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

Year 2002 *Paper* 103

Regression Analysis of Recurrent Gap Times with Time-Dependent Covariates

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 and Public Health, Johns Hopkins University

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

http://biostats.bepress.com/ucbbiostat/paper103

Regression Analysis of Recurrent Gap Times with Time-Dependent Covariates

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

Abstract

Individual subjects may experience recurrent events of same type over a relatively long period of time in a longitudinal study. Researchers are often interested in the distributional pattern of gaps between the successive recurrent events and their association with certain concomitant covariates as well. In this article, their probability structure is investigated in presence of censoring. According to the identified structure, we introduce the proportional reverse-time hazards models that allow arbitrary baseline function for every individual in the study, when the time-dependent covariates effect is of main interest. Appropriate inference procedures are proposed and studied to estimate the parameters of interest in the models. The proposed methodology is demonstrated with the Monte-Carlo simulations and applied to a well-known Denmark schizophrenia cohort study data set.

1 Introduction

In some longitudinal follow-up studies of a group of participants, individual participant may experience a series of successive occurrences of events as time progresses. The durations of these successive occurrences of events is called event history (Lindsey, 1993, p. 235). In the event history data, these events can be of different types, such as different stages of a disease. For example, in the natural history of the disease of acquired immunodeficiency syndrome (AIDS), the progression of the disease is often characterized as sero-negative→Human Immunodeficiency Virus (HIV) infection→AIDS→death (Brookmeyer & Gail, 1994). Or, these events can be of same type, for example, recurrent hospitalizations of schizophrenic patients (Eaton, et al., 1992b). When the events are considered as points occurring along the time axis, they form point processes.

There are usually two basic sampling schemes to observe the point processes of the event history data:

Incidence cohort sampling. The incidence cohort sampling forms a sample of the participants from the incidence population. That is, individual participant is included in the sample as soon as the initiating event occurs. Events of this type are frequently observed from follow-up registry data collected in hospitals or health institutions.

Prevalence cohort sampling. The prevalent cohort sampling forms a sample of the participants from a target population and their recurrent events are monitored during the follow-up period. For example, the samples could be formed by the participants tested HIV sero-positive at study-entry with recurrent opportunistic infections observed in the follow-up period. Under this sampling scheme, it is possible to observe no recurrent event within the follow-up period.

In practice, the above-mentioned schemes may interplay and form a mixed cohort of both incidence and prevalent participants.

For general point processes, as pointed out in Cox & Isham (1980, p. 11), there are three equivalent specifications to determine the processes: the intensity specification (the complete intensity function of occurrences), the duration specification (the joint distribution of durations between successive events) and the counting specification (the joint distribution of the occurrence counts in any arbitrary sets). Although the three specifications are equivalent to certain degree, the intensity and counting specifications are often more convenient to be studied in general theory development (Lawless & Nadeau, 1995). However, it is also often important in practice to study the duration specification. This is especially of interest for the event history data of the recurrent events of same type, or simply, recurrent event data, say. For example, in Eaton, et al. (1992b), the distributional pattern of durations between the successive hospitalizations serves as an important index for the progression of

the schizophrenic disease: do the durations have a pattern tending to be longer and longer, or shorter and shorter? if there is such a pattern of trend, can it be tested and its magnitude be estimated? what type of risk factors may play significant roles in describing or predicting the observed patterns?

In some ideal situations, special examples of the point processes can be adapted to analyze the recurrent event data. For example, if the independence of the durations is plausible, the point processes become the renewal processes, or if some Markovian structure is further added, the Markov renewal processes or semi-Markov processes can be used (Cox, 1962). In reality, however, the interaction of two factors imposes a unique and prominent challenge in studying the durations between the recurrent events:

Within-subject Correlation. Within-subject correlation is often hypothesized, when heterogeneity is observed as different distributional patterns of the durations for different individuals. Part of the heterogeneity can be explained by the observed predictors, such as some demographic and socio-economic status variables, or known risk factors, e.g., smoking. But some other factors can cause the heterogeneity as well. These factors are either ignored in data collection, e.g., the participants' exposure to unknown environmental risk factors, or unable to be measured accurately, e.g., the participants' genotypes. Therefore, even after accounting for the observed predictors, there may still exist unaccountable heterogeneity left among the participants. As a result, it is usually not appropriate to assume that the durations of the successive occurrences are independent.

Censoring. A participant's followup is called "censored" if its event history is not completely observed. There are many possible reasons, e.g., early termination of follow-up due to the predetermined limit of study time period, or loss to follow-up due to migration, to prevent the entire history from being fully observed.

