行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　　　　重複事件之統計分析(2/2)　　　　　　　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：☑個別型計畫　　□整合型計畫

計畫編號：NSC　90－2118－M－002－012－

執行期間：90 年 8 月 1 日至 91 年 7 月 31 日

計畫主持人：張淑惠

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：公共衛生學院公共衛生系

中　華　民　國　91 年 10 月 30 日

行政院國家科學委員會專題研究計畫成果報告

國科會專題研究計畫成果報告撰寫格式

計畫編號：NSC 90-2118-M-002-012

執行期限：90 年 8 月 1 日至 91 年 7 月 31 日

主持人：張淑惠

執行機構及單位名稱：公共衛生學院公共衛生系

中文摘要

在長期觀察研究中，一個體多次事件的發生可表示成一是計數隨機過程。本研究考慮 intensity 函數與 hazard 函數之條件迴歸模式來形容重複事件的發生與相關的因素及過去事件歷史的關係，提出對迴歸參數與 baseline intensity 及 hazard 函數較有效之估計方式，並探討其相關之統計性質。第一年度之期中報告以提及以 intensity 函數之條件迴歸模式而言仍可藉由適當定義之 filtration 建立 partial likelihood 估計得到其迴歸參數之估計 且可利用 martingale central limit theory 得到迴歸參數與 common baseline intensity 估計量之大樣本性質。然而第二年度之成果報告將針對 hazard 函數之條件迴歸模式而言以發生事件順序為分層變數亦可建立分層 partial likelihood 得到迴歸參數之估計；然而可以不分層方式建立一 profile likelihood 可得到較有效之迴歸參數與 common hazard 函數估計世，且須經由 empirical count function 得到迴歸參數與 common baseline intensity 估計量之大樣本性質。最後用模擬方式說明其相關之統計性質與表現。

關鍵詞：計數隨機過程；部分概似函數；比例危險模式；縱斷概似函數

Abstract

The occurrence of the recurrent events for a subject can be represented as a realization of a stochastic counting process. Semiparametric hazards models for serial time intervals between successive events are considered to describe the occurrence of the repeated events relayed to the time-dependent and time-independent covariates as well as the event history. In application to accidental events such as needle-stick injury in hospitals and regular events such as woman's menstrual cycles, it may be natural to assume a common baseline hazard function for all episodes of events. The common baseline hazard model includes two types of effects, the global common and episode-specific effects. Two types of likelihoods can be constructed. The first one is the partial likelihood in which the order of event is the stratification variable. Secondly, the profile likelihood is established by non-stratified approach. This project develops the estimation of regression parameters and the common cumulative baseline hazard function based on these two likelihood structures. Examples will be conducted to illustrate the performance of the proposed estimation.

KEY WORDS: counting process; partial likelihood; profile likelihood; proportional hazards model.

1 INTRODUCTION

In many studies subjects may experience two or more events under observation. The events may be repetitions of essentially the same type or a series of recurrent events of different types. The project in the second year considers modeling the hazard function for serial durations between successive events conditioning on the past counting-process history. Most literatures (Prentice, Williams and Peterson, 1981; Chang and Hsiung, 1994; Chang and Wang, 1999; Lawless and Wigg, 2001) emphasis the development of the conditional hazards models with the distinct baseline hazard functions for different episodes of events. However, in application to the occurrences of accidental or regular events such as needle-stick injury in hospitals and woman's menstrual cycles, it may be natural to restrict the baseline hazard function to be common for each episode of events. As suggested in Chang and Wang (1999), this project considers the conditional hazard models involving both structural and episode-specific parameters and the baseline hazard function is assumed to be the same for each episode of events.

