University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2002 *Paper* 112

Semiparametric Regression Analysis on Longitudinal Pattern of Recurrent Gap Times

Ying Qing Chen* Mei-Cheng Wang† Yijian Huang‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, yqchen@stat.berkeley.edu

†Department of Biostatistics, School of Hygiene & Public Health, Johns Hopkins University

‡Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle

http://biostats.bepress.com/ucbbiostat/paper112

Semiparametric Regression Analysis on Longitudinal Pattern of Recurrent Gap Times

Ying Qing Chen, Mei-Cheng Wang, and Yijian Huang

Abstract

In longitudinal studies, individual subjects may experience recurrent events of the same type over a relatively long period of time. The longitudinal pattern of the gaps between the successive recurrent events is often of great research interest. In this article, the probability structure of the recurrent gap times is first explored in the presence of censoring. According to the discovered structure, we introduce the proportional reverse-time hazards models with unspecified baseline functions to accommodate heterogeneous individual underlying distributions, when the ongitudinal pattern parameter is of main interest. Inference procedures are proposed and studied by way of proper riskset construction. The proposed methodology is demonstrated by Monte-Carlo simulations and an application to the well-known Denmark schizophrenia cohort study data set

1 Introduction

For the subjects in a longitudinal follow-up study, individual subject may experience successive events of same type over a relatively long period of time, for example, recurrent superficial tumors of cancer patients, recurrent hospitalizations of schizophrenic patients, and recurrent seizures of pediatric cerebral malaria patients. When the events are considered as points occurring along the time axis, they form point processes.

As noted in Cox & Isham (1980, p.11), there are three equivalent perspectives to study the point processes: (a) the intensity perspective (the complete intensity function of occurrences), (b) the counting perspective (the joint distribution of the occurrence counts in any arbitrary sets), and (c) the gap perspective (the joint distribution of gaps between successive events). Perspectives (a) and (b) are relatively convenient to study in general theory development. Nevertheless, the gap perspective is of important scientific interest as well. For example, in Eaton, et al. (1992), the distributional pattern of the gaps between the successive hospitalizations serves as an important index of the schizophrenic disease progression: do the gaps progress longer and longer (progressive amelioration), or shorter and shorter (progressive deterioration), over the time? if there seems to be such a progressive pattern, can it be tested and the magnitude be estimated?

The goal of this article is to develop statistical approaches to tackle the scientific questions such as the ones above. The main interest is in the longitudinal pattern of the gap times. However, the study subjects are often heterogeneous in their underlying distributions. For example, it is usually not practical to assume the underlying homogeneity of 8,811 patients in 86 psychiatric institutions across the entire nation of Denmark (Eaton, et al., 1992). Our specific approach is thus to define and estimate the longitudinal pattern parameters through regression models, when the individual underlying distributions are considered as nuisance.

Toward this end, we risk ourselves with fast growing number of nuisance parameters when the sample size increases, and therefore encounter the classical problem of Neyman-Scott type. More seriously, when censoring is present, the well-known "induced dependent censorship" further complicates the statistical modeling of recurrent event data and may lead to bias with naive application of the traditional survival analysis techniques (Gelber, Gelman & Goldhirsch, 1989; Huang, 1999; Lin, Sun & Ying, 1999). Recently, researchers have been developing various statistical methodologies trying to overcome the potential bias. A good summary of recent research development can be found in Cook & Lawless (2002).

Because of the longitudinal nature and the induced dependent censorship, the recurrent event data have distinctive features of their own. In §2.1, we first explore some of the features in probability structure of the observed gap times. In §2.2, the semiparametric regression models are introduced according to the discovered structure. In §**??**, we develop model estimation procedures by way of proper riskset construction. Numerical analyses including

some Monte-Carlo simulations and an application to the Denmark psychiatric registry data are in §4. Several remaining issues are discussed in §5. The technical proofs are collected in the Appendix.

2 Semiparametric Regression Models

2.1 *Probability structure*

Suppose there are n independent subjects in a study. Denote $i = 1, 2, \ldots, n$ the subject index, and $j = 0, 1, 2, \ldots$ the recurrent event index. For the ith subject, let T_{ij} be the gap time between the $(j-1)$st and the jth recurrent events, where $T_{i,0} = 0$, and C_i be the censoring time. It is assumed that, given any specific participant i, $i = 1, 2, \ldots, n$, $(C_i, T_{i1}, T_{i2}, \ldots)$ are independent. More discussion on the assumptions will be in later sections.

Suppose that $(t_{i0}, t_{i1}, \ldots, t_{i,m_i-1}, t^+_{i,m_i})$ is an observed sequence of the gap times, $i = 1, 2, \ldots, n$, where $M_i = m_i$ is the event index of stopping time such that

$$\sum_{k=1}^{m_i-1} t_{ij} \leqslant c_i \text{ and } \sum_{k=1}^{m_i} t_{ij} > c_i$$

for censoring time $C_i = c_i$, and $t^+_{i,m_i} = c_i - \sum_{j=0}^{m_i-1} t_{ij}$. The first $M_i - 1$ gap times are considered as "complete" duration times, while the last gap time T_{i,M_i} is always "censored." To simplify our discussion, we further assume that the underlying $(T_{i1}, T_{i2}, \ldots)$ that generate $(t_{i1}, \ldots, t_{i,m_i-1}, t^+_{i,m_i})$ are also identically distributed.

