应用概率统计 第 33 卷
第 5 期 2017 年 10 月

Chinese Journal of Applied Probability and Statistics
Oct., 2017, Vol. 33, No. 5, pp. 517-528
doi: 10.3969/j.issn.1001-4268.2017.05.008

Regression Analysis of Clustered Failure Time Data under the Additive Hazards Model *

DI RongRong
(*School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, China*)
WANG ChengYong*
(*School of Mathematics and Computer Science, Hubei University of Arts and Science, Xiangyang, 441053, China*)

Abstract: Clustered interval-censored failure time data often arises in medical studies when study subjects come from the same cluster. Furthermore, the failure time may be related to the cluster size. A simple and common approach is to simplify interval-censored data due to the lack of proper inference procedures for direct analysis. For this reason, we proposed the within-cluster resampling-based method to consider the case II interval-censored data under the additive hazards model. With-cluster resampling is simple but computationally intensive. A major advantage of the proposed approach is that the estimator can be easily implemented when the cluster size is informative. Asymptotic properties and some simulation results are provided and indicate that the proposed approach works well.

Keywords: additive hazards model; interval-censored; within-cluster resampling; semiparametric regression

2010 Mathematics Subject Classification: 62N02

Citation: Di R R, Wang C Y. Regression analysis of clustered failure time data under the additive hazards model [J]. Chinese J. Appl. Probab. Statist., 2017, 33(5): 517–528.

§1. Introduction

Case II interval-censored data is commonly encountered in biomedicine where the event time of interest is not observed directly but known only to lie between the two monitoring times. A few methods have been proposed for regression analysis of the interval-censored data. Zeng et al. [1] discussed regression analysis of case II interval-censored data, using the additive hazards model. Wang et al. [2] developed an approach which is easy to implement for case II interval-censored data and allows that the monitoring times

*The project was supported by the National Natural Science Foundation of China (Grant No. 71371066).
*Corresponding author, E-mail: wchyxf@163.com.
Received June 20, 2016. Revised November 17, 2016.

are random and continuous. It assumes that the failure time of interest follows Cox-type models [3]. However, these methods do not take into account the clustered data. In some cases, failure time data certainly comes from the same cluster. For example, failure time can be the time to disease occurrence for the patients in the same family or the same clinic. Li et al. [4] proposed an estimating equation-based approach for regression analysis of clustered interval-censored failure time data generated from the additive hazards model which does not involve the estimation of any baseline hazard function.

Another commonly used statistical method to analyse clustered failure time data is the gamma-frailty model, incorporating an unobserved random effect known as frailty into the Cox proportional hazards model. Li et al. [5] proposed a sieve estimation procedure for fitting a Cox frailty model to clustered interval-censored failure time data. A two-step algorithm for parameter estimation was developed and the asymptotic properties of the resulting sieve maximum likelihood estimators were established. Kor et al. [6] gave a method for analyzing clustered interval-censored data based on Cox's model. As pointed out in [7] that the additive hazards model describes a different aspect of the association between the failure time and covariates compared with the Cox's model and the additive model could be more plausible than the Cox's model in many applications. This is especially the situation when one is interested in the risk difference as often the case in epidemiology and public health [8].

In this paper, we consider case II interval-censored data under the additive hazards model and the situation where the correlated failure time of interest may be related to cluster size. We assume that there exist only two monitoring times independent of the failure time of interest under the given covariate process. We then use the within-cluster resampling (WCR) procedure under the additive hazards model. WCR is a method for analyzing clustered data in the presence of informative cluster size when estimation of marginal effects weighted at the cluster level is of interest. Parameter estimation with WCR is based on resampling replicate data sets, each containing one observation from each cluster. In the following, we present the approach under the additive hazards model.

The rest of the paper is organized as follows. Section 2 proposes the model and some notations used in this paper. Section 3 gives a method based on the WCR method by using the inference procedure proposed by [2] under the additive hazards model for case II failure time data, and Section 4 presents some extensive simulation studies to assess the performance of the proposed approach.

§2. Notation and Model

Suppose that there are n independent clusters and each cluster has n_{ij} exchangeable subjects for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n_i$. Let U_{ij} and V_{ij} denote the two monitoring

times for the j-th subject in the i-th cluster. Let $Z_{ij}(t)$ be the corresponding p-dimensional vector of covariates that may depend on time t, and T_{ij} denote the failure time of interest for subject j in the cluster i which is independent of monitoring times U_{ij} and V_{ij} given covariate $Z_{ij}(t)$. For each (i,j), define $\delta_{1ij} = I(T_{ij} < U_{ij})$, $\delta_{2ij} = I(U_{ij} \leqslant T_{ij} < V_{ij})$ and $\delta_{3ij} = 1 - \delta_{1ij} - \delta_{2ij}$. The observed data are $(U_{ij}, V_{ij}, \delta_{1ij}, \delta_{2ij}, \delta_{3ij}, Z_{ij}(\cdot))$.

