Event history analysis

Fetene B. Tekle and Jeroen K. Vermunt

Department of Methodology and Statistics

Tilburg University, The Netherlands

1. Introduction

In social and behavioral sciences in general and in psychology in particular, researchers are often interested in the occurrence of events such as the formation or ending of formal and informal relationships (e.g. marital unions, friendships, love relationships), the onset of and recovery from mental disorders, the entry into and exit from a job, the experience of stressful and pleasant life events (accidents, dying of a parent, being in love for the first time, etc.), and the transition across developmental stages. Mortality may also be the event of interest, though that is the more typical event in biomedical studies. Data on the occurrence of the event(s) of interest can either be collected using retrospective or prospective study designs, and will contain information on whether the event(s) of interest occurred to the individuals in the sample, and if so on the time of occurrence. In addition to information on the timing of the event(s) of interest, there will usually also be information on socio demographic covariates, risk factors, and/or the treatment or intervention received if there is any.

Event history data makes it possible to determine at what time periods the event of interest is most likely to occur, as well as to determine why some individuals experience the event earlier than others and why some do not experience the event of interest at all during the study period. Although event history data gives opportunities to answer such questions, they also pose certain challenges that cannot be dealt with using standard data analysis methods such as linear and logistic regression analysis (Tuma & Hannan, 1979; Allison, 1982;

Willett & Singer, 1993). More specifically, simple linear and logistic regression methods are not suited for dealing with two distinctive features of event history data; that is, with censoring and time-varying covariates. Censoring is a specific kind of missing data problem, namely that for some individuals it is not known when the event occurs because they did not experience the event during the observation period. Linear regression analysis of such censored data yields biased results and logistic regression analysis yields loss of information. Moreover, standard regression models lack a way to incorporate time-varying covariates, covariates that may change their value over time. In order to deal with censoring and time-varying covariates, we need special regression techniques which are known as event history models, hazard models, survival models, failure time models, and duration models.

The main distinction made in the field of event history analysis is between continuous-time methods (when the event time can take on any nonnegative value) and discrete-time methods (when the event time can take on a finite set of values). In this chapter we will focus on discrete-time techniques as a part of a course for graduate level students or as a reference for applied researchers. Although continuous-time methods are predominant in the statistical literature (Blossfeld & Rohwer, 1995; Collette, 2003; Vermunt, 1997), discrete-time methods are the more commonly used ones in psychological research as well as in other social and behavioral sciences, not only because they are conceptually simpler but also because one will seldom have real continuous-time data. Sometimes events can only occur at regular discrete time points (e.g., weekly, monthly or yearly), whereas in other situations events can occur in continuous time, but the measurement yields discrete-time data, for example, when a survey asks the age or year of the formation of a relationship, marriage, or divorce instead of the exact date. In both situations it is more appropriate to use discrete-time methods instead of methods that are developed for continuous event time. Even if the measurement scale of the event time is continuous, discrete-time techniques can be used to

approximate the results that would be obtained with continuous-time methods (Yamaguchi, 1991; Vermunt, 1997). Additionally, discrete-time techniques are computationally and conceptually simpler and thus easier to understand by social and behavioral scientists, and they can serve as a good starting point for understanding the more advanced continuous-time methods.

The remainder of this chapter is organized as follows. In the next section, an empirical example that will be used throughout this chapter is introduced. Some basic terminologies of event history analysis are presented in section 3. Section 4 explains why special regression techniques are needed for event history analysis. Section 5 presents the statistical concepts used for describing event time distributions – the hazard and survival functions – and shows how a grouped data method similar to the actuarial method can be used to estimate these functions. Section 6 deals with regression models for discrete-time event history data in which the hazard rate – or the probability of event occurrence at a particular time point – is related to covariates. Concluding remarks are given in section 7.

2. An empirical example

Throughout this chapter, a real-life example is used to illustrate the concepts and modeling approaches for event-history data. This example is introduced below.

Example: Adolescents' Relationship

The example is about adolescent's first experience with relationships. The data are taken from a small-scale survey of 145 adolescents in the Netherlands (Vinken, 1998). Vermunt (2002) used latent class analysis to construct a typology based on four events related to adolescent's first experience with relationships: 'sleeping with someone', 'going out', 'having a steady friend', and 'being very much in love'. Here, we will use the event

'sleeping with someone for the first time' to illustrate the methods of event history analysis discussed in this chapter. Besides information on the occurrence of the four events, binary time-constant (i.e., fixed) covariates, youth centrism (YC), gender (G), and education (E) are available. Youth centrism is a measure for the extent to which young people perceive their peers as a positive valued ingroup and perceive adults as a negatively valued outgroup. The dichotomous youth-centrism scale that is used here was constructed by Vinken (1998).

3. State, event, duration, risk period, risk set and censoring

In order to understand the nature of event history data and the purpose of event history analysis, it is important to understand the following concepts: state, event, risk period, and censoring (Yamaguchi, 1991). These concepts are illustrated below using the example introduced in the previous section.

The first step in an event history analysis is to define the discrete states that one wishes to distinguish. *States* are the categories of the variable, the dynamics of which one wishes to explain. At every particular point in time, each person occupies exactly one state. In our first experience with relationships example, each adolescent is either in the state 'never slept with someone' or 'has slept with someone'. An *event* is a transition from one state to another, that is, from an origin state to a destination state. In our example, the event sleeping with someone for the first time is the transition from the state 'newer slept with someone' to the state 'has slept with someone'. It is clear that in our application the event of interest can occur only once because it is not possible to exit the destination state (this is called an absorbing state). In other applications, the event(s) of interest may occur several times (this called recurrent events), such as the recovery from a depression, which is the transition between the states depressed and non-depressed.

Another important concept is the risk period. Clearly, not all persons can experience each of the events under study at every point in time. To be able to experience a particular event, one must first occupy the original state, that is, one must be at risk of the event concerned. The period that someone is at risk of a particular event – or exposed to a particular risk – is called the *risk period*. Usually it is straightforward to identify the persons at risk of the event, such as in our relationships example in which adolescents that have never slept with someone at a particular age are at risk of experiencing the event of sleeping with someone for the first time at that age. The risk period(s) for a recovery from depression are the period(s) that a subject stayed in the origin state depressed. A strongly related concept is the *risk set*. The risk set at a particular point in time is formed by all subjects who are at risk of experiencing the event concerned at that point in time.