For any fixed index $j \geqslant 1$, the shorter gap time T_{ij} is more likely to be observed as complete t_{ij}, given C_i and $(T_{i1}, \ldots, T_{i,j-1})$. In addition, although $(T_{i1}, T_{i2}, \ldots)$ are identically distributed, the observed complete gap times t_{ij}, $1 \leqslant j < m_i$ tend to be shorter as j increases. To see this, let $W_{ij} = C_i - \sum_{k=1}^{j-1} t_{ik}$, the censoring time of the jth gap time T_{ij} given $(T_{i1}, \ldots, T_{i,j-1})$. Then the complete t_{ij} is observed from the conditional distribution of T_{i1} given $T_{i1} \leqslant W_{ij}$, and hence right-truncated (Lagakos, Barraj & De Gruttola, 1988, Kalbfleisch & Lawless, 1989). As a result, larger j leads to smaller W_{ij} and hence shorter complete t_{ij}.

However, the last gap time of T_{i,M_i} is observed subject to intercept sampling (Vardi, 1982). The backward recurrence time W_{i,M_i} (Cox, 1962, p. 61) serves as its left truncation time. That is, given $W_{i,M_i} = w$, T_{i,M_i} is in fact sampled from the conditional distribution of T_{i1} given $T_{i1} \geqslant w$. Although T_{i,M_i} is subject to the truncation of different direction from the observed complete gaps, it is not counterbalanced by simply pooling together the gap times of all the subjects (Wang & Chang, 1999).

2.2 *Proportional reverse-time hazards models*

As discussed in §2.1, the complete gaps are always subject to right-truncation. In fact, researchers have extended the traditional survival techniques in reverse time to the right-truncated failure times (Lagakos, Barraj & De Gruttola, 1988; Kalbfleisch & Lawless,1991; Gross & Huber-Carol, 1992). Denote the cumulative distribution function $F_{ij}(t) = \Pr\{T_{ij} \leqslant t\}$ and its corresponding reverse-time hazard function

$$\kappa_{ij}(t) = \lim_{\Delta t \to 0+} \frac{\Pr\{t - \Delta t \leqslant T_{ij} < t | T_{ij} \leqslant t\}}{\Delta t} = \frac{d \log F_{ij}(t)}{dt},$$

for the (i, j)th gap. Then the natural extension of the Cox proportional hazards model to the right-truncated failure times is the proportional reverse-time hazards model, as recommended in Kalbfleisch & Lawless (1991) and Gross & Huber-Carol (1992):

$$\kappa(t|Z_{ij}) = \kappa_{i0}(t) \exp(\beta^{\mathrm{T}} Z_{ij}), \tag{1}$$

where Z_{ij} is p-dimensional covariate and $\beta \in \mathcal{B} \subset \mathbf{R}^{\mathrm{p}}$ is parameter for $i = 1, \ldots, n$ and $j = 1, 2, \ldots$. In fact, if the negative time scale were allowed, e.g., let $T^r_{ij} = -T_{ij}$, then T^r_{ij} would be in theory to follow the usual proportional hazards model with identical regression coefficients in model (1):

$$\lambda_{ij}(t|Z_{ij}) = \lambda_{i0}(t) \exp(\beta' Z_{ij}),$$

where $\lambda_{ij}(t) = \kappa_{ij}(-t)$, for $t \leqslant 0$.

In (1), $\{\kappa_{i0}(t); t \geqslant 0, i = 1, 2, \ldots, n\}$ are unspecified and hence the models are semiparametric. This is similar to the proportional hazards models or the log-linear models proposed for the paired failure times in Kalbfleisch & Prentice (1980, p. 190). As pointed out by one reviewer, when all the subjects are believed to share similar distributions of baseline characteristics, $\{\kappa_{i0}(t); t \geqslant 0, i = 1, 2, \ldots, n\}$ can be further modeled in a standard fashion,

$$\kappa_{i0}(t) = \alpha_i \kappa_0(t),$$

for $i = 1, 2, \ldots, n$, where $\kappa_0(t)$ is unspecified and $\alpha_1, \alpha_2, \ldots, \alpha_n$ are random effects of some parametric distribution function, as in Aalen & Husebye (1991). When the study population is highly heterogeneous, however, it may be more feasible by treating $\{\kappa_{i0}(t); t \geqslant 0, i = 1, 2, \ldots, n\}$ as nuisance parameters if its longitudinal pattern remains the major interest.

Parameter β in (1) serves as the longitudinal pattern parameter of recurrent gap times. Its interpretation is better reflected in an equivalent form of (1):

$$F(t|Z_{ij}) = F_{i0}(t)^{\exp(\beta^{\mathrm{T}} Z_{ij})}. \tag{2}$$

For example, when Z_{ij} is univariate and increases with j, β represents an assigned trend measure over the longitudinal course of the gap times (Abelson & Tukey, 1963). However,

the identifiability of β is in doubt if the longitudinal pattern is of minor interest, for an extreme example, Z_{ij}'s are always constant from gap to gap within every subject. This is usually not an issue for the scientific questions concerning the longitudinal pattern, such as the ones in §1, though, because certain distinguishable gap indicator(s) or time-dependent covariates will be included into the models naturally.

