\beta(\beta,\gamma) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{a=1}^{j}\exp(\gamma_{ka})I_j(c_i)z_{ik}\frac{\sum_{l=1}^{L}\psi(\xi_l;\theta^{(m)},O_i)\psi_{ikj}^{(1)}(\xi_l)}{\sum_{l=1}^{L}\psi(\xi_l;\theta^{(m)},O_i)}$$

$$\hat{U}_{\gamma_{kj}}(\beta,\gamma) = \frac{\partial l(\theta;O,\xi)}{\partial\gamma_{kj}} = \sum_{i=1}^{n}\sum_{h=j}^{J}\exp(\gamma_{kj})I_h(c_i)\frac{\sum_{l=1}^{L}\psi(\xi_l;\theta^{(m)},O_i)\psi_{ikh}^{(1)}(\xi_l)}{\sum_{l=1}^{L}\psi(\xi_l;\theta^{(m)},O_i)}$$

Where $\{\xi_l, l=1,\ldots,L\}$ are i.i.d. samples from the gamma distribution with mean 1 and variance $\alpha^{(m)}\alpha^{(m)}$. Then for estimation of β and γ, we can solve the equation

$$\hat{U}(\beta,\gamma) = (\hat{U}_\beta(\beta,\gamma)', \hat{U}_{\gamma_1}(\beta,\gamma_1)', \ldots, \hat{U}_{\gamma_K}(\beta,\gamma_K)')' = 0 \quad (3.3)$$

Where $\hat{U}_{\gamma_k}(\beta,\gamma) = (\hat{U}_{\gamma_{k1}}(\beta,\gamma),\ldots,\hat{U}_{\gamma_{kJ}}(\beta,\gamma))'$.

It is obvious that it would not be easy to solve all p + K * J equations in (3.3) above simultaneously. For this, we suggest the following procedure for the (m+1)th iteration.

Step 1: Determine the updated estimate $\hat{\alpha}^{(m+1)}$ of α using (3.2).

Step 2: Determine the updated estimate $\hat{\beta}^{(m+1)}$ of β by solving $\hat{U}_\beta(\beta,\gamma^{(m)}) = 0$.

Step 3: Determine the updated estimate $\hat{\gamma}_k^{(m+1)}$ of γ_k by solving the equation $\hat{U}_{\gamma_k}(\hat{\beta}^{(m+1)},\gamma_k)$, for each k = 1, ... ,K.

Step 4: Repeat steps 1-3 until convergence.

To guarantee that α is positive at each iteration, we reparameterize the frailty variance as $\alpha = \exp(\alpha^*)$ in the procedure. With a little abuse of notations, we still denote the parameter as α. The key advantage of the procedureis that, by using sieve method, one can avoid solving high dimensional or large number of equations and thus can easily apply well developed optimization algorithms in the M-step, such as Newton-Raphson algorithm. Moreover, the computation is more stable and efficient than that using nonparametric estimation of the cumulative baseline hazards functions.

Note that the number of partitions J should increase as the sample size n increases. In general, larger J leads to a better estimation, and is more computational intensive. A common practice is to choose J to be an integer greater than $n^{1/4}$ and the

I_j's such that each interval contains roughly equal numbers of observation time points [7]. In practice, it is recommended to try different choices for J and the I_j's and find out the case that works the best.

A Simulation Study

This section carries out extensive simulation studies to examine the performance of the proposed method. Our model specification mostly follows that used by *Glidden DV, et al. (1999)* [4], where a Clayton-Oakes (CO) with gamma frailty was studied for multivariate survival times subject to right censoring. In this study, we consider the situation where k = 2, so there exists two failure times T_1 and T_2. Let the covariate z be one-dimensional and generated from a binomial distribution Bin (1,0.5). We assume the cumulative baseline hazard function for T_1 and T_2 to be $\Lambda_1(t) = 0.2t$ and $\Lambda_2(t) = 0.05t$, respectively. Then the conditional hazard functions for T_1 and T_2 given ξ are

$$0.2\xi \exp(\beta' z_1 + 0.2\alpha t \exp(\beta' z_1))$$

and

$$0.05\xi \exp(\beta' z_2 + 0.05\alpha t \exp(\beta' z_2)),$$

respectively. Here, ξ is the gamma frailty which is assumed to have a gamma distribution with mean 1 and variance α. The failure times T_1 and T_2 can be generated from the inverse function of the conditional cumulative distribution function given the generated ξ. In the study, we took $s_j = j/J \times \tau$ with J = 6, and the observation times c_j's are generated from the discrete uniform distribution over $\{s_1, s_2, \ldots, s_j\}$. The results given below are based on 300 replications with L=1000 for the approximation and the sample size n=200 or 400.