Using these concepts, event history analysis can be defined as the analysis of the duration of the nonoccurrence of an event during the risk period. When the event of interest is 'sleeping with someone for the first time', the analysis concerns the duration of nonoccurrence of the experience of sleeping with someone, in other words, the time that adolescents remained in the state 'never slept with someone'. In practice, as will be demonstrated below, the dependent variable in an event history model is not duration or time itself but a transition probability or hazard rate. Therefore, event history modeling can also be defined as the analysis of the probability (or rate) of occurrence of the event of interest during the risk period. In the relationships example this concerns an adolescent's probability of sleeping with someone given that this did not happen before.

As was already indicated above, an issue that always receives a great amount of attention in discussions on event history analysis is censoring, where a distinction should be made between left and right censoring. These two forms of censoring refer to missing information on the time of nonoccurrence of the event of interest *before* and *after* the

observation (or follow-up) period, respectively. Here, we consider only the more common right-censoring problem, and we refer interested readers to Kalbfleisch and Prentice (2002), Tuma & Hannan (1979), Vermunt (2007), and Yamaguchi (1991) among others for discussions of alternative censoring mechanisms and their implications.

An observation is called (right) censored if it is known that it did not experience the event of interest during a certain amount of time (during follow-up period), but the exact time at which it experienced the event is unknown. In the recovery from depression example, a censored observation would be an individual who was in the depressed state at the end of the study or who dropped out from the study. For such a person, we know the duration of the depression till that moment, but not whether or when he or she will recover from depression, which means that the duration of nonoccurrence of recovery is only partially observed for such a person. In the relationships example, a censored observation would be an adolescent who has not experienced the event 'sleeping with someone' before the age at which the survey took place. This partial information is called the censoring time.

More formally, let T be the event time and U the censoring time. The duration of nonoccurrence of an event that can actually be observed is $Y = \min(T, U)$; that is, we observe the true event time when it is smaller than the censoring time and vice versa. Methods for event history analysis define a model for the dependent variable Y (and thus not for T). However, because it is also relevant to make a distinction between event and censoring times, an event indicator variable has to be defined. For right-censored event data, the event indicator for ith person is defined as

$$Event_i = \begin{cases} 1 \text{ if } T_i \leq U_i \\ 0 \text{ if } T_i > U_i \end{cases}.$$

In other words, $Event_i$ equal 1 if we observe the event time and 0 if the observation is censored.

While traditional regression methodology such as linear or logistic regression analysis does not provide a way of simultaneously analyzing observed and censored event times, as we will show later, event-history analysis methodology provides a way of considering both simultaneously. In the next section, we will explain when these event-history analysis methods are more appealing in relation to research problems in practice.

4. When to use event history analysis?

To determine whether the method of event-history analysis is applicable in a specific situation, one has to examine the research problem/question and a study's methodological features. A research's method of analysis calls for event-history analysis if the research's question is centered on whether and, if so, when events occur. Data can be collected prospectively or retrospectively, over a short or a long period of time, in an experiment or an observational study. The beginning time which is an initial starting point when no one under study has yet experienced the target event but everybody is in the risk set has to be identified. The time of a target event whose occurrence is being studied can be measured in years, months, days, or minutes; however, a meaningful scale needs to be chosen. For example, in the relationships example, the research question is when young adolescents have their first experience of sleeping with someone and whether predictors affect the timing of this event? Clearly, the target event is the transition to sleeping with someone for the first time. The beginning time is the time at which none of the subjects under study has experienced the event but in the risk set of the event sleeping with someone. Since data is collected retrospectively, it is not practical to precisely measure the time of the first experience of sleeping with someone in months, days, or smaller grids of time. During the data collection subjects may recall the period of the event in terms of the age at which the event happened.

Thus, it is logical to consider age in years as the unit of scale for event period in the relationships data example.

5. Describing event time distribution

5.1. Discrete versus continuous time

The manner in which the basic statistical concepts of event history analysis are defined depends on whether the time variable *T*, indicating the duration of nonoccurrence of an event, is assumed to be continuous or discrete. Of course, it seems logical to assume *T* to be a continuous variable in the sense that the event of interest may occur literally at any time defined on (0, ∞). However, in many situations this assumption is not realistic for two reasons. Firstly, in many cases, *T* is not measured accurately enough to be treated as strictly continuous. Respondents can usually give dates and times only in ranges or round numbers, even if encouraged by interviewers to be more precise. Secondly, the events of interest can sometimes only occur at finite particular points in time which are discrete (taking a finite set of values, example, $t_1, t_2, \cdots, t_L$).

Regardless of the assumption whether *T* is a discrete or continuous variable, the main aim of event history analysis is characterizing the probability distribution of the random variable *T*, the duration of nonoccurrence of an event. An additional objective is typically to gain an understanding on how risk factors and covariates affect the event times. This second objective can be addressed by modeling the probability distribution of *T* in terms of potentially explanatory variables. Even though these objectives are common to any event history analysis, the way the statistical methods are formulated depends on whether the measurement of the time variable is assumed to be discrete or continuous. The methods and discussions considered below in this chapter assume time is measured on discrete scale. We describe in the next subsection possible ways of characterizing the probability distribution of

T and give modeling strategies of the probability distribution of *T* in terms of possible explanatory variables in section 6.

5.2. Discrete Event Time distributions

Let *T* be a discrete random variable indicating the event time, and t_l the *l*th discrete time point, with $l = 1, 2,...$ and $0 < t_1 < t_2 < \cdots$. There are three equivalent ways to characterize the probability distribution of the event time *T*. The simplest is as $f(t_l) = \Pr(T = t_l)$, or as the probability of experiencing an event at $T = t_l$. Another possibility is via the survival function $S(t_l)$, which is the probability of not having the event before and in time interval t_l or, equivalently, the probability of having the event after t_l. It is defined as follows:

$$S(t_l) = \Pr(T > t_l) = 1 - \Pr(T \le t_l), \qquad l = 1, 2, \cdots$$

Another option is to use the discrete-time hazard probability $h(t_l)$, which is the conditional probability that the event occurs at $T = t_l$ given that it did not occur prior to $T = t_l$ (given $S(t_{l-1})$). Mathematically, the hazard is given as

$$h(t_l) = \Pr(T = t_l \mid T \ge t_l) = \Pr(T = t_l \mid T > t_{l-1}) = \frac{\Pr(T = t_l)}{\Pr(T > t_{l-1})} = \frac{f(t_l)}{S(t_{l-1})}. \qquad (1)$$