3 Inference Procedures

3.1 *Biased risksets*

The concept of riskset was used to develop proper inference procedures in analysis of left-truncated failure times (Woodroofe, 1985; Wang, Jewell & Tsai, 1986). Brookmeyer & Gail (1994, p. 89) extended the same concept in reverse-time to right-truncated failure times. A proper riskset is supposed to contain a random sample at risk in order to construct the parameter estimators with sound statistical properties, otherwise it is called "biased." We first consider the usual way of riskset construction for the complete gap times as right-truncated observations.

According to the definition in Brookmeyer & Gail (1994), the individuals in the riskset at the observed t_{ij}, are those "whose truncation times [w_{ik}] are greater than or equal to" t_{ij} and "whose incubation periods [i.e., observed failure times, t_{ik}] are less than or equal to" t_{ij}. That is, the seemly proper riskset at t_{ij} is

$$R_{ij} = \{k : t_{ik} \leqslant t_{ij} \leqslant w_{ik}, k = 1, \ldots, m_i - 1\}, \tag{3}$$

where $w_{ik} = c_i - \sum_{l=1}^{k-1} t_{il}$ as defined in §2.1. This leads to the partial likelihood function in reverse-time as

$$PL_i = \prod_{j=1}^{m_i-1} \frac{\exp(\beta^{\mathrm{T}} Z_{ij})}{\sum_{k \in R_{ij}} \exp(\beta^{\mathrm{T}} Z_{ik})},$$

assuming that t_{ij}'s are distinct complete gap times. If R_{ij} were proper, the members in R_{ij} would form a random sample, and each one of them would have fair probability to fail at t_{ij}. Therefore, the score function that is the derivative of $\log(PL_i)$,

$$S_i(\beta) = \sum_{j=1}^{m_i-1} \left\{ Z_{ij} - \frac{\sum_{k \in R_{ij}} Z_{ik} \exp(\beta^{\mathrm{T}} Z_{ik})}{\sum_{k \in R_{ij}} \exp(\beta^{\mathrm{T}} Z_{ik})} \right\},$$

would be zero unbiased.

However, complication arises with R_{ij}. For any specific k in R_{ij}, although t_{ik} is observed independently of w_{ik}, t_{ij} does have impact on $w_{ik} = c_i - \sum_{l<k} t_{il} = c_i - (t_{i1} + \cdots + t_{ij} + \cdots +$

$t_{i,k-1}$) for any $j < k$, i.e., increases (decreases) in t_{ij} cause decreases (increases) in w_{ik}. As a result, the occurrence ordering of t_{ik} relative to t_{ij} alone may determine its inclusion in R_{ij}, regardless of model (1). Because of this inherent longitudinal nature, R_{ij} are not random samples at risk and hence biased. The estimators constructed by $S_i(\beta)$ of the biased risksets are no longer guaranteed with the sound statistical properties as basic as consistency.

3.2 *Unbiased reduced risksets*

As discussed in §**??**, the cause of biased risksets is clear, i.e., the gaps in R_{ij} do not have fair probabilities to fail at t_{ij}. An ideal treatment of correction is to allow fair probability for $t_{ik} \in R_{ij}$ to fail at t_{ij}. That is, replacing every gap t_{ik} with t_{ij} in R_{ij},

$$|R_{ij}|t_{ij} + \sum_{l \in R_{ij}^c} t_{il} \leqslant c_i \tag{4}$$

still holds, where $|R_{ij}|$ is the size of R_{ij} and $R_{ij}^c = \{1, 2, \ldots, m_i - 1\} \setminus R_{ij}$. Then unbiased estimating functions can be constructed based on the risksets satisfying (**??**). However, this way of construction is expected to be cumbersome because (**??**) needs to be verified for at most $2^{m_i - 1}$ times to obtain the maximal R_{ij}.

More feasible approaches can be considered by limiting the number of replacements of t_{ik} in R_{ij}. This can be achieved by reducing $|R_{ij}|$. The most aggressive reduction is to include only one gap $t_{ik} \in R_{ij}$, say, at a time, in addition to t_{ij} itself. That is, a meaningful reduced riskset $\tilde{R}_{ij}$ at t_{ij} would always have two gaps, t_{ij} and t_{ik}. Similarly as discussed in §**??**, however, not every t_{ik} is eligible for $\tilde{R}_{ij}$ to be unbiased. To explore the eligibility, we plot two possible cases in Figure 1 that may appear in reality: (a) $k > j$, $\tilde{R}_{ij}$ is to include a later gap; (b) $k < j$, $\tilde{R}_{ij}$ is to include an earlier gap, respectively.

In Figure 1(a), because t_{ik} occurs later than t_{ij}, $w_{ik} \leqslant w_{ij}$ by default. To allow $t_{ik} \in \tilde{R}_{ij}$ with fair probability to fail at t_{ij}, the usual condition for riskset construction of right-truncated observations applies. That is, $t_{ik} \leqslant t_{ij} \leqslant w_{ik}$. In Figure 1(b), however, because t_{ik} occurs sooner than t_{ij}, $w_{ik} \geqslant w_{ij}$ by default. Although t_{ik} may not be bigger than t_{ij}, it does have the positive probability of being greater than w_{ij}, which t_{ij} will never have. We need to curtail w_{ik} to eliminate such an excessive probability. In fact, due to the right-truncation, the largest room left for t_{ij} to probably grow is $w_{ij} - t_{ij}$. This is also supposed to be the largest room left for t_{ik} to fairly fail at t_{ij}. Therefore, the curtailed right-truncation time for t_{ik} should be $w_{ij} - t_{ij} + t_{ik}$.