To see the challenge, consider a simplified example of two specific durations in the incidence cohort sampling, i.e., the durations of the initiating event to the first event (T_1) and the first event to the second event (T_2). In literature, sometimes it is assumed that the censoring time (C) is independent of (T_1, T_2), while T_1 and T_2 are correlated (Wang & Chang, 1999). Based on the observed data, most of the traditional statistical methods, such as such as the Kaplan-Meier product-limit estimators or the Cox regression models for censored survival data, can be adapted to study the pairs of (T_1, C)'s, because of the independence of T_1 and C. However, they may not be applied naively in analyzing $(T_2, C - T_1)$, where $C - T_1$ is the censoring time for T_2 and apparently dependent of T_2 due to the correlation between T_1 and T_2. That is, the probability of T_2 being censored depends on the magnitude of itself, even though indirectly. This phenomenon was recognized when studying the time without symptoms of disease and toxicity of treatment (TWiST) as end point (Gelber, Gelman & Goldhirsch, 1989), and later called "induced informative censoring" (Lin, Sun & Ying, 1999)

or "induced dependent censorship." (Huang, 1999). More comprehensive discussion on this phenomenon in incidence and prevalence cohort samplings can be found in Wang (1999).

Because of the potential bias in the traditional survival analysis of the duration times in presence of the induced informative censoring, researchers have recently developed some new statistical methodologies from various perspectives. The methods in Wang & Wells (1998), Huang & Louis (1998) and Lin, Sun & Ying (1999) can be used in the estimation of the joint distribution of duration times. When the duration times share identical marginal survival functions, Wang & Chang (1999) proposed an estimator of the marginal survival function. For hypothesis testing of duration times, Lin & Ying (2001) studied several two-sample testing procedures based on the estimators in Lin, Sun & Ying (1999). Chang (2000) developed a two-sample testing procedure using the accelerated failure time model. In Wang & Chen (2000), significance testing and regression procedures are proposed for the trend measures of duration times. The methods in Prentice, Williams & Peterson (1981), Chang & Wang (1999) and Huang (2000) can be used in the regression analysis of duration times.

In this paper, the focus will be on the development of a regression method of the duration times, when the recurrent events are of same type. The regression method will study the pattern of duration times, for example, to identify the treatment efficacy over time or confirm the stability of the duration times. Because of the special serial feature of the recurrent event data and the potential complication caused by the induced informative censoring, the recurrent event data are essentially different from clustered failure times collected from families or litters. Therefore, we will first study the probability structure of the observed duration times in §2.1. The regression models that fits such a structure and their estimation are proposed and studied in §2.2. Numerical analyses including simulations and real data example are in §4. Some remaining issues are discussed in §5. The technical proofs are collected in the Appendix.

2 Semiparametric Regression Models

2.1 *Probability structure*

Suppose that there are n independent participants recruited into a study. Let $i = 1, 2, \ldots, n$ be the participant index. And let $j = -1$ denote the event of study onset, $j = 0$ the index for the initiating event and $j = 1, 2, \ldots$ the indices for the subsequent events. Let T_{ij} be the time between event $j-1$ and event j, for $i = 1, 2, \ldots, n$ and $j = 0, 1, 2, \ldots$. Then T_{i0} is the time from the study onset to the initiating event for participant i. In the incidence cohort sampling, the occurrence of the initiating event usually determines the study onset and, in this case, $T_{i0} \equiv 0$. In the prevalence cohort sampling, T_{i0} represents the time from study onset to the initiating event, which may be often observed subject to right-censoring.

Furthermore, let $N_i = (T_{i0}, T_{i1}, T_{i2}, \ldots)$ denote the collection of all the duration times. Let C_i be the censoring time defined as the time from the study onset to the certain time point of censoring. To explore the probability structure of the duration times, we initially make two assumptions on T_{ij}'s. Given a specific participant, i, $i = 1, 2, \ldots, n$, say:

Conditional independence. the duration times $T_{i0}, T_{i1}, T_{i2}, \ldots$ are independently distributed;

Non-informative censoring. the censoring time C_i is independent of N_i.

The first assumption can be viewed as a general type of frailty condition – the participant himself or herself is some unspecified matching criterion for the conditional independence. The second assumption basically implies that the censoring mechanism is conditionally uninformative of the event history process.

Suppose that $(t_{i0}, t_{i1}, \ldots, t_{i,m_i-1}, t^+_{i,m_i})$ is an observed sequence of duration times of participant i. Here, $M_i = m_i$ is the stopping time of event index such that for censoring time $C_i = c_i$,

$$\sum_{k=0}^{m_i-1} t_{ij} \leqslant c_i, \text{ and } \sum_{k=0}^{m_i} t_{ij} > c_i,$$

and $t^+_{i,m_i} = c_i - \sum_{j=0}^{m_i-1} t_{ij}$. The observed duration times of t_{i0}, t_{i1}, ..., t_{i,m_i-1} are considered as "complete" duration times, while the last duration time of t_{i,m_i} is always "censored." To simplify our discussion without loss of generality, we consider the situation when the underlying $(T_{i1}, T_{i2}, \ldots)$ that generate $(t_{i0}, t_{i1}, \ldots, t_{i,m_i-1}, t^+_{i,m_i})$ are also identically distributed.