What is important is that both $f(t_l)$ and $S(t_l)$ can be expressed in terms of $h(t_l)$. Using the fact that $f(t_l) = S(t_{l-1}) - S(t_l)$, equation (1) can be rewritten as

$$h(t_l) = \frac{S(t_{l-1}) - S(t_l)}{S(t_{l-1})} = 1 - \frac{S(t_l)}{S(t_{l-1})}.$$

By rearranging this last equation, we obtain

$$S(t_l) = S(t_{l-1})[1 - h(t_l)] \qquad (2)$$

Using $S(t_0) = 1$, no individual experienced an event before and in $T = t_0$, this last equation leads to the required expression that

$$S(t_l) = \prod_{k=1}^{l} [1 - h(t_k)] .$$

where $\prod$ is a product sign and the term in [bracket] is a complementary of the hazard function , which is the conditional probability that the event occurs at time t_k given that it did not occur prior to t_k . The equation implies that the survival probability to the end of *l*th time is the product of survival probabilities at each of earlier time points.

By using equation (2) and (1), the following expression is also obtained for the probability of experiencing an event at time t_l, $f(t_l)$:

$$f(t_l) = h(t_l) \prod_{k=1}^{l-1} [1 - h(t_k)]. \qquad (3)$$

5.3. Estimating Event Time Distribution

The grouped-data or life-table method (Merrell, 1947 & Cox, 1972) and the Kaplan-Meier (Kaplan & Meier, 1958) estimator are two descriptive methods for estimating the event-time distribution from a sample. Below, we discuss the grouped-data method for discrete event times.

Grouped-data or life-table method for discrete event times

The most straightforward way to describe the event history in a sample is to construct grouped data by merging event times in groups or intervals. This method is more known as the life-table method. The life-table method enables the computation of nonparametric estimates of the survival and hazard functions in separate intervals over time. The distribution

of event times is divided into a certain number of intervals. For each interval we can then identify the number of subjects entering the respective interval without having experienced the event, the number of cases that experienced the event in the respective interval, and the number of cases that were lost or censored in the respective interval. Based on those numbers, several additional statistics can be computed. Some of these statistics are:

- **Number of Cases at Risk (risk set r_l)**: This is the number of subjects who are at risk of experiencing the event of interest within the specific interval. This number is the number of cases that entered the respective interval.
- **Proportion of cases that experience the event (hazard h_l):** This proportion is computed as the ratio of the number of cases experiencing the event within the interval divided by the number of cases at risk in the interval.
- **Cumulative Proportion Surviving (Survival Function S_l):** This is the cumulative proportion of cases surviving up to the end of respective interval.
- **Median Survival Time:** This is the survival time at which the cumulative survival function is equal to *0.5*.

5.4. Example: Life table method for Adolescents' relationship example

Table 1 shows a life table for the data on adolescent's experience on sleeping with someone example introduced in section 2. For 142 cases we have information on whether the event sleeping with someone had happened yes or no, and if when it happened. It turns out that 90.1% of these adolescents experienced the event before the time of data collection.

A natural definition of the beginning time (t_0) for an analysis of this dataset is an age at which none of the adolescents experienced the event of interested, or equivalently, an age

at which all subjects are at risk of experiencing the event. It does, however, not make sense to start at age 0 because the youngest age at which the event happened is 12, which means that the hazard rate is 0 and the survival probability is 1 for all ages prior to 12. Without loss of information we can therefore use age 11 as the beginning time for the event history analysis. By dividing into a series of rows indexing the age intervals (column 1), a life table for the relationships example in Table 1 contains information on the number of adolescents who: entered the age interval (column 2), censored during the age interval (column 3), and experienced the event 'sleeping with someone' during the age interval (column 4). The age intervals in Table 1 partition the times of the event occurrence in such way that each interval contains a range of ages that include the initial time and excludes the concluding time. The width of the intervals is set to 1 year for ease of presentation. Conventional mathematical notation [brackets] denotes inclusions and (parenthesis) denote exclusions. Thus, bracket is used for each interval's initial time while parenthesis is used for the concluding time. In total there are 14 age intervals with each having 1 year length, [11, 12), [12, 13), …, [24, 25). In general, the definition of the time intervals of a life table should be based on a relevant time unit and respect the way events occur. Data whose time unit is days, weeks, or months may require wider intervals compared to data whose time unit is years, but the grouping should always be such that it yields a series of time intervals $[t_0, t_1)$, $[t_1, t_2)$, …, $[t_{l-1}, t_l)$, $[t_l, t_{l+1})$, …, and so on. No events occur during the 0^{th} interval, which begins at time t_0 and ends just before t_1. This interval represents what is called the beginning of time. Any event occurring at t_1 or later but before t_2, is classified as an event happening during the first time interval $[t_1, t_2)$. The lth time interval, $[t_l, t_{l+1})$, begins immediately at time t_l and ends just before time t_{l+1}.

The next column in the life table contains the number of adolescents who enter each successive age interval without experiencing the event or censoring in the previous intervals. This number is the risk set for the discrete-time life-table description. As shown in column 3

of Table 1, censoring happens for some cases in the last two age groups. For these cases, the event sleeping with someone has not happened at before the time of data collection. The only known information is that those adolescents have not experienced the event until ages 24 (in the age interval [23, 24)) and 25 ((in the age interval [24, 25)), respectively. Column 4 contains the number of events that occured in each age interval. As can be seen, the risk set or the number in column 2 for each interval is the risk set of previous interval minus the sum of the numbers in column 3 and 4 of previous interval.

<<< Insert Table 1 about here >>>>>>>>>

Column 5 in Table 1 shows the proportion of each interval's risk set that experiences the event during the interval. Thus, these are conditional probabilities (or hazards) that the event occurs to an individual in *l*th interval given that the event did not occur to that person before the *l*th interval. In fact, the proportions in column 6 are maximum likelihood estimates of the discrete-time hazard given in equation (1) (Singer & Willett, 1993, 2003). That means, we estimate the discrete-time hazard for each interval in the life table by the proportion of the risk set that experiences the event within the interval. These proportions are mathematically obtained by dividing the numbers in column 4 to the numbers in column 2, i. e,.

$$\hat{h}(t_l) = \frac{\text{number of events in interval } l}{\text{number at risk in interval } l}, \tag{4}$$

where $\hat{h}(t_l)$ is an estimate of the discrete hazard probability at time *t* in interval *l*.