In summary, two gaps must satisfy one of the following two conditions in the unbiased $\tilde{R}_{ij}$: (a) for $k > j$, $t_{ik} \leqslant t_{ij} \leqslant w_{ik}$, (b) for $k < j$, $t_{ik} \leqslant t_{ij} \leqslant w_{ij} - t_{ij} + t_{ik}$. Therefore, $|\tilde{R}_{ij}|$ is 2 if either condition holds, and degenerates to 1 otherwise.

[Figure 1 about here]

3.3 *Inferences based on reduced risksets*

The unbiased reduced risksets constructed in §?? are neither necessarily existing, nor necessarily unique when existing. Denote $\{\tilde{R}_{ijk}, k = 1, 2, \ldots, m_{i-1}\}$ the entire reduced risksets at t_{ij}, and $\delta_{ijk} = I\{|\tilde{R}_{ijk}| > 1\}$ the unbiased $\tilde{R}_{ijk}$ indicator. Given $\tilde{R}_{ijk}$ and $\delta_{ijk} = 1$, the conditional likelihood contribution of $\tilde{R}_{ijk}$ is then

$$\frac{\exp(\beta^{\mathrm{T}} Z_{ij})}{\exp(\beta^{\mathrm{T}} Z_{ij}) + \exp(\beta^{\mathrm{T}} Z_{ik})}, \tag{5}$$

which resembles the ones from the Cox proportional hazards model for paired failure times as in Kalbfleisch & Prentice (1980, p. 191). Its corresponding score function is:

$$\tilde{S}_{ijk}(\beta) = Z_{ij} - \bar{Z}_{ijk}(\beta),$$

where

$$\bar{Z}_{ijk}(\beta) = \frac{Z_{ij} \exp(\beta^{\mathrm{T}} Z_{ij}) + Z_{ik} \exp(\beta^{\mathrm{T}} Z_{ik})}{\exp(\beta^{\mathrm{T}} Z_{ij}) + \exp(\beta^{\mathrm{T}} Z_{ik})}.$$

It is true that $E\{\tilde{S}_{ijk}(\beta)|\tilde{R}_{ijk}, \delta_{ijk} = 1; \beta\} = 0$. In addition, $E\{\tilde{S}_{ijk}(\beta)|\tilde{R}_{ijk}, \delta_{ijk} = 0; \beta\} = 0$. Therefore, we can use $\tilde{S}_{ijk}(\beta)$'s as "building blocks" to construct the estimating function for subject i as

$$\tilde{S}_i(\beta) = \frac{\sum_{j=1}^{m_i-1} \sum_{k=1}^{m_i-1} \delta_{ijk} \tilde{S}_{ijk}(\beta)}{\sum_{j=1}^{m_i-1} \sum_{k=1}^{m_i-1} \delta_{ijk}}.$$

If we let $\delta_{ij} = \sum_{k=1}^{m_i-1} \delta_{ijk}$, $g_{ij} = \delta_{ij} / \sum_{j=1}^{m_i-1} \delta_{ij}$ and $\bar{Z}_{ij}(\beta) = \delta_{ij}^{-1} \sum_{k=1}^{m_i-1} \delta_{ijk} \bar{Z}_{ijk}(\beta)$, straightforward algebraic manipulation shows that the ultimate set of estimating functions using all the subjects are

$$\tilde{S}(\beta) = n^{-1} \sum_{i=1}^{n} \tilde{S}_i(\beta) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i-1} g_{ij} \left\{ Z_{ij} - \bar{Z}_{ij}(\beta) \right\}.$$

It is straightforward that $\tilde{S}(\beta)$ is unbiased and thus the estimators of β can be obtained by solving $\tilde{S}(\hat{\beta}) = 0$.

With the special way of the reduced riskset construction, the martingale theory for counting processes of the usual proportional reverse-time hazards model may not be applied in any straightforward sense. However, standard asymptotic likelihood methods will be able to show the existence of $\hat{\beta}$, its uniqueness and consistency under the assumed regularity conditions in the Appendix.

In addition, since $\tilde{S}(\beta)$ is the sum of $\{\tilde{S}_i(0)\}_{i=1}^n$ as iid unbiased estimating functions, it is true that $n^{-1/2}\tilde{S}(\beta)$ is asymptotically normal with mean zero and variance $\Sigma(\beta_0)$ by the Central Limit Theorem. Following the consistency of $\hat{\beta}$ and a Taylor series expansion, we are able to establish the asymptotic normality of $\hat{\beta}$ as well. Details of technical proofs are given in the Appendix.