For any fixed index $j \geqslant 1$ of the durations, it is more likely for relatively short duration times T_{ij} to be observed as complete t_{ij}, given C_i and the durations that have occurred prior to j. In addition, although $(T_{i1}, T_{i2}, \ldots)$ share identical distribution among themselves by the assumption, the observed complete duration times t_{ij}, $1 \leqslant j \leqslant m_i$ tend to be shorter and shorter as j increases. To see this, let $W_{ij} = C_i - \sum_{k=0}^{j-1} t_{ik}$, which is the censoring time of the jth duration time T_{ij}, given all the durations prior to jth duration. Then the complete duration t_{ij} is observed from the conditional distribution of T_{i1} given $T_{i1} \leqslant W_{ij}$. That is, conditional on $W_{ij} = w_{ij}$, the observed complete t_{ij}, $1 \leqslant j \leqslant m_i - 1$ is sampled subject to the independent right-truncation of $T_{i1} \leqslant w_{ij}$ (Lagakos, Barraj & De Gruttola, 1988, Kalbfleisch & Lawless, 1989). Also as a result, as j increases, W_{ij} becomes smaller, and shorter complete t_{ij} would be observed.

However, the last duration time of t_{i,m_i} is observed subject to intercept sampling (Vardi, 1982). The backward recurrence time W_{i,m_i} can be considered as the truncation time in a left truncation model. That is, given $W_{i,m_i} = w_{i,m_i}$, the last duration time t_{i,m_i} is in fact sampled from the conditional distribution of T_{i1} given $T_{i1} \geqslant w_{i,m_i}$. It was noted that, when the censoring time is a very large constant approaching $+\infty$, its limiting distribution is length-biased (Cox, 1962, p. 61).

2.2 *Proportional reverse time hazards models*

Although the observed complete durations and the last censored duration are of opposite directions in truncation, the truncation effect cannot be canceled with each other in analysis by simply pooling them together (Wang & Chang, 1999). However, because of the unique probability structure of the complete duration times as right-truncated observations, we will be focusing on the models for them in this article.

As discussed in the previous section, the complete failure times are always right-truncated. For right-truncated data, due to their complementary structure to the left-truncated data in reverse time, researchers often adapt the usual life-table and survival analysis techniques in reverse time to study their stochastic properties. For example, Lagakos, Barraj & De Gruttola (1988) and Kalbfleisch & Lawless (1989) developed nonparametric estimation of the survival functions in reverse time. Kalbfleisch & Lawless (1991) and Gross & Huber-Carol (1992) further extended the Cox regression models (Cox, 1972) to the reverse time hazards functions.

Denote the cumulative distribution function $F(t) = \Pr\{T \leqslant t\}$, and $F_{ij}(t)$ the cumulative distribution function for the jth duration time of the ith subject, T_{ij}, $i = 1, \ldots, n, j = 1, 2, \ldots$. Furthermore, let the reverse time hazard function be

$$\kappa(t) = \lim_{\Delta t \to 0+} \frac{\Pr\{t - \Delta t \leqslant T \leqslant t | T \leqslant t\}}{\Delta t} = \frac{d \log F(t)}{dt}.$$

As in Kalbfleisch & Lawless (1991), then the proportional reverse time hazards model for the (i, j)th duration time is

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

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 $j = 1, 2, \ldots$. Here, $\kappa_{i0}(t)$'s are the unknown baseline reverse time hazards functions (Kalbfleisch & Lawless, 1991). The stochastic relationship of failure times can be seen more clearly in an equivalent form of model (1):

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

Now suppose that $Z_{ij} = z_{ij}$ would become $z_{ij} + 1$ with one unit increment, then

$$F(t|Z_{ij} = z_{ij} + 1) = F(t|Z_{ij} = z_{ij})^{\exp(\beta)}.$$

Therefore, T_{ij} is stochastically longer for $Z_{ij} = z_{ij} + 1$ when $\beta > 0$, and shorter when $\beta < 0$. $\beta = 0$ means no covariate effect.

The covariate Z_{ij} in (1) is duration-dependent. For example, Z_{ij} can be increasing with j, which may represent an assigned trend measure for the jth duration time (Abelson & Tukey,

1963). A special situation is when $Z_{ij} = j$. Then the parameter β measures the direction and magnitude of trend among the duration times. That is, $\beta > 0$ stands for longer and longer duration times, while $\beta < 0$ for shorter and shorter duration times. $\beta = 0$ means no trend.

As discussed by many authors, for the proportional reverse time hazards model, the usual at-risk process for the proportional hazards models is not "adapted" to the history process (Lagakos, Barraj & De Gruttola, 1988, Kalbfleisch & Lawless, 1991 and Gross & Huber-Carol, 1992). But if a negative time scale were allowed, i.e., 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): for $t \leqslant 0$,

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

where $\lambda_{ij}(t) = \kappa_{ij}(-t)$.

As the proportional hazards model can be considered as the continuous version of the logistic regression model of discrete failure times, the proportional reverse times hazards model is the continuous version of the complementary log-log regression model (Kalbfleisch & Lawless, 1991). Although the parameter is no longer interpreted as odds ratio or hazards ratio in the proportional reverse time hazards or the complementary log-log models, it entails simple interpretation on the cumulative distribution functions. In the later development of estimation, it will become more evident that the proportional reverse time hazards model has more advantages in utilizing the right-truncation probability structure of complete durations.