Define $T_0 = 0$. Let T_j be the random variable representing the occurrence time of the jth recurrent event and let $Y_j = T_j - T_{j-1}$ be the recurrence time of interest, $j = 1, 2, \ldots$. Let $z(u) = (z_1(u), z_2(u))$ be a vector of covariates at time u and $Z(t) = \{z(u) : u \leq t\}$ be the corresponding covariates history up to and including t. Define $N(t) = \{T_1, \ldots, T_{j-1} : T_1 < T_2 < \ldots < T_{j-1} < t \leq T_j\}$ to be the event history prior to the time t. This common baseline hazard model including both the structural and episode-specific parameters can be written as

$$\lambda(t|N(t), Z(t)) = \lambda_0(t - t_{j-1}) exp\{z_1(t)\beta + z_2(t)\gamma_j\}, \tag{1}$$

for $t > t_{j-1}$, where $j - 1 = \max\{\ell : T_\ell < t, \ell = 1, 2, \ldots\}$ is the number of recurrent events prior to time t.

Suppose that only the structural parameters are considered in the conditional hazard models, that is,

$$\lambda(t|N(t), Z(t)) = \lambda_0(t - t_{j-1}) exp\{z_1(t)\beta\}. \tag{2}$$

A few special cases of model (2) are of interest. First, if only consider the time-independent covariate in the model, the corresponding process is a renewal process with independent time intervals for all episodes of events. The statistical inference and estimation of β are based on the same approach as those used with univariate survival data. A second special case of model (2) of interest is the first-order Markov model or modulated renewal processes, where previous time intervals are time-dependent covariates in the model. The full likelihood could be decomposed into two parts. One part has exactly the same form as the partial likelihood derived from the renewal processes, but it does not have a partial likelihood interpretation as discussed in Oakes and Cui (1994) and Chang (1995). It is noted that the asymptotic distributions for the estimators of the regression parameters and the cumulative hazard in model (2) for single renewal sequences have been developed by Oakes and Cui (1994). However, Oakes and Cui (1994) established the asymptotic properties based on the large number of recurrences for a single modulated renewal process. In medical studies the collected data frequently involve large number of (subjects) sequences with finite recurrences during a time period. In addition, the stratified partial likelihood estimating method for β considered in Chang and Wang (1999) is a legitimate estimating method for β in models (1) and (2), where the order of episodes of events serves as the stratification variable. But the stratified procedure may not be efficient, especially when the censoring is heavy. The goal of this project is to develop a more efficient estimating method for β, γ_j's, and the common baseline cumulative hazard function in model (1).

In Section 2, we study the likelihood structure based on Model (1). The estimation of the regression parameters and cumulative baseline hazard are considered in section 3. The statistical properties of the parameter estimators in Model (1) are also discussed in Section 3. A simulation study is presented in Section 4 to illustrate the performance of the estimators.

2 LIKELIHOODS

Suppose that there are n subjects under observation. Let $\{N_i(\cdot), Z_i(\cdot), C_i\}$ be the event process, covariates and censoring time for subject i. Assume that the $\{N_i(\cdot), Z_i(\cdot), C_i\}$, $i = 1, 2, \ldots, n$, are independent and identically distributed. Let $K_i(t)$ be the number of events occurring in the interval $[0, t)$. Let T_{ij} be the jth event time for subject i, and $Y_{ij} = T_{ij} - T_{i,j-1}$, $i = 1, 2, \ldots, n$, $j \geq 1$. The observed recurrence times are $\{y_{i1}, \ldots, y_{i,k_i(c_i)}, y^+_{i,k_i(c_i)+1}\}$, where c_i is the realization of the random censoring time C_i and $y^+_{i,k_i(c_i)+1} = c_i - t_{i,k_i(c_i)}$ is the time from the last observed event to the censoring time. Further define $x_{ij} = y_{ij}$ for $j = 1, 2, \ldots, k_i(c_i)$ and $x_{i,k(c_i)+1} = y^+_{i,k_i(c_i)+1}$ for later discussion. Let $K = max_{1 \leq i \leq n}\{K_i(C_i)\}$.