It just as pointed out in [9], the cause for cluster sizes being informative can be complicated and usually unknown, and some latent variables may implicitly affect the baseline hazard for each cluster and/or covariates. For example, the marginal hazard function may be associated with the cluster size through the following frailty mode

$$\lambda_{ij}(t \,|\, Z_{ij}) = \lambda_0(t) + \omega_i \beta_0' Z_{ij}(t),$$

where β_0 is the unknown vector of p-dimensional regression coefficient, ω_i is the cluster-specific random effect to account for within-cluster correlation in cluster i, and $\lambda_0(t)$ is the unknown baseline hazard function. If cluster sizes are ignorable (noninformative to survival), the usual marginal additive hazards model [10] is applicable, given by

$$\lambda_{ij}(t \,|\, Z_{ij}) = \lambda_0(t) + \beta_0' Z_{ij}(t). \tag{1}$$

Motivated by the work of [2], we model the monitoring variables using Cox-type hazard functions

$$\lambda_{ij}^U(t \,|\, Z_{ij}) = \lambda_1(t) \mathrm{e}^{\gamma_0' Z_{ij}(t)}, \tag{2}$$

$$\lambda_{ij}^V(t \,|\, U_{ij}, Z_{ij}) = I(t > U_{ij}) \lambda_2(t) \mathrm{e}^{\gamma_0' Z_{ij}(t)}, \tag{3}$$

where $\lambda_1(t)$ and $\lambda_2(t)$ denote unspecified baseline functions, γ_0 is the unknown vector of regression parameters. For each i and j, define $N_{ij}^{(1)}(t) = (1 - \delta_{1ij}) I(U_{ij} \leqslant t)$, and conditional on U_{ij}, define $N_{ij}^{(2)}(t) = \delta_{3ij} I(V_{ij} \leqslant t)$ if $t \geqslant U_{ij}$ and 0 if $t < U_{ij}$. We also define

$$\lambda_{ij}^{(1)}(t \,|\, Z_{ij}) = \lambda_1(t) \mathrm{e}^{-\Lambda_0(t)} \mathrm{e}^{-\beta_0' Z_{ij}^*(t) + \gamma_0' Z_{ij}(t)} := \lambda_{10}(t) \mathrm{e}^{-\beta_0' Z_{ij}^*(t) + \gamma_0' Z_{ij}(t)} \tag{4}$$

and

$$\begin{aligned} \lambda_{ij}^{(2)}(t \,|\, U_{ij}, Z_{ij}) &= I(t > U_{ij}) \lambda_2(t) \mathrm{e}^{-\Lambda_0(t)} \mathrm{e}^{-\beta_0' Z_{ij}^*(t) + \gamma_0' Z_{ij}(t)} \\ &:= I(t > U_{ij}) \lambda_{20} \mathrm{e}^{-\beta_0' Z_{ij}^*(t) + \gamma_0' Z_{ij}(t)}, \end{aligned} \tag{5}$$

where $Z_{ij}^*(t) = \int_0^t Z_{ij}(s) \mathrm{d}s$, $\Lambda_0(t) = \int_0^t \lambda_0(s) \mathrm{d}s$, $\lambda_{10} = \lambda_1(t) \mathrm{e}^{-\Lambda_0(t)}$ and $\lambda_{20} = \lambda_2(t) \mathrm{e}^{-\Lambda_0(t)}$. Clearly models (4) and (5) satisfy the Cox proportional hazards model.

§3. The WCR-Based Procedure

When cluster sizes are informative, to estimate the unknown parameter vectors β_0 and γ_0, the estimates and inference based on equation (2) may be incorrect. To account for informative cluster sizes, this section will propose a method based on the within-cluster resampling (WCR) technique. The basic idea behind the WCR-based procedure is that one observation is randomly sampled with replacement from each of the n clusters using the WCR approach (refer to [11]). For this, let Q be a positive integer, we randomly sample one subject with replacement from each of the n clusters and suppose that the resampling process is repeated Q times. Let τ denote a known time for the length of study period, define $\delta_{1i}^q = I(T_i^q < U_i^q)$, $\delta_{2i}^q = I(U_i^q \leqslant T_i^q < V_i^q)$ and $\delta_{3i}^q = 1 - \delta_{1i}^q - \delta_{2i}^q$, the q-th resampled data set denoted by $\{U_i^q, V_i^q, \delta_{1i}^q, \delta_{2i}^q, \delta_{3i}^q, Z_i^q(t); i = 1, 2, \ldots, n, 0 \leqslant t \leqslant \tau\}$, consists of n independent observations, which can be analyzed using the models (4) and (5) for independent data set.

The within-cluster resampling estimate is constructed as the average of the Q resample-based estimates. For the q-th resampled data, to estimate β_0 and γ_0, motivated by [2], we first estimate γ_0, and for this, for $i = 1, 2, \ldots, n$ and $q = 1, 2, \ldots, Q$, we define $\widetilde{N}_i^{(1)q}(t) = I(U_i^q \leqslant t)$ and $\widetilde{N}_i^{(2)q}(t) = I(V_i^q \leqslant t)$ if $t \geqslant U_i^q$ and 0 if $t < U_i^q$ given the observed U_i^q. For $j = 0$ and 1, also define

$$S_{1,\gamma,q}^{(j)}(t,\gamma) = \frac{1}{n}\sum_{i=1}^{n} I(t \leqslant U_i^q)\mathrm{e}^{\gamma' Z_i^q(t)}(Z_i^q(t))^{\otimes j},$$
$$S_{2,\gamma,q}^{(j)}(t,\gamma) = \frac{1}{n}\sum_{i=1}^{n} I(U_i^q < t \leqslant V_i^q)\mathrm{e}^{\gamma' Z_i^q(t)}(Z_i^q(t))^{\otimes j},$$

where $a^{\otimes j} = 1$ and a for $j = 0$ and 1. We construct an estimating function $U_\gamma^q(\gamma)$ for γ_0 as

$$\sum_{i=1}^{n}\Big[\int_0^\infty \Big(Z_i^q(t) - \frac{S_{1,\gamma,q}^{(1)}(t,\gamma)}{S_{1,\gamma,q}^{(0)}(t,\gamma)}\Big)\mathrm{d}\widetilde{N}_i^{(1)q}(t) + \int_0^\infty \Big(Z_i^q(t) - \frac{S_{2,\gamma,q}^{(1)}(t,\gamma)}{S_{2,\gamma,q}^{(0)}(t,\gamma)}\Big)\mathrm{d}\widetilde{N}_i^{(2)q}(t)\Big]$$
$$= \sum_{i=1}^{n}\Big\{Z_i^q(U_i^q) - \frac{S_{1,\gamma,q}^{(1)}(U_i^q,\gamma)}{S_{1,\gamma,q}^{(0)}(U_i^q,\gamma)}\Big\} + \sum_{i=1}^{n}\Big\{Z_i^q(V_i^q) - \frac{S_{2,\gamma,q}^{(1)}(V_i^q,\gamma)}{S_{2,\gamma,q}^{(0)}(V_i^q,\gamma)}\Big\}.$$