THEOREM 1. Under the assumed regularity conditions,

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} N(0, D^{-1}(\beta_0)\Sigma(\beta_0)\{D^{-1}(\beta_0)\}^{\mathrm{T}}).$$

And a consistent estimator of $D^{-1}(\beta_0)\Sigma(\beta_0)\{D^{-1}(\beta_0)\}^{\mathrm{T}}$ can be obtained by replacing β_0 with $\hat{\beta}$, $n^{-1}\hat{D}^{-1}(\hat{\beta})\Sigma(\hat{\beta})\{\hat{D}^{-1}(\hat{\beta})\}^{\mathrm{T}}$, where $D(\beta) = E\{(\partial/\partial\beta)\tilde{S}_i(\beta)\}$.

In practice, to solve the estimating equation, a Newton-Raphson iteration algorithm can be adapted. That is, at the kth step of iteration, let the $(k+1)$st solution to the equation to be

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \hat{D}^{-1}(\hat{\beta}^{(k)})\tilde{S}(\hat{\beta}^{(k)}).$$

To our experience, this algorithm is reasonably efficient and the burden of computing is not demanding. The variance estimation of sandwich-type is also straightforward.

Since the estimating equations are constructed from the conditional score functions based on the eligible reduced risksets with equal weight, it is not expected that the proposed estimating equations would be fully efficient in general. However, if one prefers, deterministic weights can be added to the components in $\tilde{S}(\beta)$ to enable potentially more efficient estimating equations. For example, let

$$\tilde{S}^G(\beta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i-1} G_{ij}^{-1/2} g_{ij}\{Z_{ij} - \bar{Z}_{ij}(\beta)\}, \tag{6}$$

where G_{ij} is the diagonal matrix with identical diagonal elements in $E\left[g_{ij}\{Z_{ij} - \bar{Z}_{ij}(\beta)\}\right]^{\otimes 2}$.

Similar to the Cox proportional hazards model, the assumption of multiplicative form is critical to the proportional reverse-time hazards models. To assess the model adequacy, the rationale in Gill and Schumacher (1987) for the Cox proportional hazards model can be adopted. Denote $\hat{\beta}^G$ the solution to $\tilde{S}^G(\beta) = 0$. Then we can compare in difference the two corresponding estimators of $\hat{\beta}^G$ and $\hat{\beta}$, using the quadratic form

$$T_{\mathrm{GS}} = (\hat{\beta}^G - \hat{\beta})^{\mathrm{T}}\hat{V}^{-1}(\hat{\beta}^G - \hat{\beta}),$$

where $\hat{V}$ is an appropriate estimator of variance-covariance matrix of $\hat{\beta}^G - \hat{\beta}$. The statistic T_{GSL} is asymptotically χ_p if the proposed models are true.

4 Numerical Studies

Monte Carlo simulation studies are conducted to evaluate the validity of the proposed inference procedures, when the underlying models are the reverse-time hazards models as specified in (1). The following proportional reverse-time hazards models are used in our simulation studies:

$$F_{ij}(t) = F_{0i}(t)^{\exp(\beta^{\mathrm{T}} Z_{ij})}, \tag{7}$$

where $Z_{ij} = (Z_{ij,1}, Z_{ij,2})$ and (β_1, β_2) are two-dimensional vectors, for $j = 1, 2, \ldots$, $i = 1, \ldots, n$. The censoring times c_i, $i = 1, 2, \ldots, n$, are independently generated by the exponential distribution with mean μ. For subject i, the recurrence times under model (5) are independently generated by first generating u_{ij} from Uniform$(0, 1)$ distribution and then calculating

$$t_{ij} = F_{ij}^{-1}(u_{ij}) b \left[-\log \left\{ 1 - u_{ij}^{\exp(-\beta^{\mathrm{T}} Z_{ij})} \right\} \right]^{1/c},$$

The observed data thus include $(t_{i1}, \ldots, t_{m_i-1}, t_{m_i}^{+})$, $i = 1, \ldots, n$, such that

$$\sum_{j=1}^{m_i-1} t_{ij} \leqslant c_i, \text{ and } \sum_{j=1}^{m_i} t_{ij} > c_i.$$

To characterise the underlying heterogeneity among the subjects, individual baseline hazard function $F_{0i}(t)$ is chosen between the standard Weibull distribution with density of $ct^{c-1}\exp(-t^c)$, where $c > 0$, and the standard Log-normal distribution function with density of $(2\pi)^{-1/2} t^{-1} \exp\{-(\log t)^2/2\}$, with equal probabilities of 0.5. We select c to be 0.8, 1 and 2.5, to represent decreasing, constant and increasing baseline hazard functions of Weibull, respectively. Sample sizes are selected to be 100 and 250 to represent relatively small and large sample sizes, respectively. Censoring times are selected to be 10 and 15 to represent relatively short and long follow-up period, respectively. Two covariates are used: $Z_{ij,1} = j$ for trend measure, while $Z_{ij,2} = e_{ij}$ simulated from uniform distribution $U[0, 1]$ to represent some time-dependent confounding variable needed to be adjusted. True parameter $\beta_0 = (\beta_{10}, \beta_{20})$ are selected to be $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$, respectively. For each configuration, 10,000 simulations are conducted. Its empirical bias, defined as the difference between empirical mean and the true parameter, and coverage probabilities are computed. Details of results are listed in Table 1. As shown in the table, the proposed estimators are virtually unbiased and the corresponding confidence intervals have proper confidence levels.