Hence the estimate of hazard is $\hat{h}(t_1) = 0.0070$ for a time in the first interval, the estimate is zero for the second interval as none of the adolescents experienced the event in the

second interval, $\hat{h}(t_3) = 0.0142$ and so on. Note that the discrete-time hazard is a probability, which implies that its value always lies between 0 and 1. A helpful way of examining the time-dependence of the hazard probabilities is to graph their values over time. The left panel of Figure 1 plots the estimated hazard function based on the proportions in column 5 of Table 1. The risk of event occurrence (sleeping with someone for the first time) among the adolescents at ages below 16 is small. The risk in general increases with time starting from age 16 until age 19 where it drops suddenly. It again increases in the interval [20, 21) and starts to decrease thereafter. Finally, there is an increase from starting from age 23. In general the 'risky' time periods for experiencing the event are from age 16 with a high peak at age 20.

The estimated hazard probabilities at each time interval describe the distribution of event occurrence for a random sample of individuals from a homogenous population. In section 6, we show how individual difference in the hazard probabilities can be investigated using regression models including predictor variables (e.g., gender and education level).

<<<<< Insert Figure 1 about here >>>>>>>>

The proportion of adolescents who has not slept with someone till the end of each time interval (who survived) is shown in column 6 of Table 1. This proportion is an estimate of the survival function given in equation (2). Since no one has experienced the event before age 12, the estimate of survival function for the 0th interval, $\hat{S}(t_0)$, is 1. The estimate for the 1st interval is then the product of the survival function of the 0th interval and the probability of surviving (not having the event) during the 1st interval. This latter probability is just the complement of the hazard probability in the 1st interval, $1 - \hat{h}(t_1) = 1 - 0.0070$. Thus, the estimate of the survival function for the 1st interval is

$\hat{S}(t_1) = \hat{S}(t_0)[1 - \hat{h}(t_1)] = 1 \times [1 - 0.0070] = 0.9930$, implying that 99.3% of the adolescents did

not experience the event until the end of the 1st interval. In general the estimate of the hazard function for *l*th interval is

$$\hat{S}(t_l) = \hat{S}(t_{l-1})[1 - \hat{h}(t_l)]. \quad (5)$$

In general, the survivor function over time declines to 0 which is the lower bound for the survival probability. A useful way to examine the survival function is again to graph the estimates of survival function over time. The right panel of Figure 1 graphs the survival function based on the estimates in column 6 of Table 1. Unlike the hazard function which can increase, decrease or remain constant over time, the survivor function never increases. For intervals with no events occurring (for example interval [13, 14)), the survivor function remains steady at the value of previous interval. The survivor function drops rapidly in those periods where the hazard is high and the survivor function declines slowly at the time periods with low hazard.

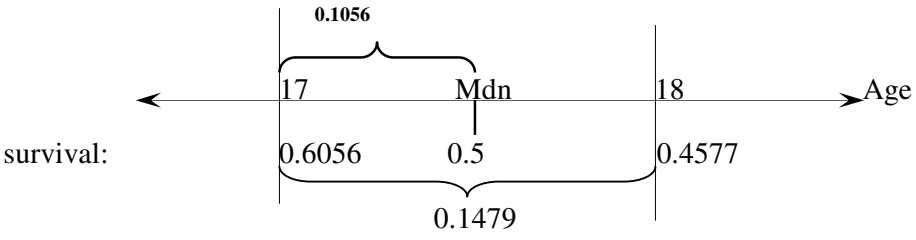
The life-table estimates of the discrete-time hazard and survival functions yield two alternative descriptions of the event time's distribution. The interest is usually also on the summary statistics or center of the distribution. If there was no censoring, the center of the event time distribution could be estimated using sample mean. However, due to censoring the event time is not known for all individuals and the sample mean cannot be used as an estimate of the center of the distribution. Instead, another measure of central tendency is often used in event history analysis: the median survival time. As a measure of center of the distribution, the estimated median survival time is the value of time (here age) for which the value of the estimated survival function is 0.5. Note that in general, the median survival time or the 50th percentile for the survival function is not necessarily the same as the point in time up to which 50% of the sample survived without the event. The median survival time

corresponds to the point in time up to which 50% of the sample survived without the event only if there were no censored observations prior to this time (which is the case in the adolescents' relationship example).

A closer look at the estimate of the survival function in column 7 shows that 0.6056 of the adolescents survived to the end of the interval [17, 18) and the proportion drops below 0.5 to 0.4577 in the interval [18, 19). Thus, the median survival time could be reported as age between 17 and 18. If needed, the estimate of median survival time can be more accurately obtained by interpolation between the two intervals that have survival estimates close to 0.5 on both sides at the top and bottom. In the current adolescents' relationship example, the two intervals are [17, 18) and [18, 19). Let t_m be the initial time for the interval when the sample survivor function is just above 0.5 (in our example, age 17), let $\hat{S}(t_m)$ represent the value of the sample survivor function of that interval, let t_{m+1} and $\hat{S}(t_{m+1})$ be the initial time and sample survivor function, respectively, for the following interval (when the survivor function is just below 0.5). Then, following linear interpolation, the estimated median survival time is

$$\hat{Mdn} = t_m + \left[\frac{\hat{S}(t_m) - 0.5}{\hat{S}(t_m) - \hat{S}(t_{m+1})}\right](t_{m+1} - t_m) \qquad (6)$$

For the current example the interpolation is illustrated using the following simple sketch.



Mdn = 17 + [(0.6056 – 0.5) / (0.6056 – 0.4577)] (18-17)

= 17 + (0.1056) / (0.1479) = 17.7.

Note that in our application there are no censored cases for the intervals prior to [18, 19) which contains the median survival time. As a result, the median survival time 17.7 is also the age at which 50% of the sample units experienced the event.

Hazard functions are the most useful tools in describing patterns of event occurrence as they are sensitive to the unique risk associated with each time period while the survivor functions accumulate information across time periods. Thus, by examining the variation over time in the magnitude of the hazard function, we identify when events are likely, or unlikely, to occur. The median survival time identifies the center of the distribution and it tells little about the distribution of event times. It is an 'average' event time but relatively less sensitive to the extreme values.