Let $\widehat{\gamma}_q$ be the solution to $U_\gamma^q(\gamma) = 0$. Next we estimate β_0 given $\widehat{\gamma}_q$. For this, we also define $N_i^{(1)q}(t) = (1 - \delta_{1i}^q)I(U_i^q \leqslant t)$, $N_i^{(2)q}(t) = \delta_{3i}^q I(V_i^q \leqslant t)$ for $i = 1, 2, \ldots, n$, $q = 1, 2, \ldots, Q$, and for $j = 0, 1$, let

$$S_{1,\beta,q}^{(j)}(t,\beta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} I(t \leqslant U_i^q)\mathrm{e}^{-\beta' Z_i^{q*}(t) + \gamma' Z_i^q(t)}(Z_i^{q*}(t))^{\otimes j},$$
$$S_{2,\beta,q}^{(j)}(t,\beta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} I(U_i^q < t \leqslant V_i^q)\mathrm{e}^{-\beta' Z_i^{q*}(t) + \gamma' Z_i^q(t)}(Z_i^{q*}(t))^{\otimes j}.$$

We propose the estimating equation $U_\beta^q(\beta, \widehat{\gamma}_q) = 0$, where $U_\beta^q(\beta, \gamma)$ is defined as

$$\sum_{i=1}^{n}(1-\delta_{1i}^q)\Big(Z_i^{q*}(U_i^q) - \frac{S_{1,\beta,q}^{(1)}(U_i^q,\beta,\gamma)}{S_{1,\beta,q}^{(0)}(U_i^q,\beta,\gamma)}\Big) + \sum_{i=1}^{n}\delta_{3i}^q\Big(Z_i^{q*}(V_i^q) - \frac{S_{2,\beta,q}^{(1)}(V_i^q,\beta,\gamma)}{S_{2,\beta,q}^{(0)}(V_i^q,\beta,\gamma)}\Big),$$

where $Z_i^{q*}(t) = \int_0^t Z_i^q(s)\mathrm{d}s$. Then we can estimate β_0 by $\widehat{\beta}_q$ defined as the root of $U_\beta^q(\beta, \widehat{\gamma}_q) = 0$. Furthermore, Wang et al. [2] showed that $\sqrt{n}(\widehat{\beta}^q - \beta_0)$ can be asymptotically approximated by a normal vector with mean zero and a covariance matrix of $\widehat{\beta}_q$ that can be consistently estimated by $\widehat{\Sigma}_q := (\widehat{A}_\beta^q)^{-1}\widehat{\Gamma}_q(\widehat{A}_\beta^q)^{-1}/n$, where $\widehat{A}_\beta^q$ and $\widehat{\Gamma}_q$ will be defined in the Appendix, thus $\widehat{\beta}_q$ is consistent.

As it is known to all that sample mean can reduce the system error, after repeating this procedure Q times, the WCR estimator for β_0 can be constructed as the average of the Q resample-based estimators, which is

$$\widehat{\beta}_{\mathrm{wcr}} = \frac{1}{Q}\sum_{q=1}^{Q}\widehat{\beta}_q.$$

Under some regularity conditions, it can be shown that $\sqrt{n}(\widehat{\beta}_{\mathrm{wcr}} - \beta_0)$ converges in distribution to a zero-mean normal random vector, and the variance-covariance matrix of $\widehat{\beta}_{\mathrm{wcr}}$ can be consistently estimated by

$$\widehat{\Sigma}_{\mathrm{wcr}} = \frac{1}{Q}\sum_{q=1}^{Q}\widehat{\Sigma}_q - \frac{1}{Q}\sum_{q=1}^{Q}(\widehat{\beta}_q - \widehat{\beta}_{\mathrm{wcr}})(\widehat{\beta}_q - \widehat{\beta}_{\mathrm{wcr}})'.$$

The proof of this result is sketched in the Appendix.

§4. Simulation

An extensive simulation study was conducted to assess the finite sample performance of the estimates proposed in the previous sections. For simplicity, here only consider noninformative cases. In the simulation study, the true covariate Z_{ij} generated from the Bernoulli distribution $\mathrm{B}(1, 0.5)$. Given the Z_{ij}'s, the failure times of interest were assumed to follow model (1) with $\lambda_0(t) = 2$ or $\lambda_0(t) = 4$, the observation times U_{ij}'s and V_{ij}'s, generated from (2) and (3) with $\lambda_1(t) = 4$, $\lambda_2(t) = 2$ or $\lambda_1(t) = 8$, $\lambda_2(t) = 4$. The cluster size n_i was randomly generated from uniform distribution $\mathrm{U}\{1, 2, 3, 4, 5, 6, 7\}$. The results given below are based on 400 replications with $Q = 400$ resamples and the number of clusters $n = 200$ or 400.

Table 1 and Table 2 present the results on estimation of (γ_0, β_0) with true values $(\gamma_0, \beta_0) = (0, 0), (0, 0.2), (0, -0.2), (0.2, 0), (0.2, 0.2)$ or $(0.2, -0.2)$. The results include the estimated biases (Bias) given by the averages of the point estimates minus the true values, the averages of the standard error estimates (SEE), the sampling standard errors of the

point estimates (SSE) and the 95% percent empirical coverage probabilities (CP). The results indicate that the proposed estimate seems to be approximately unbiased and the proposed variance estimate also seems to be reasonable, and all estimates become better when the sample size increases.