Define $y_{(1),j} < y_{(2),j} < \ldots < y_{(d_j),j}$ to be the d_j ordered distinct recurrence times for the jth recurrent event. Let $Z_{(i),j}$ denote the covariates history up to and including $t_{(i),j}$, the ordered event time corresponding to $y_{(i),j}$. With the help of the above notation and under the independent censoring assumption,

$$\lambda(t \mid N_i(t), Z_i(t), C_i > t) = \lambda(t \mid N_i(t), Z_i(t)),$$

the full likelihood for model (1) can be expressed as,

$$\mathcal{L}(\beta; \gamma_j, j \geq 1; \lambda_0(\cdot)) = \prod_{j \geq 1} \prod_{i=1}^{d_j} \lambda_0(y_{ij}) e^{\{z_1(t_{(i),j})\beta + z_2(t_{(i),j})\gamma_j\}} e^{\{-\int_{t_{(i),j-1}}^{t_{(i),j}} \lambda_0(u - t_{(i),j-1}) e^{\{z_1(u)\beta + z_2(u)\gamma_j\}} du\}}. \quad (3)$$

2.1 Profile Likelihood

As discussed in section 1, we can not construct a partial likelihood which is a product of a series of conditional probabilities. As suggested by Breslow (1974), substituting the Breslow estimator of the baseline cumulative hazard function, $\Lambda_0(t)$,

$$\hat{\Lambda}_0(t) = \sum_{m \geq 1} \sum_{y_{(i),m} \leq t} \frac{1}{\sum_{j \geq 1} \sum_{\ell \in R_j(y_{(i),m})} exp\{z_{\ell 1}(t_{\ell,j-1} + y_{(i),m})\beta + z_{\ell 2}(t_{\ell,j-1} + y_{(i),m})\gamma_j\}},$$

in $\mathcal{L}(\beta; \gamma_j, j \geq 1; \lambda_0(\cdot))$ and simplifying yields a profile likelihood proportional to

$$\mathcal{L}_{p1}(\beta; \gamma_j, j \geq 1) =$$
$$\prod_{j\geq 1}\prod_{i=1}^{d_j} \frac{exp\{z_{(i),1}(t_{(i),j})\beta + z_{(i),2}(t_{(i),j})\gamma_j\}}{\sum_{m\geq 1}\sum_{\{\ell\in R_m(y_{(i),j})\}} exp\{z_{\ell 1}(t_{\ell,m-1} + y_{(i),j})\beta + z_{\ell 2}(t_{\ell,m-1} + y_{(i),j})\gamma_m\}}, \quad (4)$$

where $R_j(u)$ is the risk set defined at u for the jth recurrent event among those who have had $j-1$ recurrent events – namely, $R_m(y_{(i),j}) = \{\ell : x_{\ell m} \geq y_{(i),j}$ and $k_\ell(t_{\ell,m-1}+y_{(i),j}) \geq m-1\}$.

2.2 Partial Likelihood

To construct the partial likelihood for Model (1), we first define the "complete history" of the first j recurrent events up to and including time $y_{(i),j}$,

$$\begin{aligned}\mathcal{H}_j(y_{(i),j}) &= \{\{(Z_{(1),1}, y_{(1),1}), \ldots, (Z_{(d_1),1}, y_{(d_1),1})\}; \ldots;\\ &\quad \{(Z_{(1),j-1}, y_{(1),j-1}), \ldots, (Z_{(d_{j-1}),j-1}, y_{(d_{j-1}),j-1})\};\\ &\quad \{(Z_{(1),j}, y_{(1),j}), \ldots, (Z_{(i),j}, y_{(i),j})\}\},\end{aligned}$$

for $i \leq d_j$ and $j \leq K$. The complete history, $\mathcal{H}_j(y_{(i),j})$, consists of full information of counting processes and covariates for the first $j-1$ recurrent events, and information of counting processes and covariates up to and including $y_{(i),j}$ for the jth recurrent events.

Thus, the partial likelihood, $\mathcal{L}_{p2}$, can be simply expressed as

$$\prod_{j\geq 1}\prod_{i=1}^{d_j} \frac{exp\{z_{(i),1}(t_{(i),j})\beta + z_{(i),2}(t_{(i),j})\gamma_j\}}{\sum_{\{\ell\in R_j(y_{(i),j})\}} exp\{z_{\ell 1}(t_{\ell,j-1} + y_{(i),j})\beta + z_{\ell 2}(t_{\ell,j-1} + y_{(i),j})\gamma_j\}}.$$

The partial likelihood is the same as derived in Chang and Wang (1999) by the model with various baseline hazards for different episodes of events.