3 Inference Procedures

3.1 *Biased risksets*

The concept of riskset is useful in analyzing the right-censored and left-truncated data to construct the Kaplan-Meier product-limit estimators and estimate the Cox regression models, for example, see Woodroofe (1985) and Wang, Jewell & Tsai (1986). Brookmeyer & Gail (1994, p. 89) used the same term for the right-truncated data, because it carries analogy with the life-table analysis in reverse time, although its definition is different from the left-truncated data. We first consider the usual risksets for the complete duration times as right-truncated observations, and however argue that these risksets are "biased" in estimating the parameters in model (1). A biased riskset means that not all its members share same baseline distribution function even after accountable covariate adjustment.

According to the definition of risksets in Brookmeyer & Gail (1994), the individuals in the riskset of certain time, t, are those "whose truncation times are greater than or equal to"

t and "whose incubation periods [i.e., observed survival times] are less than or equal to" t. Therefore, with the assumption of conditional independence for subject i, the seemly correct riskset at t_{ij} is

$$R_{ij} = \{k : t_{ik} \leqslant t_{ij} \leqslant w_{ik}, k = 1, \ldots, m_i - 1\}, \tag{2}$$

where w_{ik} are defined as in §2.1. Then the partial likelihood function in reverse time is

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

assuming that t_{ij}'s are distinct complete duration times.

If R_{ij} were unbiased, adjusted for the covariates in the reverse time hazards model, the members in R_{ij} would have comparability, which allows themselves to have fair chance to compete to fail at t_{ij}. This is because they would share the same baseline reverse time hazard function, and more importantly, the members in the unbiased riskset would be a random or unbiased sample from the corresponding underlying risk population. Thus, the score function contributed by the ith subject, which 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, there is complication with the riskset defined in (2). That is, although t_{ik} is independent of w_{ik} for any specific k, the truncation time $w_{ik} = c_i - \sum_{l<k} t_{il}$ is apparently a function of t_{ij} defining the riskset R_{ij} as long as $j < k$. Therefore the defined riskset R_{ij} does not fit the ordinary riskset for right-truncated observations. In fact, the members in R_{ij} do not have the comparability as needed for the riskset to be unbiased, either.

For a hypothetical example, suppose the censoring time $C_i = 10$, $T_{i0} = 1$ and $(T_{i1}, T_{i2}, T_{i3}, \ldots)$ are as in Table 1.

[Table 1 about here]

As shown in Table 1, the earlier durations are more likely to be included in the risksets of later durations, or equivalently, the later durations are less likely to be included in those of earlier ones, because the truncation times becomes shorter and shorter for later durations. Therefore the comparability of the subjects within R_{ij} does not exist any longer.

Nevertheless, if the comparability in R_{ij} were preserved, then one subject would be able to be replaced by the other in the riskset. For example, the truncation time for T_{i2} is $W_{i2} = 8$. Hence $T_{i2} = 5$ is observed and T_{i1} in its riskset. However, if T_{i1} were replaced by the value

of T_{i2}, which is 5, then W_{i2} would be 4. As a result, T_{i2} would not be fully observed but censored at 4. Again, this essentially means T_{i1} can by no means be comparable to T_{i2}. The comparability in R_{ij} is violated and therefore R_{ij} is a biased riskset, which would lead to biased inference procedures for the model parameters.

3.2 *Unbiased reduced risksets*

As discussed in the above section, the cause that leads to the biased risksets is clear, i.e., some of the subjects in the riskset, R_{ij}, say, do not have their fair share of chance to compete to fail at t_{ij}. Therefore, an ideal treatment to correct such bias is to allow all the subjects in the riskset being capable of competing to fail at t_{ij}. That is, by replacing every subject in R_{ij} with T_{ij},

$$|R_{ij}|T_{ij} + \sum_{k \in R_{ij}^c} T_{ik} \leqslant C_i$$

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 equations can be constructed based on the risksets satisfying this criterion. However, this way of construction is expected to be cumbersome because the criterion becomes less likely to be satisfied as $|R_{ij}|$ becomes larger and consequently it will disqualify many risksets as unbiased.

Naturally, more effective approaches can be considered by reducing the number of replacements of T_{ik} in R_{ij}, which can be achieved by actually reducing $|R_{ij}|$. The most aggressive reduction is to include only one duration, t_{ik} say, in R_{ij} at a time, in addition to t_{ij} itself. That is, a reduced riskset $\tilde{R}_{ij}$ at t_{ij} would have at most two members, t_{ij} and t_{ik}. As previously discussed, we know that not every t_{ik} is eligible to be included to form the reduced unbiased riskset. Then the question arises naturally: what kind of t_{ik}'s are eligible? For illustration purpose, in Figure 1, we plot two hypothetical cases of t_{ij} and t_{ik} and further explore the eligibility conditions of t_{ik} to be in the reduced unbiased riskset of t_{ij}:

Case 1. For $k > j$, $\tilde{R}_{ij}$ is to include a later duration;

Case 2. for $k < j$, $\tilde{R}_{ij}$ is to include an earlier duration.

