Frailty models with a cure fraction for modeling clustered discrete survival data

Yunchan Chi
Department of Statistics, National Cheng-Kung University
No.1 University Road,
Tainan, Taiwan
E-mail: ycchi@stat.ncku.edu.tw

Chia-Min Chen
Department of Statistics, National Cheng-Kung University
No.1 University Road,
Tainan, Taiwan
E-mail: cmchen@stat.ncku.edu.tw

1. Introduction

Clustered survival data with a cure fraction arise naturally from biomedicine, econometrics and sociology studies. For example, the one-stage non-submerged dental implant study conducted by Chi-Mei Medical Center in Taiwan (Wen *et al.*, 2008) is to identify the risk factors associated with dental implant failure based on a 7-year follow-up data set composed of 324 patients (742 Implants). Here, a patient is referred to as a cluster and cluster sizes vary from 1 to 11. Implant failure was defined as if there were functional signs of pain or discomfort, inflammation or infection during the clinical examination, implant mobility, radiolucency or radiographically detectable bone loss recurrent. Since Wen *et al.* (2008) applied the life-table method to estimate the failure rate, the failure times are only available in yearly intervals. Therefore, the time until failure can be considered as grouped or discrete survival time. Moreover, cure is possible for dental implant. Therefore, it is desirable to use the mixture cure rate models to identify the risk factors associated with dental implant failure based on discrete survival time data.

The mixture cure rate models have been well developed for univariate and multivariate (or clustered) continuous right-censored data. In the literature, two approaches have been proposed for multivariate or clustered continuous right-censored data. One is referred to as marginal regression approach and this approach is useful if a covariate’s population average effect is of primary interest and the correlation structure is not of interest. For example, Peng et al. (2007) extended univariate mixture cure rate models for multivariate continuous survival data by modeling the marginal distribution as a proportional hazards model with logistic regression for cure fraction. Another is long-term survivor mixture model with frailty or random effect to take into account for the correlation structure within each cluster. For example, Chatterjee and Shih (2001) and Wienke et al. (2003) considered the shared frailty model and correlated frailty model, respectively, for bivariate continuous survival data.

Recently, Zhao and Zhao (2008) proposed discrete-time survival models with long-term survivors for univariate grouped or discrete-time survival data. Chi, Chen and Su (2010) proposed using a marginal regression approach to estimate the regression parameter in discrete-time survival models for clustered discrete survival data. When the joint survival distribution is obtained from a frailty model, the random effect frailty model is more appropriate to capture the association between the correlated survival times. Therefore, frailty is introduced in this paper to characterize the correlation between discrete survival times within clusters. In particular, the positive association between survival times is incorporated by imposing a common gamma frailty effect for clustered discrete survival data. The accuracy of the estimators of the parameters in the mixture cure gamma frailty model is examined by simulation. In addition, the implementation of the mixture cure gamma frailty model to a dental implant study is presented.

2. Mixture cure frailty model

Let X_{jk} and U_{jk} be the survival time and censoring time for the *k*th individual in the *j*th cluster, $j=1,2,...,N$ and $k=1,2,...,n_j$. The observed survival time is denoted by $T_{jk}=\min\{X_{jk},U_{jk}\}$ with a right-censored indicator $\delta_{jk}=I(X_{jk}<U_{jk})$. Let Z_{jk} denote a q by 1 vector of covariates for the *k*th individual in the *j*th cluster. Next, the m distinct observed survival times are denoted by t_i with $t_1<t_2<\cdots<t_m$.

For j = 1, to account for heterogeneity, Price and Manatunga (2001) considered the use of frailty mixture models $S(t)=(1-p)+pE(\exp(-H_0(t)w)$, which is proposed by Longini and Halloran (1996), where $H_0(t)$ is the cumulative baseline hazard function and w is the frailty. They used different frailty distributions to model the leukemia remission data. Chatterjee and Shih (2001) extended the univariate mixture cure rate models to bivariate continuous survival times, which follow a shared gamma frailty model. Based on this model, they constructed a full likelihood function for bivariate continuous survival times to obtain parameter estimate. Since the shared gamma frailty model only explains correlations within clusters, Wienke et al. (2003) used correlated frailty model to account for both correlations within cluster and population heterogeneity. For bivariate continuous survival times, the maximum likelihood estimators can be obtained from the full likelihood function. Their procedures are fully parametric approach and do not consider covariates.