To obtain the estimates $\hat{\alpha}$ and $\hat{\beta}$, we consider four different cases. In the tables 1 and 2, we include the Averages (AVE) and the Sample Standard Deviations (SSE). The true value of β was taken as 0.5 or -0.5. And the frailty variance α was set to be 1.2 or 0.8. One can see from these tables that the proposed estimate $\hat{\beta}$ and $\hat{\alpha}$ seems to relatively unbiased.

	Parameters	n	J	True	AVE	BIAS	SSE	MSE
Case 1	α	200	6	0.8	0.877	0.077	0.126	0.132
	β			0.5	0.473	-0.027	0.163	0.189
Case 2	α	200	6	0.8	0.749	0.051	0.112	0.115
	β			-0.5	-0.554	-0.054	0.177	0.181
Case 3	α	200	6	1.2	1.281	0.081	0.147	0.154
	β			0.5	0.435	0.065	0.181	0.185
Case 4	α	200	6	1.2	1.131	-0.069	0.153	0.157
	β			-0.5	-0.578	-0.078	0.172	0.179

Table 1: Results on estimations of the regression coefficients with n=200.

	Parameters	n	J	True	AVE	BIAS	SSE	MSE
Case 1	α	400	6	0.8	0.826	0.026	0.086	0.093
	β			0.5	0.485	-0.015	0.116	0.129
Case 2	α	400	6	0.8	0.773	0.027	0.082	0.086
	β			-0.5	-0.514	-0.014	0.137	0.125
Case 3	α	400	6	1.2	1.245	0.045	0.116	0.107
	β			0.5	0.471	0.029	0.135	0.128
Case 4	α	400	6	1.2	1.161	-0.039	0.108	0.119
	β			-0.5	-0.593	-0.007	0.126	0.131

Table 2: Results on estimations of the regression coefficients with n=400.

As mentioned in [7], one could estimate β based on the observed current status data on T_1 and T_2 by assuming that β in the model is the same for T_1 and T_2. However, this marginal approach is less efficient than the estimation procedure developed here. We also found the EM-algorithm developed is sensitive to the initial value of γ_k. To obtain good convergence rate and increase convergence speed, it is important and necessary to have good initial values.

An Application

Now we apply the proposed method to an animal tumorigenicity experiment conducted by the National Center for Toxicological Research (NCTR). It is a 2-year rodent carcinogenicity study of chloroprene consisting of F344/N rats and B6C3F1 rats with both sexes. There are 50 male and 50 female rats in both control and high dose group. Rats in the high dose group were exposed to chloroprene at the concentrations of 80 ppm, 6 h per day, 5 days per week for up to 2 years. The animal either died during the study or was sacrificed at the end of the study. At the death or sacrifice, the presence or occurrence of different types of tumors was determined through a pathological examination. Due to the nature of the pathological examination process, the tumor occurrence time was not exactly observed but instead known only to be smaller or greater than the death or sacrifice time. Therefore, the tumor occurrence time of lung and adrenal cancers is bivariate current status data. Following *Dunson DB, et al. (2002)* [9], the main objective of the study is to investigate the effect of choloroprene on lung and adrenal cancer by comparing the tumor growth rates between the control and high dose group.

Let T_1 and T_2 denote the occurrence times of adrenal and lung tumors, we model the joint distribution of T_1 and T_2 using the proposed Clayton-Oaks model with Gamma frailty. Application of the proposed EM-algorithm gave $\hat{\beta}\hat{\beta}$= 0.5717 with an estimated standard error of 0.2851. Therefore, the animals in the high dose group had significantly higher occurrence rate of both adrenal and lung tumors than those in the control group. To illustrate this occurrence rate difference, Figure 1 shows the estimated survivals functions of the lung tumor occurrence time for animals in the control and high dose groups.