As shown in the figure of Case 1, the truncation time w_{ik} of t_{ik} is shorter than that of t_{ij} by default because t_{ik} happens later than t_{ij}. Therefore, the usual condition of riskset for right truncated observations applies. That is, $t_{ik} \leqslant t_{ij} \leqslant w_{ik}$, which allows t_{ik} to fairly compete with t_{ij} to fail in $\tilde{R}_{ij}$. In Case 2, the truncation time of w_{ik} of t_{ik} is however longer than that of t_{ij} by default because t_{ik} now happens earlier than t_{ij}. Although t_{ik} may still be shorter than t_{ij}, it does have excessive advantage of being potentially longer than w_{ij}, which t_{ij} will never have. Therefore, we need to modify, i.e., to curtail w_{ik} to reduce such

an excessive advantage. In fact, due to the right-truncation, the largest room left for t_{ij} to potentially grow is $w_{ij} - t_{ij}$. This is also the largest room left for t_{ik}, in order to allow t_{ik} to fairly compete with t_{ij} to fail. Therefore, the curtailed right-truncation time for t_{ik} should be $w_{ij} - t_{ij} + t_{ik}$.

In summary, two members must be satisfying one of the two following criteria to let $\tilde{R}_{ij}$ be unbiased:

Criterion 1. When $k > j$, $t_{ik} \leqslant t_{ij} \leqslant w_{ik}$;

Criterion 2. When $k < j$, $t_{ik} \leqslant t_{ij} \leqslant w_{ij} - t_{ij} + t_{ik}$.

Therefore $|\tilde{R}_{ij}|$ is 2 for reduced unbiased riskset if either condition holds, and 1 otherwise. Later in the section of Discussion, we will point out that the above two conditions are essentially to establish the so-called "comparability" of t_{ij} and t_{ik} in Bhattacharya, Chernoff & Yang (1983), Efron & Petrosian (1999) and Wang & Chen (2000).

[Figure 1 about here]

3.3 *Inference based on reduced risksets*

The reduced unbiased riskset identified in the preceding section is neither necessarily existing, nor necessarily unique even when existing. Denote $\tilde{R}_{ijk}, k = 1, 2, \ldots, m_{i-1}$, for all the possible reduced risksets, and $\delta_{ijk} = I\{|\tilde{R}_{ijk}| > 1\}$, which is an indicator of unbiased $\tilde{R}_{ijk}$. With the assumptions in §2.1, given $\delta_{ijk} = 1$, T_{ij} and T_{ik} are independent and able to compete to fail at t_{ij}, accounting for appropriate covariate adjustment specified in model (1). Therefore, the conditional probability for T_{ij} to fail at t_{ij} is

$$\frac{\exp(\beta^{\mathrm{T}} Z_{ij})}{\exp(\beta^{\mathrm{T}} Z_{ij}) + \exp(\beta^{\mathrm{T}} Z_{ik})}, \tag{3}$$

and its corresponding score function is the derivative of the log of (3):

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

where

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

Thus, $E\{\tilde{S}_{ijk}(\beta)|\tilde{R}_{ijk}, \delta_{ijk} = 1; \beta\} = 0$. It is also true, although trivial, $E\{\tilde{S}_{ijk}(\beta)|\tilde{R}_{ijk}, \delta_{ijk} = 0; \beta\} = 0$. Therefore, we can use $\tilde{S}_{ijk}(\beta)$'s as "building blocks" to construct an estimating

function for subject i as

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

As a result, the set of estimating functions of parameter β using observations of all n subjects are

$$\tilde{S}(\beta) = n^{-1}\sum_{i=1}^{n}\tilde{S}_i(\beta).$$

If we let

$$\bar{Z}_{ij}(\beta) = \frac{\sum_{k=1}^{m_i-1}\delta_{ijk}\bar{Z}_{ijk}(\beta)}{\sum_{k=1}^{m_i-1}\delta_{ijk}},$$

and $\delta_{ij} = \sum_{k=1}^{m_i-1}\delta_{ijk}$, straightforward algebraic manipulation shows that $\tilde{S}(\beta)$ is actually

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

where $g_{ij} = \delta_{ij}/\sum_{j=1}^{m_i-1}\delta_{ij}$. The estimating function of $\tilde{S}(\beta)$ is also unbiased, since

$$E\{\tilde{S}(\beta)\} = n^{-1}\sum_{i=1}^{n}E\{\tilde{S}_i(\beta)\}$$

$$= \quad n^{-1}\sum_{i=1}^{n}E\left[E\left\{\tilde{S}_i(\beta)|\tilde{R}_{ijk}, m_i, j = 1,\ldots,m_i-1, k = 1,\ldots,m_i-1\right\}\right] = 0$$

Therefore an estimator of parameter β can be obtained by solving

$$\tilde{S}(\hat{\beta}) = 0. \tag{4}$$

Noticeably, the proposed estimating function restricts the use of itself when $Z_{ij} \equiv Z_i$ for every $i = 1, 2, \ldots, n$. This is because, for any subject-specific covariate, which is time-independent over the entire study, the estimating function degenerates. The associated parameters will disappear in the estimating function because of the product form of the model. Therefore the proposed inference procedure is not able to estimate the subject-specific covariates effects if only main effects are included. Nevertheless, if the interaction between the subject-specific and duration-specific covariates are of major interest, the proposed inference procedure is still able to make inference of the interaction terms. This echoes the similarity in the conditional inference procedures of the fixed-effect logistic regression models for matched case-control studies in epidemiology.