3 ESTIMATION

We first introduce assumptions and notation that facilitate the development of the asymptotic properties. Suppose that a longitudinal study is conducted over a finite time interval $[0, c_0]$ so that all the censoring times are bounded, $C_i \leq c_0 < \infty$.

First, consider a simpler case where model (1) involves only finite episode-specific parameters γ_j, $j = 1, \ldots, J$. The following assumptions are considered: **(A)** $Z_i(t) = \{Z_{i1}(t), Z_{i2}(t)\}$ are bounded uniformly in t for each i; **(B)** $\int_0^{c_0} \lambda_0(t)dt < \infty$; **(C)** $\lambda_0(t) > 0$

for $t \in (0, c_0]$; **(D)** $inf_j \Pr(C_i \geq c_0 | K_i(C_i) \geq j-1) > 0$; **(E)** $\lambda(t|N_i(t), Z_i(t)) = 0$ if $K_i(t) \geq J$ for an integer $J \geq 1$.

The assumption (E) with assumptions (A) and (B) implies that $K_i(c_0)$ $(i = 1, \ldots, n)$ are bounded by a constant J. From assumptions (A), (C) and (E), it can be shown that $\Pr(K_i(C_i) \geq J) \geq 0$, which ensures that the number of the observed Jth episodes increases as the sample size n increases.

Let $\hat{\beta}_{p1}$ be the estimator of β derived from maximizing the profile likelihood $\mathcal{L}_{p1}$, and $\hat{\beta}_{p2}$ is the estimator of β based on the partial likelihood $\mathcal{L}_{p2}$. The asymptotic properties for $\hat{\beta}_{p2}$ and $\hat{\gamma}_{j,p2}$'s are the same as discussed in Wang and Chang (1999).

We first consider Model (2), a special case of (1). Without γ_j (i.e. all $\gamma_j = 0$), the profile score functions for β and γ_j are

$$U^{p1}_{n,\beta}(\beta, \gamma) = \frac{1}{\sqrt{n}} \sum_{j \geq 1} \sum_{i=1}^{n} \int \left[z_{i1}(\tilde{t}_{ij}) - \frac{S^{(1)}_{1n}(\beta, \gamma; y)}{S^{(0)}_{1n}(\beta, \gamma; y)} \right] dM_{ij}(y), \tag{5}$$

$$U^{p1}_{n,\gamma_j}(\beta, \gamma) = \frac{1}{\sqrt{n}} \sum_{j \geq 1} \sum_{i=1}^{n} \int \left[z_{i2}(\tilde{t}_{ij}) I(K_i(\tilde{t}_{ij}) \geq j-1) - \frac{S^{(1)}_{2nj}(\beta, \gamma; y)}{S^{(0)}_{2nj}(\beta, \gamma; y)} \right] dM_{ij}(y), \tag{6}$$

where $\tilde{t}_{ij} = y + t_{i,j-1}$, $dM_{ij}(y) = dN_{ij}(y) - I(X_{ij} \geq y, K_i(\tilde{t}_{ij}) \geq j-1)\lambda_0(y) exp\{z_{i1}(\tilde{t}_{ij})\beta + z_{i2}(\tilde{t}_{ij})\gamma_j\}$, $N_{ij}(y) = I(X_{ij} \leq y, Y_{ij} = X_{ij}, K_i(\tilde{t}_{ij}) \geq j-1)$;
$S^{(v)}_{1n}(\beta, \gamma; y) = \frac{1}{n} \sum_{m \geq 1} \sum_{\ell \in R_m(y)} z_{\ell 1}(\tilde{t}_{\ell m})^{\otimes} exp\{z_{\ell 1}(\tilde{t}_{\ell m})\beta + z_{\ell 2}(\tilde{t}_{\ell m})\gamma_m\}$, and $S^{(v)}_{2nj}(\beta, \gamma; y) = \frac{1}{n} \sum_{m \geq 1} \sum_{\ell \in R_m(y)} z_{\ell 2}(\tilde{t}_{\ell m})^{\otimes} I(K_\ell(\tilde{t}_{\ell j}) \geq j-1) exp\{z_{\ell 1}(\tilde{t}_{\ell m})\beta + z_{\ell 2}(\tilde{t}_{\ell m})\gamma_j\}$ for $v = 0, 1, 2$. The assumptions (A) and (B) imply that the condition $E(K_i(c_0)) < \infty$ and then