Table 1 Simulation results for estiamtion of β_0 and γ_0 with $\lambda_0 = 2$, $\lambda_1 = 4$, $\lambda_2 = 2$

		$n = 200$				$n = 400$			
(γ_0, β_0)		BIAS	SEE	SSE	CP	BIAS	SEE	SSE	CP
(0,0)	$\widehat{\gamma}$	0.0010	0.0603	0.0610	0.9575	0.0001	0.0436	0.0427	0.9475
	$\widehat{\beta}$	-0.0015	0.2360	0.2381	0.9475	0.0020	0.1704	0.1708	0.9475
(0,0.2)	$\widehat{\gamma}$	-0.0029	0.0580	0.0602	0.9550	0.0037	0.0424	0.0430	0.9475
	$\widehat{\beta}$	0.0245	0.2485	0.2442	0.9375	-0.0020	0.1767	0.1780	0.9450
(0,-0.2)	$\widehat{\gamma}$	0.0005	0.0619	0.0604	0.9475	0.0001	0.0436	0.0427	0.9475
	$\widehat{\beta}$	0.0034	0.2463	0.2283	0.9375	0.0040	0.1642	0.1639	0.9425
(0.2,0)	$\widehat{\gamma}$	-0.0030	0.0612	0.0606	0.9500	-0.0018	0.04180	0.0429	0.9475
	$\widehat{\beta}$	0.0061	0.2447	0.2405	0.9450	0.0037	0.1760	0.1734	0.9475
(0.2,0.2)	$\widehat{\gamma}$	0.0026	0.0640	0.0611	0.9400	-0.0008	0.0431	0.0430	0.9450
	$\widehat{\beta}$	0.0050	0.2654	0.2556	0.9400	0.0027	0.1808	0.1819	0.9450
(0.2,-0.2)	$\widehat{\gamma}$	-0.0030	0.0612	0.0606	0.9500	-0.0008	0.0431	0.0430	0.9450
	$\widehat{\beta}$	0.0012	0.2282	0.2316	0.9500	-0.0048	0.1681	0.1667	0.9500

Table 2 Simulation results for estiamtion of β_0 and γ_0 with $\lambda_0 = 4$, $\lambda_1 = 8$, $\lambda_2 = 4$

		$n = 200$				$n = 400$			
(γ_0, β_0)		BIAS	SEE	SSE	CP	BIAS	SEE	SSE	CP
(0,0)	$\widehat{\gamma}$	-0.0017	0.0635	0.0602	0.9400	0.0001	0.0436	0.0427	0.9475
	$\widehat{\beta}$	-0.0021	0.5167	0.4735	0.9350	0.0041	0.3401	0.3417	0.9475
(0,0.2)	$\widehat{\gamma}$	-0.0017	0.0603	0.0602	0.9400	0.0001	0.0436	0.0427	0.9475
	$\widehat{\beta}$	-0.0017	0.5323	0.4837	0.9375	0.0081	0.3491	0.3492	0.9475
(0,-0.2)	$\widehat{\gamma}$	-0.0047	0.0612	0.0603	0.9550	0.0037	0.0424	0.0430	0.9475
	$\widehat{\beta}$	0.0135	0.4807	0.4626	0.9425	-0.0163	0.3340	0.3317	0.9475
(0.2,0)	$\widehat{\gamma}$	0.0026	0.0640	0.0611	0.9400	-0.0008	0.0436	0.0431	0.9450
	$\widehat{\beta}$	-0.0012	0.5143	0.4891	0.9450	0.0027	0.3459	0.3480	0.9450
(0.2,0.2)	$\widehat{\gamma}$	-0.0030	0.0612	0.0606	0.9500	-0.0018	0.0418	0.0429	0.9475
	$\widehat{\beta}$	0.0172	0.5036	0.4913	0.9425	-0.0202	0.3610	0.3549	0.9425
(0.2,-0.2)	$\widehat{\gamma}$	-0.0030	0.0612	0.0606	0.9500	-0.0008	0.0431	0.0432	0.9450
	$\widehat{\beta}$	0.0066	0.4712	0.4721	0.9400	0.0011	0.3396	0.3404	0.9475

For comparison, we also consider the correlated failure times model used in [4], that

is,

$$\lambda_{ij}(t \mid Z_{ij}, b_i) = \lambda_0(t) + \beta_0' Z_{ij} + b_i \tag{6}$$

with $\lambda_0(t) = 2$. The latent variables b_i's were assumed to follow a normal distribution with zero mean and variance equal to 1/4. The covariates Z_{ij}'s were generated from the Bernoulli distribution with success probability $p = 0.5$. The monitoring variables U_{ij}'s and V_{ij}'s were generated from (2) and (3) with $\lambda_1(t) = 4$ and $\lambda_2(t) = 2$. The cluster size n_i was generated from the uniform distribution $\mathrm{U}\{2, 3, 4\}$ and the number of clusters $n = 200$. The results based on 1 000 replications and the WCR method with $Q = 400$ resamples for each step. The true regression parameter γ_0 was taken to be 0.25, 0 and β_0 was 0.25, 0 and -0.25. Simulated results are listed in Table 3, and all the results listed below "Li, Wang and Sun" are extracted directly from the paper of [4]. These results indicate that the proposed procedure actually better performance than the method given by [4]. The proposed method seems to give smaller biases and standard errors. This is because the individuals are related and the WCR method take into account the correlation compared with the method given by [4]. So the WCR method is more effective.