The life-table method for estimating hazard and survival functions using discrete-time event data is similar to the actuarial method for estimating these functions using continuous-time event data. These methods differ only in the assumption about the occurrence of events and censorings within intervals. The discrete-time method assumes events and censorings occur at the endpoint of the time interval, which means that all those who entered the interval are in the risk set throughout the interval. The actuarial method, however, assumes that both events and censorings are distributed equally throughout the interval, which means that the risk set changes during an interval. In the actuarial method, the risk set is estimated as the average number of person at risk during the interval, which leads to slight modifications of equations (4) and (5) for this method. More details about the actuarial method are given elsewhere (Singer & Willett, 2003 (pp. 480); Mould, 1976; and Cutler & Ederer, 1958. among others).

The description and estimation of the event time distributions discussed so far assumes a single homogeneous group of subjects. However, in practice researchers are also

interested in identifying factors that affect the occurrence and timing of the event of interest. For example, in intervention or experimental studies, the interest is to estimate the effects of the intervention on the probability of occurrence of the event over time. Therefore, we move from the discussion of method for describing event time distribution to regression models for which allow including predictors such as socio demographic covariates and experimental conditions. The objective is to determine the relationship between those predictors and the likelihood of event occurrence. The specific modeling approach again depends on whether the event time is treated as discrete or continuous. Thus, we pursue the discussion below with the assumption of discrete event times and we refer readers to Blossfeld & Rohwer (1995), Collette(2003), and Vermunt (1997) among others for well developed methodologies on continuous event time models. Regression models for discrete event time data are better known in literature as discrete-time survival models (DTSM). Even though originally discrete- and continuous-time models were formulated separately by Cox (Cox, 1972), asymptotically the two models are equivalent. That is, as the discrete time interval get smaller and smaller, a DTSM becomes more and more similar to a continuous-time model (Thompson, 1977; Allison, 1982; D'agostino et al., 1990; Yamaguchi, 1991; Peterson, 1991).

6. Discrete Time Event History Models

Cox was the first to propose models for censored data (Cox, 1972). He proposed discrete-time models for tied event data alongside his formulation of proportional hazard modeling for continuous-time data. A tie refers to the situation where several subjects experience the event at the same specific time, which is something that in a strict continuous-time framework should not occur. When there are many ties in the data, a discrete-time model is more appropriate. The discrete-time modeling approach involves using a logistic regression model

for a person-period dataset. For all time points until a person either experiences the event or is censored, the dependent variable is an indicator of whether or not an individual experienced the event at that time point (Allison, 1982; Singer & Willett, 1993).

As discussed in the previous section, in contrast to the other distributions for event history data, the hazard function is the most important element in event history analysis because of at least three reasons. First, it shows the risk of event occurrence at each time period, it tells us whether and when an event is likely to occur. Second, the event history analysis is able to deal with censored cases because the hazard always includes both censored and non-censored cases. Third, the sample survival function cannot be computed directly for a given time point when there is censoring, but the survival function can be estimated indirectly from the hazard function. In general, the mathematical relationships described earlier between the distributions of event history data can be used to obtain estimates of other distributions ($S(t_l)$ and $f(t_l)$) if the estimate of hazard is known. By using the hazard $h(t_l)$ as the left-hand side variable in a regression model, one can relate hazard distribution to the covariates of interest. Because hazards are (conditional) probabilities bounded by 0 and 1, they can be transformed using logit link function so that the transformed hazard is unbounded and can easily be regressed on covariates and time variables. That means, instead of modeling the hazard probability directly the logit of the hazard probability is used as the left-hand side variable in a (generalized) linear model. The logit hazard at time $T = t_l$ for a person with covariate value *X* is given by:

$$\text{logit}[h_X(t_l)] = Log_e\left[\frac{h_X(t_l)}{1-h_X(t_l)}\right] = \alpha_l D_l + \beta X \ , \qquad (7)$$

where $Log_e[c]$ is the natural logarithm of *c*, $h_X(t_l)$ is the hazard or conditional event probability at time t_l for a person with covariate value X, D_l is a dummy variable indicating

the time period ($D_l = 1$, if $T = t_l$ and 0 otherwise), α_l is the intercept parameter at time point $T = t_l$ representing the logit hazard when $X = 0$ (baseline logit hazard at time t_l), and β is the slope parameter that shows the effect of the covariate X on the logit hazard. When there are more covariates, model (7) can be extended by including the covariates and corresponding β parameters at the right-hand side of the equation. That means, for a person with covariate values $X_1, X_2, \cdots, X_p$ the logit hazard is

$$\text{logit}[h_{X_p}(t_l)] = Log_e\left[\frac{h_{X_p}(t_l)}{1-h_{X_p}(t_l)}\right] = \alpha_l D_l + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \alpha_l + \sum_{k=1}^{p} \beta_k X_k , \quad (8)$$

where $h_{X_p}(t_l)$ is the hazard or conditional probability at time t_l for a person with p covariate values $X_1, X_2, \cdots, X_p$ and β_k is the parameter that shows the effect of the covariate X_k on the logit hazard controlling for other covariates in the model, $k = 1, \cdots, p$.

Equation (8) is a discrete-time event history model or DTSM. When the value of all covariates equals, the logit hazard in (8) reduces to the baseline logit hazard

$$\text{logit}(h_0(t_l)) = Log_e\left[\frac{h_0(t_l)}{1-h_0(t_l)}\right] = \alpha_l D_l , \quad (9)$$

where $h_0(t_l)$ is the baseline hazard or conditional probability with covariate values equal to zero at time t_l .