To account for the correlation, the mixture cure frailty model for the *k*th individual in the *j*th cluster, based on the given covariates Z_{jk} and a frailty u_j is proposed in this paper, as

$$S(t\mid Z_{jk},u_j)=1-p+p[S^*(t\mid Z_{jk})]^{u_j}\ .$$

Note that the individuals within the same cluster share the same frailty. If $S^*(t\mid Z_{jk})$ follows a proportional hazards model, the conditional improper survival function given frailty u_j can be expressed as

$$S(t \mid Z_{jk}, u_j) = 1 - p + p\exp[u_j \exp(\beta Z_{jk}) \ln S_0^*(t)], \qquad (2.1)$$

where β is a q by 1 vector of unknown regression parameters and $S_0^*(t)$ is an unknown baseline survival function. Thus a semiparametric frailty model is considered here. Note that, for discrete survival times, the baseline survival function can be expressed as a product of hazard rates, that is $S_0^*(t) = \prod_{i|t_i \le t}(1-\lambda_{i0}^*)$, where λ_{i0}^* is the baseline hazard rate evaluated at time t_i.

To estimate the unknown parameters *p*, *β*, and λ_{i0}^*, the conditional likelihood function for discrete survival times can be constructed as follows. For the *k*th individual in the *j*th cluster, if δ_{jk} = 1, then the contribution to the conditional likelihood function given the frailty u_j is

$$f(t \mid Z_{jk}, u_j) = p\exp[u_j \exp(\beta Z_{jk}) \ln S_0^*(t-1)] - p\exp[u_j \exp(\beta Z_{jk}) \ln S_0^*(t)].$$

Whereas, if δ_{jk} = 0, then the contribution is

$$S(t \mid Z_{jk}, u_j) = 1 - p + p\exp[u_j \exp(\beta Z_{jk}) \ln S_0^*(t)].$$

Thus, the likelihood function conditional on the frailty u_j is

$$L_{jk}(\theta \mid u_j) = \left[f(t \mid Z_{jk}, u_j)\right]^{\delta_{jk}} \left[S(t \mid Z_{jk}, u_j)\right]^{1-\delta_{jk}},$$

where $\theta = (\beta, p, \lambda_{10}^*, \ldots, \lambda_{m0}^*)'$ is the vector of all parameters in the mixture cure frailty model. Because the survival times within the same cluster conditional on frailties are assumed to be independent, the likelihood function for the *j*th cluster conditional on the frailty u_j is $L_j(\theta \mid u_j) = \prod_{k=1}^{n_j} L_{jk}(\theta \mid u_j)$. Then, the unconditional likelihood function can be derived when the distribution of frailty is specified. For example, if u_j follows gamma distribution, the likelihood function conditional on the frailty u_j is

$$L_j(\theta) = \int \prod_{k=1}^{n_j} L_{jk}(\theta \mid u_j) \frac{\sigma^\sigma}{\Gamma(\sigma)} u_j^{\sigma-1} e^{-\sigma u_j} du_j,$$

which does not include the unobserved information u_j. In general, the unconditional likelihood function can be constructed by

$$L(\theta) = \prod_{j=1}^{N} L_j(\theta) = \prod_{j=1}^{N} \int \prod_{k=1}^{n_j} L_{jk}(\theta \mid u_j) h(u_j) du_j,$$

where $h(u_j)$ is the density function of u_j. Then the maximum likelihood estimator of θ can be derived through Newton-Raphson algorithm when the distribution of the frailty is specified.

3. Simulation study

To assess the performance of the maximum likelihood estimators in the mixed cure frailty model, a simulation study is conducted. To generate clustered survival data from the mixed cure frailty model displayed in (2.1), the covariate of each subject in the cluster is generated first. A binary covariate Z_{jk} for the *k*th individual in the *j*th cluster is considered and generated from Bernoulli distribution with

$P(Z_{jk}=1)=0.5$. Similarly, the cure status C_{jk} for the *k*th individual in the *j*th cluster is generated from Bernoulli distribution with cured probability of 0.1, that is *p* = 0.9. Then the gamma frailty u_j controls within-cluster dependence of the *j*th cluster is generated from a one-parameter gamma distribution $f(u_j)=\sigma^{\sigma}u_j^{\sigma-1}e^{-\sigma u_j}/\Gamma(\sigma)$ with $\sigma=2$. Therefore, the expectation of u_j is 1 and the variance of u_j is 2. Note that a smaller value of σ induces a more strong correlation between survival times within the cluster. Next, the baseline survival function for discrete survival times employed here is the same as in Zhao and Zhou (2008). The baseline hazard rates at each time point are specified as λ_{10}^*=0.200, λ_{20}^*=0.375, λ_{30}^*=0.300, λ_{40}^*=0.714, and λ_{50}^*=1 at t_1=1, t_2=2, t_3=3, t_4=4, and t_5=5, respectively.