Table 3 Compared simulation results for estiamtion with $\lambda_0 = 2$, $\lambda_1 = 4$, $\lambda_2 = 2$, $n = 200$

		Li, Wang and Sun				WCR			
(γ_0, β_0)		BIAS	SEE	SSE	CP	BIAS	SEE	SSE	CP
(0,0)	$\widehat{\gamma}$	0.0019	0.1106	0.1081	0.948	0.0020	0.0602	0.0597	0.951
	$\widehat{\beta}$	0.0315	0.8804	0.8273	0.975	-0.0047	0.2460	0.2298	0.937
(0,0.25)	$\widehat{\gamma}$	0.0073	0.1102	0.1065	0.946	0.0020	0.0602	0.0597	0.951
	$\widehat{\beta}$	-0.0225	0.8835	0.8303	0.973	0.0030	0.2618	0.2430	0.936
(0,-0.25)	$\widehat{\gamma}$	-0.0051	0.1126	0.1100	0.952	-0.0050	0.0590	0.0595	0.950
	$\widehat{\beta}$	0.0172	0.8776	0.8255	0.938	-0.0055	0.2257	0.2191	0.941
(0.25,0)	$\widehat{\gamma}$	0.0035	0.1154	0.1126	0.941	0.0007	0.0594	0.0603	0.946
	$\widehat{\beta}$	0.0850	0.9610	0.9150	0.940	0.0044	0.2454	0.2393	0.943
(0.25,0.25)	$\widehat{\gamma}$	0.0122	0.1138	0.1107	0.948	0.0007	0.0594	0.0603	0.946
	$\widehat{\beta}$	0.0731	0.9634	0.9133	0.971	0.0107	0.2609	0.2526	0.942
(0.25,-0.25)	$\widehat{\gamma}$	-0.0036	0.1177	0.1151	0.941	0.0007	0.0594	0.0603	0.946
	$\widehat{\beta}$	0.0992	0.9643	0.9112	0.963	-0.0013	0.2309	0.2271	0.952

Finally, it can be seem from the Tables 1 – 3 that $\widehat{\gamma}$ seems to have smaller standard error than that of $\widehat{\beta}$ for all the estimates. This is because that completely observed data can be used for the estimate of γ, while only incompletely observed data for the estimate of β.

Appendix: Proofs of the Asymptotic Normality of $\widehat{\beta}_{\mathbf{wcr}}$

Proof For $i = 1, 2, \ldots, n$, we first define

$$M_i^{(1)q}(t) = N_i^{(1)q}(t) - \int_0^t I(s \leqslant U_i^q)\lambda_{10}(s)\mathrm{e}^{-\beta_0' Z_i^{q*}(s)+\gamma_0' Z_i^q(s)}\mathrm{d}s,$$
$$M_i^{(2)q}(t) = N_i^{(2)q}(t) - \int_0^t I(U_i^q < s \leqslant V_i^q)\lambda_{20}(s)\mathrm{e}^{-\beta_0' Z_i^{q*}(s)+\gamma_0' Z_i^q(s)}\mathrm{d}s,$$
$$\widetilde{M}_i^{(1)q}(t) = \widetilde{N}_i^{(1)q}(t) - \int_0^t I(s \leqslant U_i^q)\lambda_1(s)\mathrm{e}^{\gamma_0' Z_i^q(s)}\mathrm{d}s,$$
$$\widetilde{M}_i^{(2)q}(t) = \widetilde{N}_i^{(2)q}(t) - \int_0^t I(U_i^q < s \leqslant V_i^q)\lambda_2(s)\mathrm{e}^{\gamma_0' Z_i^q(s)}\mathrm{d}s,$$

which are martingales.

Since $\widehat{\beta}_q$ is the solution of the estimating equation $U_\beta^q(\beta, \widehat{\gamma}_q) = 0$. By the Taylor's expansion, we have

$$-U_\beta^q(\beta_0, \widehat{\gamma}_q) = U_\beta^q(\widehat{\beta}_q, \widehat{\gamma}_q) - U_\beta^q(\beta_0, \widehat{\gamma}_q) = \frac{\partial U_\beta^q(\beta^*, \widehat{\gamma}_q)}{\partial \beta^*}(\widehat{\beta}_q - \beta_0),$$

where β^* is on the line segment between $\widehat{\beta}_q$ and β_0. Rewriting the above equation yields that

$$\sqrt{n}(\widehat{\beta}_q - \beta_0) = \Big(-\frac{1}{n}\frac{\partial U_\beta^q(\beta^*, \widehat{\gamma}_q)}{\partial \beta^*}\Big)^{-1}\Big(\frac{1}{\sqrt{n}}U_\beta^q(\beta_0, \widehat{\gamma}_q)\Big).$$