$$E(S^{(v)}_{1n}(\beta, \gamma; y)) \leq \mathbf{b_1} \frac{1}{n} \sum_{\ell=1}^{n} E\Big(\sum_{m=1}^{K_\ell(C_\ell)} I(Y_{\ell m} \geq y) \Big) \tag{7}$$

$$\leq \mathbf{b_1} \frac{1}{n} \sum_{\ell=1}^{n} E(K_\ell(C_\ell)) \leq \mathbf{b_1} J < \infty, \tag{8}$$

where $z^{\otimes v}_{\ell f} exp\{z_{\ell 1}(\tilde{t}_{\ell m})\beta + z_{\ell 2}(\tilde{t}_{\ell m})\gamma_m\}$ is bounded uniformly by a constant $\mathbf{b_1}$ from assumption (A). Similarly, $E(S^{(v)}_{2nj}(\beta, \gamma; y)) < \infty$ and the second moments of $S^{(v)}_{1n}(\beta, \gamma; y)$ and $S^{(v)}_{2nj}(\beta, \gamma; y)$ are also uniformly bounded from assumption (E). Define $s^{(v)}_1(\beta, \gamma, y) = E(S^{(v)}_{1n}(\beta, \gamma; y))$ and $s^{(v)}_{2j}(\beta, \gamma, y) = E(S^{(v)}_{2nj}(\beta, \gamma; y))$. From the above discussion, $S^{(v)}_{1n}(\beta, \gamma; y)$ and $S^{(v)}_{2nj}(\beta, \gamma; y)$ are convergent in probability to $s^{(v)}_1(\beta, \gamma, y)$ and $s^{(v)}_{2j}(\beta, \gamma, y)$, respectively. Therefore, one has the following approximations

$$U^{p1}_{n,\beta}(\beta, \gamma) = \frac{1}{\sqrt{n}} \sum_{j \geq 1} \sum_{i=1}^{n} \int \left[z_{i1}(\tilde{t}_{ij}) - \frac{s^{(1)}_{1n}(\beta, \gamma; y)}{s^{(0)}_{1n}(\beta, \gamma; y)} \right] dM_{ij}(y) + o_p(1), \text{ and} \tag{9}$$

$$U^{p1}_{n,\gamma_j}(\beta, \gamma) = \frac{1}{\sqrt{n}} \sum_{j \geq 1} \sum_{i=1}^{n} \int \left[z_{i2}(\tilde{t}_{ij}) I(K_i(\tilde{t}_{ij}) \geq j-1) - \frac{s^{(1)}_{2nj}(\beta, \gamma; y)}{s^{(0)}_{2nj}(\beta, \gamma; y)} \right] dM_{ij}(y) + o_p(1) \tag{10}$$

It is noted that the above two profile score functions can be viewed as a generalized version of the partial score function from univariate survival data. Using the history $\mathcal{H}_j(\cdot)$ as defined in section 2.1, asymptotical normality of $\hat{\beta}_{p1}$ and $\hat{\gamma}_j$ (j=1, ..., J) can be developed by the counting processes and martingale techniques in survival analysis and the corresponding variance-covariance matrix of $\hat{\beta}_{p1}$ and $\hat{\gamma}_j$ (j=1, ..., J) is estimated by the second derivative of the logarithm of the profile likelihood $\mathcal{L}_{p1}$.

Now consider the general model defined in Model (1) involving infinite episode-specific parameters ($\{\gamma_j\}$) in which the number of $\{\gamma_j\}$ increases with the sample size n. That is, the estimation of β will be established without assumption (E). As discussed in Chang and Wang (1999), $\hat{\gamma}_{j,p1}$ for large $j \geq 1$ may not be consistent since assumption (C) does not guarantee $\Pr(K_i(C_i) \geq j) > 0$ as $j \to \infty$. In order to avoid the effect of the inconsistent estimation of γ_j's on the estimation of β, consider a substratified profile score function for β which consists of two parts: the first part is the profile score function using the first J recurrences provided that $\Pr(K_i(C_i) \geq J) > 0$ and the second part is the stratified partial score function (considered by Chang and Wang, 1999) using the rest of the data.