With the special way of constructing the reduced risksets, the standard martingale theory for counting processes (Andersen, et al., 1993) in analyzing the usual right-truncated data

(Kalbfleisch & Lawless, 1991 and Gross & Huber-Carol, 1992) may not be applied here in any straightforward sense. However, in the Appendix, we will be able to show the existence of the solution of $\hat{\beta}$ in (4), its uniqueness and consistency as well under the following regularity conditions:

Condition 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, $\Pr\{\sum_{(i,j,k)} \delta_{ijk} > 0\} = 1$.

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

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

Condition 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.

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 by the central limit theorem. Following the consistency of $\hat{\beta}$ and a Taylor series expansion, we will be able to establish the asymptotic normality of $\hat{\beta}$ under the stated regularity conditions. Details of technical proofs are given in the Appendix.

Theorem 1. Under the above 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 $\hat{\beta}$:

$$n^{-1}\hat{D}^{-1}(\hat{\beta})\Sigma(\hat{\beta})\{\hat{D}^{-1}(\hat{\beta})\}^{\mathrm{T}}.$$

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

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

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

In practice, to solve the estimating equation in (4), 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 effective and the burden of computing is not demanding. The variance estimation is also straightforward.

4 Numerical Studies

4.1 *Simulations*

Monte Carlo simulation studies are conducted to evaluate the proposed estimation procedures for the reverse time hazards models. We evaluate our estimation procedure in the scenario when the underlying model is indeed with the reverse time hazards model as specified in (1) with constant β for every individual.

The following reversed proportional hazards models are used in our simulation studies:

$$F_{ij}(t) = F_{0i}(t)^{\exp(\beta^{\mathrm{T}} Z_{ij})}, \tag{5}$$

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

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

In the simulations the time to the initial event, t_{i0}, is set to be zero. The observed data thus include $(t_{i1}, \ldots, t_{m_i-1}, t_{m_i}^{+})$, $i = 1, \ldots, n$, such that

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

For simplicity, the baseline hazard function is chosen to be independent of i. More specifically, it is Weibull distribution function defined by

$$F_{0i}(t) = 1 - \exp\{-(t/b)^c\},$$

where $b, c > 0$. We select c to be 0.8, 1 and 2.5, while letting b be always 1, to represent decreasing, constant and increasing baseline hazard functions, respectively. Sample sizes are selected to be 100 and 250 to represent relatively small, moderate 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 2. As shown in the table, the proposed estimators are virtually unbiased and the corresponding confidence intervals have proper confidence intervals.

[Table 2. about here]

4.2 *Data analysis example*

In 1938, the country of Denmark started systematic registration of mental health patients admitted to hospitals for treatment. The registration includes all the cases from 86 psychiatric institutions in the entire nation of Denmark. In schizophrenic epidemiology, one of the important scientific questions to be addressed is the progression of schizophrenia, which is characterized by the trend pattern of between-hospitalization durations (Eaton, et al., 1992a, Mortensen, et al., 1994). So naturally, in the Denmark schizophrenia study, the recurrence time is the time between two consecutive hospitalizations. The recurrence times are measured by days. They are 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

during the period April 1, 1970, through March 25, 1988. The distributional pattern of the recurrence times can be used as an index for the deterioration or amelioration of the disease.

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. However, it is not clear whether or not their proposed regression estimator is unique because of the discreteness of their proposed estimating functions. In addition, computer-intensive methods such as bootstrapping have to be used to estimate the variance because it involves unknown nonparametric component.

In order to contrast with the results in Wang & Chen (2000), we first choose the same index of trend measure of $Z_{ij} = j$ and $Z_{ij} = \sqrt{j}$ in model (1), as used in their paper. We found that the β-estimates are -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 trend, which is consistent with what were reported in Wang & Chen (2000). 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 trend for both onset age groups, although the later onset age group may show a stronger trend. 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.

Furthermore, we added one more potential confounding variable to adjust in the model, which is the age at each hospital admission. Then it is interesting to find that the direction of schizophrenia progression is reversed: the β-estimates are 0.0228 ($s.e. = 0.0015$, $p < 0.0001$) for $Z_{ij} = j$ and 0.2757 ($s.e. = 0.0135$, $p < 0.0001$) for $Z_{ij} = \sqrt{j}$, respectively. This means that the overall schizophrenia progression might be progressive amelioration if hospital admission age is appropriately adjusted. Similar results are obtained for both separate onset age groups.