Note that $-n^{-1}\partial U_\beta^q(\beta, \gamma)/\partial\beta$ is equal to

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \int_0^\infty \Big(\frac{S_{1,\beta,q}^{(2)}(t,\beta,\gamma)}{S_{1,\beta,q}^{(0)}(t,\beta,\gamma)} - \frac{(S_{1,\beta,q}^{(1)}(t,\beta,\gamma))^{\otimes 2}}{(S_{1,\beta,q}^{(0)}(t,\beta,\gamma))^2}\Big)\mathrm{d}N_i^{(1)q}(t)\\
&+ \frac{1}{n}\sum_{i=1}^n \int_0^\infty \Big(\frac{S_{2,\beta,q}^{(2)}(t,\beta,\gamma)}{S_{2,\beta,q}^{(0)}(t,\beta,\gamma)} - \frac{(S_{2,\beta,q}^{(1)}(t,\beta,\gamma))^{\otimes 2}}{(S_{2,\beta,q}^{(0)}(t,\beta,\gamma))^2}\Big)\mathrm{d}N_i^{(2)q}(t)\\
=\ &\frac{1}{n}\sum_{i=1}^n \int_0^\infty (Z_i^{q*}(s) - \overline{Z}_{(1)}^q(\beta,\gamma,t))^{\otimes 2} I(U_i^q \geqslant t)\mathrm{e}^{-\beta' Z_i^{q*}(t)+\gamma' Z_i^q(t)}\frac{\mathrm{d}\overline{N}^{(1)q}(t)}{S_{1,\beta,q}^{(0)}(t,\beta,\gamma)}\\
&+ \frac{1}{n}\sum_{i=1}^n \int_0^\infty (Z_i^{q*}(s) - \overline{Z}_{(2)}^q(\beta,\gamma,t))^{\otimes 2} I(U_i^q \leqslant t < V_i^q)\mathrm{e}^{-\beta' Z_i^{q*}(t)+\gamma' Z_i^q(t)}\frac{\mathrm{d}\overline{N}^{(2)q}(t)}{S_{2,\beta,q}^{(0)}(t,\beta,\gamma)},
\end{aligned}$$

where

$$\overline{Z}_{(1)}^q(\beta,\gamma,t) = \frac{S_{1,\beta,q}^{(1)}(t,\beta,\gamma)}{S_{1,\beta,q}^{(0)}(t,\beta,\gamma)}, \quad \overline{Z}_{(2)}^q(\beta,\gamma,t) = \frac{S_{2,\beta,q}^{(1)}(t,\beta,\gamma)}{S_{2,\beta,q}^{(0)}(t,\beta,\gamma)}, \quad \overline{N}^{(1)q}(t) = \frac{1}{n}\sum_{i=1}^n N_i^{(1)q}(t),$$

and $\overline{N}^{(2)q}(t) = n^{-1}\sum_{i=1}^n N_i^{(2)q}(t)$. It can be easily seen that $-n^{-1}\partial U_\beta^q(\beta_0, \widehat{\gamma}^q)/\partial\beta_0$ is positive definite and $-n^{-1}\partial U_\beta^q(\beta_0, \gamma_0)/\partial\beta$ converges in probability to a deterministic and

positive definite matrix denoted by A_β, which can be consistently estimated by $\widehat{A}_\beta^q := -n^{-1}\partial U_\beta^q(\beta^*, \widehat{\gamma}_q)/\partial\beta^*$.

Averaging over $q = 1, 2, \ldots, Q$ resamples, it yields that

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_{\mathrm{wcr}} - \beta_0) &= \frac{1}{Q}\sum_{q=1}^{Q}\sqrt{n}(\widehat{\beta}_q - \beta_0)\\ &= \frac{1}{Q}\sum_{q=1}^{Q}\Big(-\frac{1}{n}\frac{\partial U_\beta^q(\beta^*, \widehat{\gamma}_q)}{\partial\beta^*}\Big)^{-1}\Big(\frac{1}{\sqrt{n}}U_\beta^q(\beta_0, \widehat{\gamma}_q)\Big)\\ &= A_\beta^{-1}\frac{1}{\sqrt{n}Q}\sum_{q=1}^{Q}U_\beta^q(\beta_0, \widehat{\gamma}_q) + o_p(1).\end{aligned}$$

Use the Taylor series expansions of $U_\beta^q(\beta_0, \widehat{\gamma}_q)$ and $U_\gamma^q(\widehat{\gamma}_q)$ around γ_0, we have

$$\begin{aligned}U_\beta^q(\beta_0, \widehat{\gamma}_q) - U_\beta^q(\beta_0, \gamma_0) &= \frac{\partial U_\beta^q(\beta_0, \gamma^t)}{\partial\gamma^t}(\widehat{\gamma}_q - \gamma_0),\\ -U_\gamma^q(\widehat{\gamma}_0) = U_\gamma^q(\widehat{\gamma}_q) - U_\gamma^q(\gamma_0) &= \frac{\partial U_\gamma^q(\gamma^s)}{\partial\gamma^s}(\widehat{\gamma}_q - \gamma_0),\end{aligned}$$

where both γ^t and γ^s are on the line segment between $\widehat{\gamma}_q$ and γ_0. By the consistency of $\widehat{\gamma}_q$ and rewriting the above equations yields that $\sqrt{n}^{-1}U_\beta^q(\beta_0, \widehat{\gamma}_q)$ is equal to

$$\begin{aligned}&\frac{1}{\sqrt{n}}\Big\{U_\beta^q(\beta_0, \gamma_0) + \frac{\partial U_\beta^q(\beta_0, \gamma^t)}{\partial\gamma^t}(\widehat{\gamma}_q - \gamma_0)\Big\}\\ &= \frac{1}{\sqrt{n}}\Big\{U_\beta^q(\beta_0, \gamma_0) + \frac{1}{n}\frac{\partial U_\beta^q(\beta_0, \gamma^t)}{\partial\gamma^t}\Big(-\frac{1}{n}\frac{\partial U_\gamma^q(\gamma^s)}{\partial\gamma^s}\Big)^{-1}U_\gamma^q(\gamma_0)\Big\}\\ &= \frac{1}{\sqrt{n}}\{U_\beta^q(\beta_0, \gamma_0) + A_\gamma^q(B_\gamma^q)^{-1}U_\gamma^q(\gamma_0)\} + o_p(1)\\ &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{a_{1i}^q(\beta_0, \gamma_0) + a_{2i}^q(\beta_0, \gamma_0) + A_\gamma^q(B_\gamma^q)^{-1}(b_{1i}^q(\gamma_0) + b_{2i}^q(\gamma_0))\} + o_p(1),\end{aligned}$$