[Table 1. about here]

In 1938, systematic registration was started in Denmark for the mental health patients admitted to the hospitals for treatment. The registration includes all the cases from 86 psychiatric institutions in the entire nation of Denmark. In investigating the schizophrenic

epidemiology of such a population, the average schizophrenia progression over a relatively long period of time is of critical research interest. Specifically in the Denmark schizophrenia study, the gap times between two consecutive hospitalizations are thus selected as endpoints to study the longitudinal pattern of schizophrenia progression (Eaton, et al., 1992). In fact, the gap times are measured by days and collected from 8,811 patients (5,493 males and 3,318 females) who were admitted to the hospitals due to schizophrenic symptoms for the first time in their lives between April 1, 1970 and March 25, 1988. Given the nature of dataset of this magnitude over a wide geographic region, heterogeneity among subjects is highly probable and any assumption on the underlying characteristics may need to be scrutinized.

In Wang & Chen (2000), a testing procedure was proposed and applied to this data set and detected that there is similar deterioration patterns of the disease among the patients with onset ages less than 20 and those with above. A regression model based on the semi-parametric accelerated failure time model was also used to estimate the magnitude of the pattern. In order to contrast with their findings, we first choose the same index of trend measure of $Z_{ij} = j$ and $Z_{ij} = \sqrt{j}$ in model (1), as used in Wang & Chen (2000). We obtain the β-estimates of -0.0196 and -0.1684, with standard errors ($s.e.$) of 0.0010 and 0.0069, respectively. Both of associated p-values are extremely small, and their negative signs suggest deterioration pattern, which is consistent with the reported results. When model (1) is applied separately to the group with onset age $\leqslant 20$ and otherwise, the β-estimates are -0.0155 ($s.e. = 0.0018$, $p < 0.0001$) and -0.0213 ($s.e. = 0.0012$, $p < 0.0001$) for $Z_{ij} = j$, respectively. This means there is same deterioration pattern for both onset age groups, although the later onset age group may show a stronger pattern. Similar conclusions are reached for $Z_{ij} = \sqrt{j}$: the β-estimates are -0.1585 ($s.e. = 0.0145$, $p < 0.0001$) and -0.1713 ($s.e. = 0.0078$, $p < 0.0001$) for the younger and older onset age groups respectively.

Although it is not of main interest in this article, the grouping effect, or the time-independent covariate effect, is not estimated because of the "stratification" nature of our proposed models and inference procedures. However, similar to the conditional logistic regression models for the matched case-control study, we are still able to estimate the interaction terms of the time-independent and -dependent covariates. For example, we estimate that the interaction of longitudinal pattern measure and the onset age grouping is 0.00071 ($s.e. = 0.00017$, $p < 0.0001$). This suggests that the longitudinal patterns of schizophrenia progression are significantly different between the two onset age groups, although they share same direction of progressive deterioration separately.

5 Discussion

Because of the longitudinal nature, it is well known that the recurrence times as a type of serial multivariate survival times have different statistical structure from those of parallel

multivariate survival times, such as collected in the family studies. For example, pervious works such as Wei, Lin & Weissfeld (1989) may be more applicable to the data sets of the latter type, as noted in Pepe & Cai (1993). Other works such as Prentice, William & Peterson (1981) and Chang & Wang (1999) focus more on the conditional analysis of recurrence times. More recently, marginal approaches such as in Huang (2000) are also explored to model the recurrence times, although maybe under different contexts. The focus of our article is to model and estimate the longitudinal pattern parameter of the recurrence times, when the study population is considered highly heterogeneous and under censoring.

To accommodate censoring and heterogeneity, this paper utilizes the comparability concept to construct appropriate risksets of gap times as truncated observations. Similar to the usual univariate right-truncated data, the proportional hazards model does not serve as a natural model, but instead, the proportional reverse-time hazards model is proven to be a more proper candidate. The comparability condition for the reduced risksets identified in this paper subsequently fits the model and overcomes all the heterogeneous baseline distribution functions as nuisance.

However, the comparability condition does have limitations to certain degree. One major limitation is that the complete recurrence times are only considered as "comparable" pairwise. So it is of greater interest but non-trivial to extend to the comparability condition to more than paired recurrence times, which will allow us to gain more efficiency in estimation. In addition, similar to the conditional inference procedures of the fixed-effect logistic regression models for matched case-control studies, the proposed inference procedure does not aim to estimate the subject-specific covariate effects, if the population heterogeneity is related to such subject-specific covariates. A straightforward approach is to use first gap times only, although more effective and more efficient approaches are needed.

APPENDIX

Asymptotics

Martingale theory has been useful in developing asymptotic theory for the inference procedures of the Cox proportional hazards models (Andersen & Gill, 1982). However, martingales are concerned with future events conditioning on the entire history up to the time points at which risksets are constructed. Within the current framework, however, the usual martingale theory is not able to be used in straightforward terms and alternative techniques are applied in developing asymptotic properties in this article. In the following development, without loss of generality, we further assume that β is a scaler. It should not be difficult to extend all the results to the multivariate situation.

The following regularity conditions are assumed:

(1) There exist an $l \in \{1, 2, \ldots, n\}$ and enough big constant $C_0 > 0$ such that $\int_0^{C_0} \kappa_{0l}(s)ds < \infty$. In addition, $\Pr\{\sum_{(i,j,k)} \delta_{ijk} > 0\} = 1$.