Because of the "stratification" nature of our proposed estimating functions, the grouping effect, or the time-independent covariate effect, is not estimable by this methodology. 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 trend measure and the onset age grouping is 0.00071 ($s.e. = 0.00017$, $p < 0.0001$). This may suggest that the trend measures are significantly different for the two onset age groups, although they share same direction of progressive amelioration with adjustment of age at hospital admission.

5 Discussion

This paper explores the comparability condition from the perspective of appropriate risksets of truncated data. 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 a better choice. The comparability condition for the reduced risksets identified in this paper subsequently fits the model, because the baseline hazard functions in reverse time become nuisance. This is the same philosophy as in Wang & Chen (2000), which is to identify comparable recurrence times, but with resort to different models. Actually during the process of identifying the comparability condition, it is not difficult to draw the similar kind of pictures to the Figures 1 and 2 in Wang & Chen (2000) for the comparable pairs of complete recurrence times. As a result, the comparability condition identified in this paper is equivalent to that in Wang & Chen (2000).

The comparability condition identified in this article has advantages in consistently estimating the parameters of interest. However, it 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. The other major limitation is that the effect of time-independent or subject-specific covariates is not estimable, although its interaction with any time-dependent covariates are still estimable. Therefore, a more elaborated approach to analyzing the recurrence time data is in need when the effects of both time-dependent and -independent covariates on the distribution of durations are indeed of scientific interest at same time.

In §1, we discuss in brief about the sampling schemes of duration times in practice. Although the prevalence sampling scheme is more often encountered in practice than the incidence sampling scheme, the former sampling scheme entails more complicated probability structure than the latter one and thus more difficult to be analyzed. This article mainly focused on the duration-specific covariate effect AFTER certain initial event occurs, which is certainly of interest to be relaxed in the future research development.

There are also some other statistical issues left in this article not discussed. One important issue is the evaluation of efficiency of the proposed estimators. Although various types of weighting scheme can be applied to the proposed estimating equations, the existence of the optimal weights, and, if yes, the properties of such optimal weights, are not rigorously investigated. The other important issue is the model adequacy. In this article, the proportional reverse time hazards models are used. Therefore the proportionality between the reverse time hazard functions needs to be justified. In addition, compared with the accelerated failure time models in Wang & Chen (2000), the proportional reverse time model is not necessarily to be a better choice, although it has its own merits. A formal statistical procedure of selection may be of help to the practitioners with choosing the better one.

APPENDIX
Asymptotics

Martingale theory has been useful in developing asymptotic theory for the inference procedures of the Cox proportional hazards models (Andersen & Gill, 1982; Fleming & Harrington, 1991). 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.

As shown in §3.3, the estimating function in (4) 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

$$n^{1/2}(\hat{\beta} - \beta_0) = \left\{ n^{-1} \cdot \frac{\partial \tilde{S}(\hat{\beta}_0)}{\partial\beta} + n^{-1} \cdot \frac{1}{2} \cdot \frac{\partial^2 \tilde{S}(\hat{\beta}^*)}{\partial\beta^2} \cdot (\hat{\beta} - \beta_0) \right\}^{-1} \cdot \left\{ -n^{-1/2} \tilde{S}(\beta_0) \right\}. \quad (6)$$

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

Abelson, R. P. & Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics* **34**, 1347-1369.

Andersen, P. K., Borgan, Ø, Gill, R. D., & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Bhattacharya, P. K., Chernoff, H. & Yang, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Annals of Statistics* **11**, 505-514.

Brookmeyer, R. & Gail, M. H. (1987). Bias in prevalent cohort. *Biometrics* **43**, 739-749.

Brookmeyer, R. & Gail, M. H. (1994). AIDS Epidemiology. New York: Oxford University Press.

Chang, S.-H. (2000). A two-sample comparison for multiple ordered event data. *Biometrics* **56**, 183-189.

Chang, S.-H. & Wang, M.-C. (1999). Conditional regression analysis of recurrent time data. *Journal of the American Statistical Association* **94**, 1221-1230.

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

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society* Ser. B, **34**, 187-220.

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

Eaton, W. W., Bilker, W., Haro, J. M., et al. (1992a). Long-term course of hospitalization for schizophrenia: Part II. Change with passage of time. *Schizophrenia Bulletin* **18**, 229-241.

Eaton, W. W., Mortensen, P. B., Herrman, H. et al. (1992b). Long-term course of hospitalization for schizophrenia: Part I. Risk for hospitalization. *Schizophrenia Bulletin* **18**, 217-228.

Efron, B. & Petrosian, V. (1999). Nonparametric methods for doubly truncated data, *Journal of the American Statistical Association* **94**, 824-834.

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

Gross, S. T. & Huber-Carol, C. (1992). Regression models for truncated survival data. *Scandinavian Journal of Statistics* **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. *Journal of the American Statistical Association* **95**, 619-627.

Huang, Y. & Louis, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* **85**, 785-798.