Note that $\frac{h_{X_p}(t_l)}{1-h_{X_p}(t_l)}$ is the odds of event occurrence at time t_l and the model in equation (8) represents the log-odds of event occurrence as a function of time period and covariates. By using an inverse function of the natural logarithm, exponential function, in equation (8) the odds of an event can be expressed as

$$\frac{h_{X_p}(t_l)}{1-h_{X_p}(t_l)} = \exp(\alpha_l D_l)\exp\left(\sum_{k=1}^{p}\beta_k X_k\right) \quad (10)$$

By using equation (9), this last equation can be rewritten as

$$\frac{h_{X_p}(t_l)}{1-h_{X_p}(t_l)} = \frac{h_0(t_l)}{1-h_0(t_l)}\exp\left(\sum_{k=1}^{p}\beta_k X_k\right). \quad (11)$$

Thus, the discrete-time event history model implies that the odds of having the event at each discrete time point are $\exp\left(\sum_{k=1}^{p}\beta_k X_k\right)$ times for subjects characterized by covariate values $X_1, X_2, \cdots, X_p$ compared to subjects in the baseline group (subjects characterized by covariate values $X_1 = 0, X_2 = 0, \cdots, X_p = 0$). The model also implies that, controlling for other covariates, an increase in one unit of X_k increases (or decreases depending on the sign of β_k) the odds of having the event $\exp(\beta_k)$ times. When the covariate effects are time-independent (that is, when there are no time-covariate interactions), the model is a proportional odds model. That means, the ratio of odds of having an event among a group characterized by particular covariate values and the baseline group, $\frac{h_{X_p}(t_l)}{1-h_{X_p}(t_l)} / \frac{h_0(t_l)}{1-h_0(t_l)}$, is constant over time. The proportional odds model is similar to the Cox proportional hazards model for continuous event times with respect to this property. While the ratio of odds is time-constant for the proportional odds model, the ratio of hazard rates is time-constant for the Cox's proportional hazards model. As discussed in more detail below, the proportional odds assumption can be tested and relaxed by including interaction effects of time and covariates in the model.

An advantage of the discrete-time event history model compared with the continuous time event history model is that we can use the usual logistic regression options of most

available computer programs for the estimation of the parameters. The structure of the input data, however, differs between the usual logistic regression analysis and the use of logistic regression for discrete-time event history analysis. While the former uses one observation for each sample subject, the latter uses multiple observations for each subject. Accordingly the event history data for the logistic regression must be arranged in a specific way as described in the next subsection.

6.1. Construction of input data

The discrete-event time model described above uses a logistic regression approach. However, unlike standard logistic regression analysis, such an analysis it is not based on a person-oriented dataset, but instead it requires a person-period dataset in which each person may have a different number of records depending on the duration or stay in the risk set. In a typical person-oriented dataset, each person has one record (case) of data. Because researchers often keep event history data in a person oriented dataset format, conversion to a person-period dataset that contains for each person as many records as the time (period) he or she stays in the risk set without experiencing the event or censoring is needed. Table 2 illustrates such a conversion using 3 subjects from the adolescents' relationships data. The smallest event time for the adolescents' relationship data is 12 and the maximum is 24. Among the three adolescents whose data on some of the variables are shown in Table 2, the event times are known for the first and third (ID 2 and 12 reported event time at age 15). An adolescent with ID 9 is censored, i.e., event has not occurred during data collection and his age is 24. In the converted person-period dataset, subjects have different number of records depending on how long they stay before they experience the event or censoring. The period variable is included in the dataset to indicate the time to which the corresponding record

refers for each adolescent from age 12 and above. Since the first event occurs in the dataset at age 12, interest on event history is restricted to ages 12 and above. The period variable takes on the values 12, 13, 14 and 15 for each of adolescent with ID 2 and 12, to indicate that these four records describe their corresponding status in the periods from age 12 until the occurrence of the event at age 15. For the adolescent with ID 9, the variable period takes on the values 12 through 24, to indicate that those are the ages represented in the 13 records. The set of dummy variables D_{12} through D_{24} are also created to represent each time period in the logistic regression model for the discrete time event history model. The dummy variable $D_l = 1, \text{if period} = t_l$ and 0 otherwise. A dichotomous event indicator is created for the occurrence of 'sleeping with someone for the first time' to indicate whether a person experiences the event during the time period concerned (0 = no event, 1 = event). For each individual, the event indicator is 0 in every record except the last. Noncensored adolescents experience the event in their last period, so the variable event takes on the value 1 in that last period as shown for ID 2 and 12. Censored adolescents never experience at the periods shown in the data, so the variable event remains 0 throughout the records as shown for ID 9. Values of the time-constant covariates are repeated in each time period. Only values for the covariate gender are shown in Table 2 due to space limitation. SPSS and SAS® syntaxes to convert the person-oriented dataset to person-period dataset are given in the appendix.

The person-period dataset contains all information on survival time, including the information for censored observations. Once a person-period dataset is created, existing procedures in general statistical packages can be directly used for event history analysis without any modifications for censoring. As a descriptive analysis, a cross-tabulation of the variables period and event which can be obtained using CROSSTABS procedure of SPSS (PASAW statistics) or SAS procedure FREQ (SAS Institute) gives estimates of hazard or conditional probability at each time period as shown in column 5 of Table 1.

<<<<< Insert Table 2 about here >>>>>>>>>>

6.2. Maximum likelihood estimates

The likelihood function (Myung, 2003) for the discrete event time model expresses the probability that we would observe the specific pattern of event occurrence actually observed over time. The likelihood function is composed of as many terms as there are records in the person-period dataset. The probability that *i*th individual experiences the event at time period t_l, given that the event has not occurred before, is $h_i(t_l)$. During the time period when individual *i* experiences the event (the variable event =1 in the person-period dataset), he or she contributes $h_i(t_l)$ to the likelihood function. The probability that *i*th individual does not experience the event in time period t_l, given that the event has not occurred before, is $[1-h_i(t_l)]$. During the time periods when individual *i* does not experience the event (when the variable event = 0), he or she contributes $[1-h_i(t_l)]$ to the likelihood function. The contributions are mathematically expressed using the likelihood function as

$$\text{Likelihood} = \prod_{i=1}^{v}\prod_{l=1}^{L_i} h_i(t_l)^{event_{il}}[1-h_i(t_l)]^{(1-event_{il})}, \tag{12}$$

where $\prod$ is a product sign, *n* is the total number of subjects (sample size), L_i is the number of records for *i*th person in the person-period dataset and $event_{il}$ is the value on the variable event (either 0 or 1) for *i*th person at *l*th period. Note that the two product signs ensure that the likelihood function multiplies the contributions of each record in the person-period dataset across all individuals via the first product sign and across all time periods for each individual via the second product sign. Since event is coded with either 0 or 1, only one of the

two terms ($h_i(t_l)$ or $[1-h_i(t_l)]$) contribute to the likelihood function at each record in person-period dataset. In time period when the event does occur, only the first term remains while the second term becomes 1. In time periods when the event does not occur, only the second term remains while the first term becomes 1.