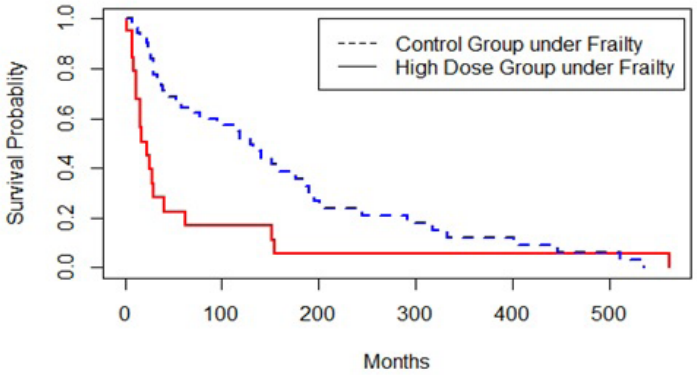

Figure 1: Estimated marginal survival functions for time to lung tumor.

To compare the performance of the proposed Clayton-Oaks model with traditional marginal proportional hazards model. The Mean Square Error (MSE) was calculated for both models. For the control group of the lung tumor, we obtained MSE=0.8741 under Clayton-Oaks model and 1.1681 under the marginal proportional hazards model. For the high dose group of the lung tumor, the corresponding values are 0.8254 and 1.3517, respectively. These results suggest that the proposed Clayton-Oaks model fits the data better than the marginal proportional hazard model for the lung tumor. The results for the adrenal tumor are similar.

Discussion and Concluding Remarks

In this paper, we proposed an EM-algorithm to estimate the regression coefficients under Clayton-Oakes model for multivariate current status data. The simulation study and real data application indicated that the method works well for practical situations. The Clayton-Oakes model is a very promising model for analysis of multivariate failure time data, it characterize the dependence of multiple failure time by using a gamma frailty. This model is different from the commonly used marginal Cox model and normal frailty Cox model. The analysis of CO model would benefit from methods which can diagnose the appropriateness of the model.

Actually, the gamma frailty assumption is not essential, and the proposed procedure can be carried out for other choices of frailty distributions, for example, log-normal frailty, positive stable frailty, or inverse Gaussian frailty. It is noteworthy that the proposed frailty model does not provide the marginal model directly, but the marginal model can be derived by integrating out the frailty.

The EM-algorithm developed in this paper can also be extended to case II interval censored data. However, the presented methodology has several limitations, which lead to some directions for future research. The first limitation is that one may have a hard time conducting model selection between our proposed CO model and the log-normal frailty Cox model proposed by *Chen MH, et al. (2009)* [7]. Moreover, in our model, we assume that for in each cluster i, the failure times $\{T_1,\ldots,T_k\}$ are independent. In practice, this may not be true. However, to consider the dependence between k failure times, we have to introduce a multivariate gamma frailty which will significantly increase the computational difficulty.

References

1. Sun J (2006) The statistical analysis of interval-censored data. Springer.
2. Zeng D, Gao F, Lin D (2017) Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. Biometrika 104: 505-525.
3. Sun T, Ding Y (2019) Copula-based semiparametric regression method for bivariate data under general interval censoring. Biostatistics.
4. Glidden DV, Self SG (1999) Semiparametric likelihood estimation in the Clayton-Oakes failure time model. Scand. J Statist 26: 363-372.
5. Wen CC and Chen YH (2013) A frailty model approach for regression analysis of bivariate interval-censored survival data. Statistica Sinica 23: 383-408.
6. Wang N, Wang L, McMahan CS (2015) Regression analysis of bivariate current status data under the gamma-frailty proportional hazards model using the EM algorithm. Computational Statistics & Data Analysis 83: 140-150.
7. Chen MH, Tong X, Sun J (2009) A frailty model approach for regression analysis of multivariate current status data. Statistics in Medicine 2: 3424-3436.
8. Nielsen GG, Gill RD, Andersen PK, Sorensen TIA (1992) A counting process approach to maximum likelihood estimation in frailty models. Scand J Statist 19: 25-43.
9. Dunson DB, Dinse GE (2002) Bayesian models for multivariate current status data with informative censoring. Biometrics 58: 79-88.