Kalbfleisch, J. D. & Lawless, J. F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association* **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. *Statistica Sinica* **1**, 19-32.

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

Lawless, J. F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association* **82**, 808-815.

Lawless, J. F. & Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158-168.

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.

Lin, D. Y. & Ying, Z. (2001). Nonparametric tests for the gap time distributions of serial events based on censored data. *Biometrics* **57**, 369-375.

Lindsey, J. K. (1993). *Models for Repeated Measurements.* New York: Oxford University Press.

Mortensen, P. B. & Eaton, W. W. (1994). Predictors for readmission risk in schizophrenia. *Psychological Medicine* **24**, 223-232.

Pepe, M. S. & Cai, J. (1993). Some graphical display and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **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. *Annals of Statistics* **10**, 616-620.

WANG, M.-C. (1999). Gap time biases in incident and prevalent cohorts. *Statistica Sinica* **9**, 909-1010.

WANG, M.-C. & CHANG, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association* **94**, 146-153.

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. *Annals of Statistics* **14**, 1597-1605.

WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics* **12**, 163-177.

Table 1: A hypothetical example of biased riskset

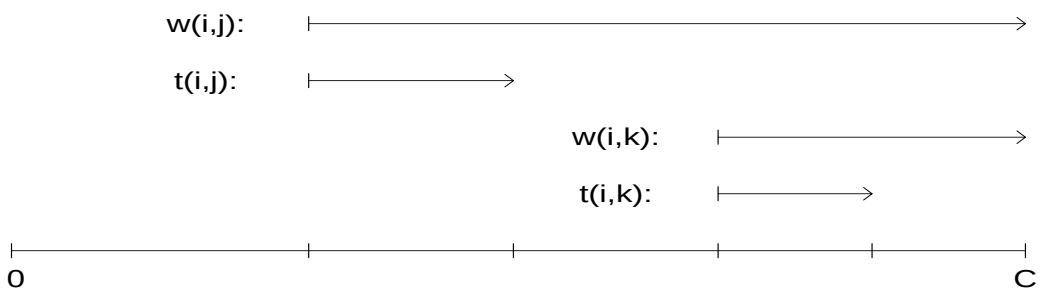
Duration j	1	2	3	4	5
Actual T_{ij}	1	5	2	6	...
Observed t_{ij}	1	5	2	1+	
$W_{ij} = C_i - \sum_{l<j} t_{ij}$	9	8	3	1	
R_{ij}		t_{i1}	t_{i1}		
		t_{i2}	t_{i3}		

Table 2: Summary of Simulation Studies

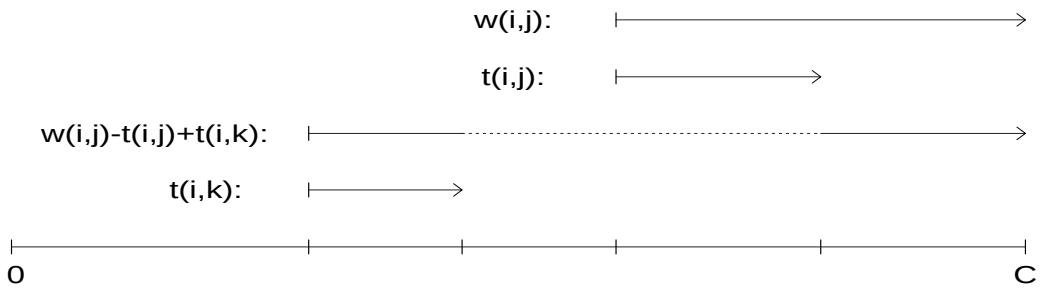
n	μ[a]	c[b]		(β_0^1, β_0^2) (0,0)		(1,0)		(1,0)		(1,1)	
				$\hat{\beta}_0^1$	$\hat{\beta}_0^2$	$\hat{\beta}_{01}$	$\hat{\beta}_0^2$	$\hat{\beta}_0^1$	$\hat{\beta}_{02}$	$\hat{\beta}_0^1$	$\hat{\beta}_0^2$
100	10.8		Bias[c]	-0.0060	-0.0033	0.0059	0.0061	-0.0083	-0.0048	-0.0006	0.0048
			Cov. Pr.[d]	0.9526	0.9472	0.9475	0.9529	0.9476	0.9503	0.9500	0.9506
	1.0		Bias	0.0023	0.0026	0.0042	0.0006	0.0029	-0.0010	0.0081	-0.0030
	ov. Pr.			0.9511	0.9473	0.9508	0.9498	0.9476	0.9403	0.9530	0.9522
	7	-0.0100		0.0039	-0.0077	0.0024	0.0115	-0.0055	-0.0072		
	8	0.9511		0.9483	0.9519	0.9507	0.9526	0.9494	0.9488		
	5	-0.0002		3.0029	-0.0018		-0.0010	0.0011			
	.9472	0.9474		0.9508	0.9489						
	-0.0053	-0.0022		-0.0075							
		0.9519		0.9514							
	83	0.0018									

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}



(a) k>j



(b) k<j