where

$$\begin{aligned}a_{1i}^q(\beta, \gamma) &= \int_0^\infty\Big(Z_i^{q*}(t) - \frac{s_{1,\beta,q}^{(1)}(t, \beta, \gamma)}{s_{1,\beta,q}^{(0)}(t, \beta, \gamma)}\Big)\mathrm{d}M_i^{(1)q}(t),\\ a_{2i}^q(\beta, \gamma) &= \int_0^\infty\Big(Z_i^{q*}(t) - \frac{s_{2,\beta,q}^{(1)}(t, \beta, \gamma)}{s_{2,\beta,q}^{(0)}(t, \beta, \gamma)}\Big)\mathrm{d}M_i^{(2)q}(t),\\ b_{1i}^q(\gamma) &= \int_0^\infty\Big(Z_i^{q*}(t) - \frac{s_{1,\gamma,q}^{(1)}(t, \gamma)}{s_{1,\gamma,q}^{(0)}(t, \gamma)}\Big)\mathrm{d}\widetilde{M}_i^{(1)q}(t),\\ b_{2i}^q(\gamma) &= \int_0^\infty\Big(Z_i^{q*}(t) - \frac{s_{2,\gamma,q}^{(1)}(t, \gamma)}{s_{2,\gamma,q}^{(0)}(t, \gamma)}\Big)\mathrm{d}\widetilde{M}_i^{(2)q}(t),\end{aligned}$$

A^q_γ and B^q_γ are limits of $\widehat{A}^q_\gamma(\beta,\gamma) = n^{-1}\partial U^q_\beta(\beta,\gamma)/\partial\gamma$ and $\widehat{B}^q_\gamma(\gamma) = -n^{-1}\partial U^q_\gamma(\gamma)/\partial\gamma$ at (β_0,γ_0). $s^{(j)}_{l,\beta,q}(t,\beta,\gamma)$ and $s^{(j)}_{l,\gamma,q}(t,\gamma)$ denote the limits of $S^{(j)}_{l,\beta,q}(t,\beta,\gamma)$ and $S^{(j)}_{l,\gamma,q}(t,\gamma)$, respectively, for $l = 1,2$ and $j = 0,1$. Note that

$$U^q_\beta(\beta_0,\gamma_0) = \sum_{i=1}^n \{a^q_{1i}(\beta_0,\gamma_0) + a^q_{2i}(\beta_0,\gamma_0)\} + o_p(1),$$
$$U^q_\gamma(\gamma_0) = \sum_{i=1}^n \{b^q_{1i}(\gamma_0) + b^q_{2i}(\gamma_0)\} + o_p(1),$$

It is easy to show that $(\sqrt{n}Q)^{-1}\sum_{q=1}^Q U^q_\beta(\beta_0,\widehat{\gamma}_q)$ converge to a normal distribution as $n\to\infty$, changing the order of summation as

$$\begin{aligned}
&\frac{1}{\sqrt{n}Q}\sum_{q=1}^Q U^q_\beta(\beta_0,\widehat{\gamma}_q)\\
&= \frac{1}{\sqrt{n}Q}\sum_{q=1}^Q \{U^q_\beta(\beta_0,\gamma_0) + A^q_\gamma (B^q_\gamma)^{-1} U^q_\gamma(\gamma_0)\} + o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{1}{Q}\sum_{q=1}^Q \{a^q_{1,i}(\beta_0,\gamma_0) + a^q_{2,i}(\beta_0,\gamma_0) + A^q_\gamma (B^q_\gamma)^{-1}(b^q_{1,i}(\gamma_0) + b^q_{2,i}(\gamma_0))\} + o_p(1)\\
&:= \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i(\beta_0,\gamma_0) + o_p(1),
\end{aligned}$$

where $U_i(\beta_0,\gamma_0)$, $i = 1,2,\ldots,n$ are independent with mean zero and finite variance. It thus follows from the multivariate Central Limit Theorem that $(\sqrt{n}Q)^{-1}\sum_{q=1}^Q U^q_\beta(\beta_0,\widehat{\gamma}_q)$ is asymptotically normal with zero mean. Combining with Slutsky's theorem, $\sqrt{n}(\widehat{\beta}_{\rm wcr} - \beta_0)$ converges in distribution to a zero-mean normal random vector and covariance matrix can be consistently estimated by $n\widehat{\Sigma}_{\rm wcr}$.

Wang et al.[2] showed that $\sqrt{n}(\widehat{\beta}_q - \beta_0)$ can be asymptotically approximated by a normal vector with mean zero and a covariance matrix that can be consistently estimated by $n\widehat{\Sigma}_q = (\widehat{A}^q_\beta)^{-1}\widehat{\Gamma}_q(\widehat{A}^q_\beta)^{-1}$, where

$$\widehat{\Gamma}_q = \frac{1}{n}\sum_{i=1}^n \widehat{\alpha}^q_i(\widehat{\beta}_q,\widehat{\gamma}_q)(\widehat{\alpha}^q_i(\widehat{\beta}_q,\widehat{\gamma}_q))'$$

with

$$\widehat{\alpha}^q_i(\widehat{\beta}_q,\widehat{\gamma}_q) = \widehat{a}^q_{1i}(\widehat{\beta}_q,\widehat{\gamma}_q) + \widehat{a}^q_{2i}(\widehat{\beta}_q,\widehat{\gamma}_q) + \widehat{A}^q_\gamma(\widehat{\beta}_q,\widehat{\gamma}_q)(\widehat{B}^q_\gamma(\widehat{\gamma}_q))^{-1}\{\widehat{b}^q_{1i}(\widehat{\gamma}_q) + \widehat{b}^q_{2i}(\widehat{\gamma}_q)\}.$$