(2) There exists a finite $M > 0$ for a neighborhood U_0 at β_0 such that

$$\sup_{(i,j),\beta \in U_0} [E\{Z_{ij} \exp(\beta' Z_{ij})\}] < M.$$

(3) There exist $\Sigma(\beta_0)$ and positive-definite $D(\beta_0)$ such that

$$\left\| \hat{\Sigma}(\beta_0) - \Sigma(\beta_0) \right\| \to 0$$

and

$$\left\| \hat{D}(\beta_0) - D(\beta_0) \right\| \to 0,$$

respectively, where

$$\hat{\Sigma}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{\sum_j \sum_k \delta_{ijk}} \sum_j \sum_k \frac{\delta_{ijk} \exp(\beta Z_{ik})(Z_{ij} - Z_{ik})}{\exp(\beta Z_{ij}) + \exp(\beta Z_{ik})} \right\}^{\otimes 2}$$

and

$$\hat{D}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sum_j \sum_k \delta_{ijk}} \sum_j \sum_k \frac{\delta_{ijk} \exp(\beta Z_{ij}) \exp(\beta Z_{ik})(Z_{ij} - Z_{ik})^{\otimes 2}}{\{\exp(\beta Z_{ij}) + \exp(\beta Z_{ik})\}^2}.$$

Here, $v^{\otimes 0} = 1$, $v^{\otimes 1} = v$ and $v^{\otimes 2} = vv^{\mathrm{T}}$, and $\| \cdot \|$ defines the Euclidean norm.

As shown in §3.3, the estimating function $\tilde{S}(\beta)$ is unbiased. According to the conditions in Foutz (1977) and later used in Pepe & Cai (1993), if the following conditions are satisfied:

Condition F.1. the partial derivatives of $\tilde{S}(\beta)$ with respect to β exist and are continuous;

Condition F.2. the matrix $n^{-1}(\partial/\partial\beta)\{\tilde{S}(\beta_0)\}$ is non-singular with probability converging to 1 as $n \to \infty$;

Condition F.3. and the matrix $n^{-1}(\partial/\partial\beta)\{\tilde{S}(\beta)\}$ converges in probability to the function $A(\beta) = \lim_{n\to\infty} E[n^{-1}(\partial/\partial\beta)\{\tilde{S}(\beta)\}]$ uniformly in β,

then there exists a neighborhood such that a unique consistent solution to $\tilde{S}(\beta) = 0$ exist with probability converging to 1. It is straightforward to verify conditions F.2 and F.3 implied by regularity conditions 2 and 3 in §3.3, respectively. And since F.1 is an obvious fact, the consistency and uniqueness are then established.

By the Taylor series expansion, we know that in the neighborhood of β_0

$$\tilde{S}(\hat{\beta}) - \tilde{S}(\beta_0) = \frac{\partial \tilde{S}(\hat{\beta}_0)}{\partial \beta} \cdot (\hat{\beta} - \beta_0) + \frac{1}{2} \cdot \frac{\partial^2 \tilde{S}(\hat{\beta}^*)}{\partial \beta^2} \cdot (\hat{\beta} - \beta_0)^2,$$

where β^* lies between β_0 and $\hat{\beta}$. Straightforward algebraic manipulation shows that

$$n^{1/2}(\hat{\beta} - \beta_0) = \left\{ n^{-1} \cdot \frac{\partial \tilde{S}(\hat{\beta}_0)}{\partial \beta} + n^{-1} \cdot \frac{1}{2} \cdot \frac{\partial^2 \tilde{S}(\hat{\beta}^*)}{\partial \beta^2} \cdot (\hat{\beta} - \beta_0) \right\}^{-1} \cdot \left\{ -n^{-1/2} \tilde{S}(\beta_0) \right\}. \quad (8)$$

By regularity condition 2 in §3.3, $n^{-1}\{(\partial^2/\partial\beta^2)\tilde{S}(\beta)\}$ is uniformly bounded in the neighborhood of β_0. Therefore, $n^{-1}\{(\partial^2/\partial\beta^2)\tilde{S}(\beta^*)\}(\hat{\beta} - \beta_0)$ converges to 0 in probability.

Because of the way of constructing $\tilde{S}(\beta)$, $n^{-1}(\partial/\partial\beta)\tilde{S}(\beta_0)$ is an average of n iid random variables with finite variance. Therefore, by the Weak Law of Large Numbers (WLLN), it converges in probability to $D(\beta_0) = E\{(\partial/\partial\beta)\tilde{S}_i(\beta_0)\}$, $i = 1, 2, \ldots, n$. In addition, all the $\tilde{S}_i(\beta_0)$'s are iid zero-mean random variables, so by the central limit theorem, $n^{-1/2}\tilde{S}(\beta)$ converges in distribution to a normal with mean zero and variance of $\Sigma(\beta_0)$. Because of the positive-definity of $D(\beta_0)$, it is straightforward to establish the asymptotic normality of $\hat{\beta}$ as specified in Theorem 1. Using the result in Andersen & Gill (1982) and the consistency of $\hat{\beta}$, the consistency of the variance estimators in Theorem 1 is also implied.