4 SIMULATION

In the previous sections, we have shown the asymptotic normality of $\hat{\beta}$ and the $\{\hat{\gamma}_{j,p1}\}$ for the special case with finite recurrent events. When model (1) involving infinite γ_j's, three-type estimating procedures of β are discussed in the previous section. A simulation study is conducted to illustrate the performance of estimators of β and γ_j's with three different estimating procedures. A sample with size=50 is generated from the Weibull distribution with the conditional hazard function of the jth event time, $\lambda_{0j}(t_j - t_{j-1})exp\{-\beta z_1 - \gamma_j z_2)\}$, where $\lambda_{0j}(t_j - t_{j-1}) = (2\alpha)(t_j - t_{j-1})$, $z_1 \sim Bin(1, 0.5)$ and $z_2 \sim Bin(1, 0.7)$. The true parameter values are $\beta = 1$; $\gamma_1 = -2$, $\gamma_2 = -2$, $\gamma_3 = -3$, $\gamma_4 = -4$, $\gamma_5 = -5$ and $\gamma_j = -1$ for $j > 5$. The independent censoring times are generated from the uniform distribution on $[0, c_0]$, with $c_0 = 4$. The average number of events for a subject is 2.5. Tables 1 gives the simulation results based on 500 replicates of samples generated from the above simulation procedures.

The display in Table 1 includes the mean estimates of β and γ_j's and the Monte Carlo variance estimates (var) for the corresponding $\hat{\beta}$ and $\hat{\gamma}_j$'s. From Table 1, we can see that the MSEs of $\hat{\beta}_{p1}$ based on the profile likelihood and substratified method are less than that based on the partial likelihood.

Table 1: Simulation with Sample Size=50

Parameter	Profile estimate(mse)	Partial estimate(mse)	Substratified(k=4) estimate (mse)
β	1.09 (0.108)	1.096 (0.14)	1.083 (-0.108)
γ_1	-2.04 (0.348)	-2.129 (0.665)	-2.052 (0.365)
γ_2	-2.06 (0.474)	-2.108 (1.337)	-2.086 (0.519)
γ_3	-3.19 (0.993)	-7.506 (—)	-3.214 (1.07)
γ_4	-5.20 (153.925)	-81.651 (—)	-5.19 (153.908)
γ_5	-5.98 (518.434)	-76.714 (—)	-77.15 (—)
γ_6	88.56 (—)	10.026 (—)	9.401 (—)
γ_7	-83.42 (—)	-8.406 (—)	-8.592 (—)
γ_8	21.96 (—)	10.279 (—)	9.777 (—)
γ_9	14.45 (—)	-272.471 (—)	-272.496 (—)
γ_{10}	44.89 (—)	102.884 (—)	102.481 (—)

5 Conclusion

This project considers the extended semi-Markov model with the structural and episode-septic parameters. The estimation of the structural and episode-specific parameters and the cumulative hazard can be obtained based on the profile likelihood when the number of events is bounded. In addition, the corresponding large sample properties can be established as the same as the traditional univariate survival data through the empirical count processes and the martingale representation. When the number of events may be infinite under the condition that the expected number of events is finite, a substratification estimating approach for estimating β can be developed to handle the inconsistent estimates of γ_j.

REFERENCES

Dabrowska, D.M., Sun, G-W, and Horowitz (1994) , "Cox regression in a Markov renewal model: an application to the analysis of bone marrow transplant data, " *JASA* **89**, 867-877.

Lawless, J.F. and Wigg, M.B. (2001), "Analysis of repeated failures or durations, with application to shunt failures for patients with paediatrics hydrocephalus" *Appl.Statist.*, **50**, 449-465.

Oakes,D. and Cui, L. (1994) "On semiparametric inference for modulated renewal processes", *Biometrika*, **81**, 83-90.