For C_{jk} = 0, the survival time is generated from the frailty cure model, $S_{jk}(t)=[S_0^*(t)]^{u_j\exp(\beta Z_{jk})}$, with β equal to 0.3581. The censoring time U_{jk} for the *k*th subject in the *j*th cluster is generated from three censoring distributions: (i) the censoring points are (3, 4, 5, 6, 7) with probability (0.2, 0.2, 0.2, 0.2, 0.2); (ii) the censoring points are (1, 2, 3, 4, 5) with probability (0.2, 0.2, 0.2, 0.2, 0.2); (iii) the censoring points are (1, 2, 3, 4, 5) with probability (0.6, 0.1, 0.1, 0.1, 0.1). Finally, for the *k*th subject in the *j*th cluster, if the cure status C_{jk}=1, the observed survival time is setting to $T_{jk}=U_{jk}$ with the right-censored indicator δ_{jk} = 0, otherwise the observed survival time T_{jk} = min$\{X_{jk},U_{jk}\}$ with the right-censored indicator δ_{jk} = $I(X_{jk}<U_{jk})$.

To understand the effect of equal cluster sizes on parameter estimation accuracy, the number of individuals within each cluster is chosen to be 2, 3, and 5 with the number of clusters chosen to be 300, 200 and 120, respectively. Therefore, total number of subjects is 600 for each configuration. The simulation is repeated 500 times for all configurations. Tables I and II display the biases and standard errors of the estimators from the data with cluster size = 2 and 5, respectively. It can be seen that, the biases of the estimated baseline hazard rates $\hat{\lambda}_{i0}^*$, uncured probability *p* and β are very reasonable under three censoring proportions. As the censoring proportion increases, the standard errors of $\hat{\lambda}_{i0}^*$, $\hat{p}$, $\hat{\sigma}$ and $\hat{\beta}$ become larger. As cluster size increases, the biases of $\hat{\sigma}$ and $\hat{\beta}$ decrease.

Moreover, to understand the impact of unequal cluster sizes on parameter estimation accuracy, the cluster size of each cluster is generated from discrete uniform distribution with 5 possible values 1, 2, 3, 4, and 5, and associated probability of 0.2, 0.2, 0.2, 0.2, 0.2, respectively. Table III displays the biases and standard errors of the estimators for three censoring settings based on 500 simulation runs with 300 clusters in each run. It can be seen that, the biases of the estimated baseline hazard rates $\hat{\lambda}_{i0}^*$, uncured probability *p* and β are very reasonable under three censoring proportions. Likewise, as the censoring proportion increases, the standard errors of $\hat{\lambda}_{i0}^*$, $\hat{p}$, $\hat{\sigma}$ and $\hat{\beta}$ become larger.

4. Example

The model displayed in (2.1) is applied to one-stage non-submerged dental implant study described in Section 1. Gender and implant case type are considered as risk factors for demonstration purposes. There are

149 male and 175 female patients with 359 and 383 implants, respectively, the average cluster sizes of male and female patients are about 2.2 and 2.4, respectively, and the numbers of implant failures of male and female patients are 77 and 59, respectively. The Turnbull estimates of the survival functions for the two groups shows a slightly larger difference in the late period of the study. Implant case types is classified as 4 groups, Single (single tooth replacement, 224 implants), FPD (458 implants), FDB (Fixed detachable bridge, 19 implants) and overdenture (41 implants). Likewise, the Turnbull estimates of the survival functions for single and FPD implant case type shows a slightly larger difference in the late period of the study.

The estimates from marginal regression approach and gamma frailty model are listed in Table V. Both approaches identify that that female has longer implant survival time. However, only the marginal regression approach identifies that FPD case type has longer implant survival time. Hence, diagnostic methods for identifying the model are need.

Table I. The biases and standard errors of the estimators from the data with cluster size = 2.

	Case I (20.29%)			Case II (43.26%)			Case III (61.63%)		
	Estimate	*Bias*	*SE*	*Estimate*	*Bias*	*S.E.*	*Estimate*	*Bias*	*SE*
λ^*_{10}	0.2003	0.0003	0.0229	0.2004	0.0004	0.0236	0.2016	0.0016	0.0268
λ^*_{20}	0.3742	–0.0008	0.0384	0.3749	–0.0001	0.0465	0.3783	0.0033	0.0639
λ^*_{30}	0.3002	0.0002	0.0451	0.3010	0.0010	0.0598	0.3088	0.0088	0.0905
λ^*_{40}	0.7081	–0.0059	0.0784	0.7079	–0.0061	0.1053	0.7202	0.0062	0.1426
σ	2.1934	0.1934	0.7507	2.3866	0.3866	1.3654	2.8543	0.8543	2.7839
β	0.3635	0.0054	0.1341	0.3644	0.0063	0.1495	0.3640	0.0059	0.1746
p	0.8984	–0.0016	0.0152	0.8982	–0.0018	0.0247	0.8964	–0.0036	0.0378

The value inside the parentheses is the average censoring proportion.