By using the notation $h_{X_p}(t_l)$ for the conditional probability of an ith individual with *p* covariates from equation (8) in equation (12) instead of $h_i(t_l)$, the likelihood function can be written as a function of unknown parameters α's for the baseline hazards and β's for the effects of the covariates. The objective of maximum likelihood estimation is to find estimates of the parameters that maximize the likelihood function. In practice the logarithm of the likelihood (log-likelihood) function is used as it is mathematically more tractable. Thus, the values of α's and β's that maximize the log-likelihood function should be obtained. In fact the routines for logistic regression model available in most standard statistical packages (e.g., SPSS, SAS, Stata, etc) provide estimates of the parameters of the discrete event time model that maximize the log-likelihood function when a proper person-period dataset is used. Thus, statistical routines of logistic regression to regress the event indicator variable on the dummy variables for time indicators and the selected *p* covariates in the person-period dataset can be used to obtain the maximum likelihood estimates of the parameters in the discrete event time model.

6.3. Example: Discrete event time model for Adolescents' relationship data

To illustrate the procedures of fitting, interpreting and testing statistical statements (hypothesis) for the discrete event time model, we use the adolescents' relationship data example. The person-period dataset with records of 1093 for the event of 'sleeping with

someone' for the first time and the covariates youth-centrism (YC), gender (G), education (E) and dummy variable D_l as indicator of time period t_l (see section 6.1) is used to fit the following models:

Model 1: $\text{logit}(h_0(t_l)) = \alpha_l D_l$

Model 2: $\text{logit}(h_X(t_l)) = \alpha_l D_l + \beta_1 \text{YC}$

Model 3: $\text{logit}(h_X(t_l)) = \alpha_l D_l \quad + \beta_2 \text{G} + \beta_3 \text{E}$

Model 4: $\text{logit}(h_X(t_l)) = \alpha_l D_l + \beta_1 \text{YC} + \beta_2 \text{G} + \beta_3 \text{E}$

Model 1 contains only the time periods and it describes the hazard profile over time. The model is estimated using the dummy variables for time with a no intercept option of logistic model in standard software. Model 1 helps to find the risk of event occurrence at each time period and identify the important periods (ages) where the event is common among the adolescents. The parameters α's for each time period are the logit of the hazard (log odds). From these parameters the odds of event occurrence can be obtained by exponentiation of the parameters. The conditional probabilities or hazards of event occurrence can be obtained from the parameter estimates using:

$$h(t_l) = \frac{1}{1 + \exp[-(\alpha_l)]} \ . \tag{13}$$

The estimates of the parameters of discrete event time model and the corresponding hazards for the baseline or reference group (when covariate values YC, gender and education equal to 0) under each of the four models fitted for the adolescents' relationship data are shown in Table 3. An increase in hazard shows higher risk of event occurrence. A closer look at hazards (based on estimates of parameters) in model 1 shows that this baseline model (a

model without a covariate) gives exactly the result on hazard probability of a life table analysis we have in Table 1. Thus, using the baseline model the hazard at each time period computed in life table can be obtained.

The baseline model estimates the overall population profile of the risk across time and indicate when events are more likely to occur. In order to know whether the hazard profile is different for adolescents with different values on the covariate youth centric, the dummy variable youth-centric (YC =1 if youth centric and 0 if not youth centric) is considered in addition to the time period variables as given in model 2. The parameter estimates are given under model 2 in Table 3. The estimates of the parameters $\alpha_{12}, \alpha_{13}, \cdots, \alpha_{23}, \alpha_{24}$ under model 2 represent the baseline log odds of hazard profile at each of the time periods for the non youth centric group of adolescents (a reference group with YC = 0). The parameter β_1 is a shift parameter that displaces the baseline log odds of hazard profile for the youth centric group (YC = 1). The estimated log odds for YC is 0.310 with corresponding odds ratio of 1.363 ($\exp(0.310)$). This indicates that youth centric adolescents are about 1.4 times more likely than non youth centric adolescents to experience the event 'sleeping with someone for the first time' (if at risk).

Model 3 contains the covariates gender (G) and education (E) in addition to the time period dummy variables. The baseline (or reference group) contains subjects with value 0 for gender and education, i.e., low educated female adolescents. The estimates of the parameters α's for the reference group is shown as log odds under model 3 in Table 3. The corresponding hazard (probability) of event occurrence at each time for the reference group are also shown in Table 3 under the column model 3. The parameter β_2 in model 3 is a shift parameter that displaces the baseline log odds of hazard profile for male adolescents (gender = 1) keeping the values of education constant. The estimated log odds for gender in model 3 is -0.542 with

the corresponding odds ratio of 0.581 ($\exp(-0.542)$), controlling for education. Similarly, the parameter β_3 in model 3 is a shift parameter that displaces the baseline log odds of hazard profile for highly educated adolescents (education = 1) keeping gender constant. The estimated log odds for education is -0.112 with the corresponding odds ratio of 0.894 ($\exp(-0.112)$), controlling for gender.

Model 4 includes the covariates YC, G and E in addition to the time period dummy variables. The baseline (or reference group) consists subjects with value 0 for YC, G and E, i.e., non youth-centric low educated female adolescent. The estimates of the parameters α's for the reference group is shown as log odds under model 4 in Table 3. The corresponding hazard (probability) of event occurrence at each time for the reference group are shown in the last columns of Table 3. The parameter β_1 in model 4 is a shift parameter that displaces the baseline log odds of hazard profile for the youth centric group (YC = 1) keeping gender and education constant. The estimated log odds for YC is 0.483 with the corresponding odds ratio of 1.620 ($\exp(0.483)$), controlling gender and education. Similarly, the parameter β_2 in model 4 is a shift parameter that displaces the baseline log odds of hazard profile for male adolescents (gender = 1) keeping the values of youth-centric and education constant. The estimated log odds for gender is -0.592 with the corresponding odds ratio of 0.553 ($\exp(-0.592)$), controlling YC and education. Thus, the odds of 'sleeping with someone for the first time' for boys is 0.553 times that for girls. The estimate of β_3 is interpreted in a similar fashion for education by keeping YC and gender constant.