For each resampled data, $\mathsf{Var}\,(\widehat{\beta}_q)$ can be consistently estimated by $\widehat{\Sigma}_q$. Average over the Q resamples, the resulting estimator denoted by $Q^{-1}\sum_{q=1}^Q \widehat{\Sigma}^q$ is also consistent. For the consistent estimator of the covariance matrix of $\widehat{\beta}_{\rm wcr}$, similar to [11], we first write

$$\mathsf{Var}\,(\widehat{\beta}_q) = \mathsf{E}(\mathsf{Var}\,(\widehat{\beta}_q \,|\, \text{data})) + \mathsf{Var}\,(\mathsf{E}(\widehat{\beta}_q \,|\, \text{data})).$$

By the fact of $\mathsf{E}(\widehat{\beta}_q \,|\, \text{data}) = \widehat{\beta}_{\text{wcr}}$, it yields that

$$\mathsf{Var}\,(\widehat{\beta}_{\text{wcr}}) = \mathsf{Var}\,(\widehat{\beta}_q) - \mathsf{E}(\mathsf{Var}\,(\widehat{\beta}_q \,|\, \text{data})).$$

Since

$$\mathsf{E}(\mathsf{Var}\,(\widehat{\beta}_q \,|\, \text{data})) = \mathsf{E}\Big(\frac{1}{Q}\sum_{q=1}^{Q}(\widehat{\beta}_q - \widehat{\beta}_{\text{wcr}})(\widehat{\beta}_q - \widehat{\beta}_{\text{wcr}})'\Big),$$

it can be estimated as the covariance matrix based on the Q resampling estimators $\widehat{\beta}_q$, $q = 1, 2, \ldots, Q$, that is

$$\Omega = \frac{1}{Q}\sum_{q=1}^{Q}(\widehat{\beta}_q - \widehat{\beta}_{\text{wcr}})(\widehat{\beta}_q - \widehat{\beta}_{\text{wcr}})'.$$

Thus the estimated variance-covariance matrix of $\widehat{\beta}_{\text{wcr}}$ is

$$\widehat{\Sigma}_{\text{wcr}} = \frac{1}{Q}\sum_{q=1}^{Q}\widehat{\Sigma}_q - \frac{1}{Q}\sum_{q=1}^{Q}(\widehat{\beta}_q - \widehat{\beta}_{\text{wcr}})(\widehat{\beta}_q - \widehat{\beta}_{\text{wcr}})'.$$

To show the consistency of $\widehat{\Sigma}_{\text{wcr}}$, it is easy to see that $\Omega - \mathsf{E}(\Omega) \to 0$ in probability as $n \to \infty$. Actually, this can be easily shown by applying the same arguments as those in the proof of [9]. This completes the proof. □

Acknowledgements The authors gratefully acknowledge the recommendations of the associate editor and the reviewers that led to an improved revision of an earlier manuscript.

References

[1] Zeng D L, Cai J W, Shen Y. Semiparametric additive risks model for interval-censored data [J]. *Statist. Sinica*, 2006, **16(1)**: 287–302.

[2] Wang L M, Sun J G, Tong X W. Regression analysis of case II interval-censored failure time data with the additive hazards model [J]. *Statist. Sinica*, 2010, **20(4)**: 1709–1723.

[3] Cox D R. Regression models and life-tables (with discussion) [J]. *J. Roy. Statist. Soc. Ser. B*, 1972, **34(2)**: 187–220.

[4] Li J L, Wang C J, Sun J G. Regression analysis of clustered interval-censored failure time data with the additive hazards model [J]. *J. Nonparametr. Stat.*, 2012, **24(4)**: 1041–1050.

[5] Li J L, Tong X W, Sun J G. Sieve estimation for the Cox model with clustered interval-censored failure time data [J]. *Statist. Biosci.*, 2014, **6(1)**: 55–72.

[6] Kor C T, Cheng K F, Chen Y H. A method for analyzing clustered interval-censored data based on Cox's model [J]. *Stat. Med.*, 2013, **32(5)**: 822–832.

[7] Lin D Y, Oakes D, Ying Z L. Additive hazards regression with current status data [J]. *Biometrika*, 1998, **85(2)**: 289–298.

[8] Kulich M, Lin D Y. Additive hazards regression for case-cohort studies [J]. *Biometrika*, 2000, **87(1)**: 73–87.

[9] Cong X J, Yin G S, Shen Y. Marginal analysis of correlated failure time data with informative cluster sizes [J]. *Biometrics*, 2007, **63(3)**: 663–672.

[10] Lin D Y, Ying Z L. Semiparametric analysis of the additive risk model [J]. *Biometrika*, 1994, **81(1)**: 61–71.

[11] Hoffman E B, Sen P K, Weinberg C R. Within-cluster resampling [J]. *Biometrika*, 2001, **88(4)**: 1121–1134.

加法风险模型下聚类区间删失数据的回归分析

邸荣荣 王成勇

(武汉大学数学与统计学院, 武汉, 430072) (湖北文理学院数学与计算机科学学院, 襄阳, 441053)

摘 要: 聚类区间删失失效时间常出现于医学研究中研究对象来自同一个类中的情形. 此外, 失效时间可能与类的大小相关. 由于缺乏直接分析所需的推演过程, 因此常见的简单方式就是简化区间删失数据. 鉴于此, 本文提出了类内重抽样方法来考虑加法风险模型下的 II 型区间删失问题. 类内重抽样的方法简单但需要大量计算, 这一方法的主要优势在于在类的大小相关时, 估计变量易于实现. 渐近性质和部分模拟结果的讨论验证了该方法的有效性.

关键词: 加法风险模型; 区间删失; 类内重抽样; 半参数回归

中图分类号: O213.9