Table II. The biases and standard errors of the estimators from the data with cluster size = 5

	Case I (20.05%)			Case II (43.08%)			Case III (61.47%)		
	Estimate	*Bias*	*SE*	*Estimate*	*Bias*	*S.E.*	*Estimate*	*Bias*	*SE*
λ^*_{10}	0.2003	0.0003	0.0226	0.1999	–0.0001	0.0242	0.2000	0.0000	0.0268
λ^*_{20}	0.3756	0.0006	0.0367	0.3746	–0.0004	0.0415	0.3745	–0.0005	0.0552
λ^*_{30}	0.3008	0.0008	0.0424	0.2994	–0.0006	0.0529	0.3013	0.0013	0.0734
λ^*_{40}	0.7138	–0.0002	0.0616	0.7107	–0.0033	0.0858	0.7118	–0.0022	0.1133
σ	2.1535	0.1535	0.5805	2.2278	0.2278	0.7334	2.3670	0.3670	1.1840
β	0.3544	–0.0037	0.1195	0.3592	0.0011	0.1378	0.3633	0.0052	0.1606
p	0.8995	–0.0005	0.0146	0.8985	–0.0015	0.0235	0.8983	–0.0017	0.0352

The value inside the parentheses is the average censoring proportion.

Table III. The biases and standard errors of the estimators from the data with variable cluster sizes.

	Case I (20.12%)			Case II (43.05%)			Case III (61.50%)		
	Estimate	*Bias*	*SE*	*Estimate*	*Bias*	*S.E.*	*Estimate*	*Bias*	*SE*
λ_{10}^*	0.1989	–0.0011	0.0178	0.1982	–0.0018	0.0188	0.1984	–0.0016	0.0212
λ_{20}^*	0.3716	–0.0034	0.0299	0.3705	–0.0045	0.0336	0.3720	–0.0030	0.0457
λ_{30}^*	0.2955	–0.0045	0.0322	0.2936	–0.0064	0.0429	0.2945	–0.0055	0.0588
λ_{40}^*	0.7072	–0.0068	0.0537	0.7062	–0.0078	0.0735	0.7035	–0.0105	0.1050
σ	2.1000	0.1000	0.4528	2.1548	0.1548	0.5878	2.2153	0.2153	0.8126
β	0.3629	0.0048	0.1038	0.3687	0.0106	0.1183	0.3711	0.0130	0.1381
p	0.8999	–0.0001	0.0115	0.9010	0.0010	0.0193	0.9010	0.0010	0.0270

The value inside the parentheses is the average censoring proportion.

Table IV. Estimates from marginal regression approach and gamma frailty model.

	Methods			
	Marginal approach		*Gamma frailty model*	
Covariates	*Estimate* (*SE*)	*p-value*	*Estimate* (*SE*)	*p-value*
Gender (female =0)				
male=1	0.393 (0.192)	0.041	1.563 (0.772)	0.043
Cure rate	0.655 (0.037)		0.515 (0.058)	
σ			6.297 (1.170)	
Implant case type (single =0)				
FPD	–0.466 (0.216)	0.031	–0.643 (0.384)	0.094
FBD	–0.992 (0.761)	0.192	0.234 (0.608)	0.884
Overdenture	0.521 (0.362)	0.150	0.927 (1.148)	0.419
Cure rate	0.639 (0.042)		0.541 (0.059)	
σ			4.263 (0.941)	

REFERENCES (RÉFERENCES)

1. Chatterjee, N. and Shih, J. (2001). A bivariate cure-mixture approach for modeling familial association in disease. *Biometrics*, **57**: 779–786.

2. Chi, Y., Chen, C. M. and Su, P. F. (2010). Marginal regression approach in cure models with clustered discrete survival data. *Technical Report*.

3. Longini I. M, Halloran M. E. (1996). A frailty mixture model for estimating vaccine efficacy. *Applied. Statistics*, 45: 165–173.

4. Peng, Y. (2003). Fitting semiparametric cure models. *Computational Statistics and Data Analysis*, **41**: 481–490.

5. Price, Y, Dear, K. B. G. and Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, **17**: 813–830.

6. Wen, M. J., Tseng, C. C., Lee, C. K., (2008). Life table analysis for evaluating curative-effect of one-stage non-submerged dental implant in Taiwan. *Journal of Data Science*, 6: 591-599.

7. Wienke, A., Lichtenstein, P. and Yashin, A. I. (2003). A bivariate frailty model with a cure fraction for modeling familial correlation in disease. *Biometrics*, **59**: 1178–1183.

8. Zhao, X., Zhou, X. (2008). Discrete-time survival models with long-term survivors. *Statistics in Medicine*, 27: 1261-1281.