REFERENCES

AALEN, O. O. & HUSEBYE, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statist. Med.* **10**, 1227-1240.

ABELSON, R. P. & TUKEY, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Ann. Math. Statist.* **34**, 1347-1369.

ANDERSEN, P. K. & GILL, R. D. (1982). Cox regression model for counting processes - a large sample study. *Ann. Statist.* 10, 1100-1120.

BROOKMEYER, R. & GAIL, M. H. (1994). *AIDS Epidemiology.* New York: Oxford University Press.

CHANG, S.-H. & WANG, M.-C. (1999). Conditional regression analysis of recurrent time data. *J. Amer. Statist. Assoc.* **94**, 1221-1230.

COOK, R. J. & LAWLESS, J. F. (2002). Analysis of repeated events. *Statist. Methods Med. Res.* **11**, 141-166.

COX, D. R. (1962). *Renewal Theory.* New York: Wiley.

COX, D. R. & ISHAM, V. (1980). *Point Processes.* London: Chapman & Hall.

EATON, W. W., BILKER, W., HARO, J. M., HERRMAN, H., MORTENSEN P. BB, FREEMAN, H., BURGESS, P. (1992). Long-term course of hospitalization for schizophrenia: Part II. Change with passage of time. *Schizo. Bull.* **18**, 229-241.

FOUTZ, R. V. (1977). Unique consistent solution to likelihood equations. *J. Amer. Statist. Assoc.* **72**, 147-148.

GELBER, R. D., GELMAN, R. S. & GOLDHIRSCH, A. (1989). A quality-of-life-oriented endpoint for comparing therapies. *Biometrics* **45**, 781-795.

GILL, R. & SCHUMACHER, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289-300.

GROSS, S. T. & HUBER-CAROL, C. (1992). Regression models for truncated survival data. *Scand. J. Statist.* **19**, 193-213.

HUANG, Y. (1999). The two-sample problem with induced dependent censorship. *Biometrics* **55**, 1108-1113.

HUANG, Y. (2000). Multistate accelerated sojourn times model. *J. Amer. Statist. Assoc.* **95**, 619-627.

Kalbfleisch, J. D. & Lawless, J. F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *J. Amer. Statist. Assoc.* **84**, 360-372.

Kalbfleisch, J. D. & Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statist. Sinica* **1**, 19-32.

Kalbfleisch, J. D. & Prentice, R. L. (1980). *Statistical Analysis of Failure Time Data.* New York: Wiley.

Lagakos, S. W., Barraj, L. M. & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data with application to AIDS. *Biometrika* **75**, 515-523.

Lin, D. Y., Sun, W. & Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* **86**, 59-70.

Pepe, M. S. & Cai, J. (1993). Some graphical display and marginal regression analyses for recurrent failure times and time dependent covariates. *J. Amer. Statist. Assoc.* **88**, 811-820.

Prentice, R. L., Williams, B. J. & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**, 373-379.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616-620.

Wang, M.-C. & Chen, Y. Q. (2000). Nonparametric and semiparametric trend analysis of stratified recurrent times. *Biometrics* **56**, 789-794.

Wang, M.-C., Jewell, N. P. & Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* **14**, 1597-1605.

Wei, L. J., Lin, D. Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.* **84**, 1065-1073.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **12**, 163-177.

Table 1: Summary of Simulation Studies

n	μ[a]	c[b]		(0,0) $(\hat{\beta}_{01}, \hat{\beta}_{02})$		(1,0) $(\hat{\beta}_{01}, \hat{\beta}_{02})$		(1,0) $(\hat{\beta}_{01}, \hat{\beta}_{02})$		(1,1) $(\hat{\beta}_{01}, \hat{\beta}_{02})$	
100	10.8		Bias[c]	0.0118	0.0062	0.0079	0.0095	-0.0102	0.0046	0.0044	0.0018
			Cov. Pr.[d]	0.9534	0.9509	0.9506	0.9493	0.9515	0.9502	0.9478	0.9457
	1.0		Bias	0.0052	-0.0045	0.0004	0.0053	-0.0038	0.0039	0.0041	0.0045
			Cov. Pr.	0.9517	0.9491	0.9517	0.9521	0.9422	0.9479	0.9507	0.9527
	2.5		Bias	-0.0005	0.0020	-0.0004	-0.0117	-0.0051	-0.0057	0.0147	0.0014
			Cov. Pr.	0.9488	0.9502	0.9493	0.9468	0.9461	0.9479	0.9519	0.9460
.001	0.0006		-0.0158	-0.0033	-0.0017	0.0033	0.0090	0.0034			
517		0.9494	0.9519	0.9471	0.9546	0.9586	0.9531				
54	0.0163	0.0133	-0.0028	0.0067	-0.0051						
80	0.9527	0.9501	0.9575	0.9551							
066	0.0098		0.0119	0.0027							
	0.9505	0.9513									

μ^{a} is mean censoring time. c^{b} is the shape parameter of the baseline hazard function. Bias^{c} is the average $\hat{\beta}$'s minus β_0. Cov. Pr.^{d} is the coverage probability of the 95% confidence intervals. All the entries are computed from 10,000 simulations.

Figure 1: Illustrative example of two members in the reduced riskset of t_{ij}