The likelihood ratio test (e.g., Rao, 1973) is used to test the significance of the effects of the covariates in each of the models. The -2 Log Likelihood statistic (-2LL), which is often displayed in outputs from commonly used statistical software (e.g., SPSS) for logistic

regression routine, is the deviance statistic for the discrete event history models. A deviance statistic is always greater than 0 and the smaller the deviance the better the fit of the model to the observed data. The last row of Table 3 shows the deviance statistics for the four models we fit for the adolescents' relationship data.

<<<<< Insert Table 3 about here >>>>>>>>

A likelihood ratio test compares deviance statistics for the comparison of two nested models. Two models are called nested when both contain the same parameters and one of the models has at least one additional parameter. That is, one is obtained by having a constraint on the additional parameters in the other model. In the current example, model 1 can be obtained from all other models by imposing constraints that the parameters corresponding to the covariates youth-centric, gender and education are set to zero. Thus, model 1 is nested within all other models. Model 4 contains all parameters involved in other models. Thus, model 4 is the biggest model with all parameters of interest. For the likelihood ratio test, the difference in deviance statistics for two nested models will asymptotically have a chi-square distribution on k degrees of freedom (df), where k is the number of additional parameters in a bigger model in terms of parameters. When the difference in deviance statistic is larger than the critical value for chi-square distribution we reject the null hypothesis that the model with less number of parameters (reduced model) and conclude that some of the additional covariates in the model with more number of parameters (full model) have significant effect on the log odds of event occurrence. When the difference in the deviance statistic is small, we fail to reject the reduced parsimonious model and we conclude that the reduced model is as good as the full model. In the current example, the difference between deviance statistics for model 1 and model 2 is 2.046 (=645.979-643.933) which is smaller than the critical value of chi-square distribution with df =1 at 5% level of significance ($\chi^2_{1(0.05)} = 3.84$). Thus, model 2 is

not statistically better than model 1 and the covariate youth-centric alone has no effect on the log odds of event occurrence. However; it turns out that the covariate youth-centric indeed has an effect after the variables gender and education are controlled as discussed below. The difference in deviance statistics between model 1 and model 3 is 22.680 (=645.979 – 623.299). Because there are 2 more parameters in model 3 compared to model 1 (β's for gender and education), the *df* for the chi-square is now 2. Since the difference in deviance, 22.680 is greater than the chi-square value ($\chi^2_{2(0.05)} = 5.99$), we reject the reduced model (model 1) and conclude that model 3 gives a better fit of the data in such a way that at least one of the covariates involved in the model (gender and education) have significant effect on the log odds of event occurrence. Having model 3 which contains gender and education, we may be interested to know the effect of youth-centrism given that gender and education are controlled. We can compare models 3 and 4 for that purpose. Note that comparison between models 1 and 2 gives the effect of youth-centrism without controlling the variables gender and education. However, the comparison between models 3 and 4 helps to test whether youth-centrism has an effect on hazard probabilities after controlling the covariates gender and education.

The difference in deviance statistics between model 3 and model 4 is 3.943 (=623.299 - 619.356). Since there is only one more parameter in model 4 compared to model 3 (β_1 for youth-centric), the *df* for the chi-square is 1. Because the difference in deviance, 3.943, is slightly greater than the chi-square value ($\chi^2_{1(0.05)} = 3.84$) we conclude that YC has significant effect after controlling the covariates gender and education.

6.4. Polynomial specification of time period

The dummy variables for time period that included in the discrete event time model helps to maintain the shape of the baseline logit hazard function. The use of $T_1, T_2 \cdots, T_L$ as a representation of the *L* discrete time points in the model puts no specification on the shape of the hazard functions and further makes the interpretation of the parameters in the discrete event history model easier. Each of the coefficients of the dummy variables for time periods, α_l, is interpreted as the population value of logit hazard in time period *l* for the baseline group, for $l = 1, 2, \ldots, L$. The use of dummy variables representation for time periods in the model is encouraged as it does not put any constraint on the shape of the baseline model and facilitates interpretation of the coefficients. However, when there are many discrete time periods, *L* is large; the model needs the inclusion of many dummy variables representation for the time periods. This leads the model to be over parameterized and lack of parsimony (Efron, 1988; Fahrmeir and Wagenpfeil, 1996; Singer & Willett, 2003 (PP. 408)). Thus, using an alternative approach for the representation of time periods in the discrete event time model is required when the there are many time points. The option of considering the time periods as if they are continuous covariate and a specification of polynomial model for the baseline logit hazard function gives a more parsimonious model provided that the fit of the model to the data is not compromised (Mantel & Hankey, 1978 and Singer & Willett, 2003). The variable period in person-period dataset whose values represent the time period that the record describes, as shown in Table 2, can easily be used as a continuous covariate in polynomial representation of time in the discrete event time models. The polynomial representation could be linear, quadratic, cubic or higher degree polynomials. The choice could also be a logarithmic transformation of time or any other kind of function of time depending on the theoretical or practical motivation for such functions. In situations where the polynomial model is not pre-specified, search for the appropriate polynomial model can begin from the most simple one to the more complex models guided by statistical test for model comparison.

As outlined by Singer & Willett (2003 pp. 410), a formal goodness of test should confirm that the selected polynomial fits the data as good as the model with dummy variable representation of time period ('general' model). That means, a likelihood ratio test should confirm that there is no statistically significance difference between a polynomial model representation of time period for baseline logit hazard and model 1 for example in adolescents' relationship data example. Table 4 displays the deviance statistics and the differences in deviance statistics for the likelihood ratio test to identify an appropriate polynomial representation of time periods for the adolescents' relationship data. The *df* for the test is the difference in the number of parameters in the models to be compared. The difference in deviance statistics between the linear and the general model, 58.988, is greater than the chi-square value ($\chi^2_{11(0.05)} = 19.68$). Thus, the fit of the linear model is not as good as the general model. The next candidate model is the quadratic model. The difference in deviance statistics between quadratic and general models, 17.312, is less than the chi-square value ($\chi^2_{10(0.05)} = 18.31$). Thus, the fit of the quadratic model is as good as the general model. We found the same result for cubic model. However, a comparison between the quadratic and cubic models shows that the cubic model is not significantly better than the quadratic model (the difference in deviances, 2.581, is less than chi-square value ($\chi^2_{1(0.05)} = 3.84$)). Thus, quadratic polynomial representation of the time periods is parsimonious in its number of parameters while it fits the data as good as the general model for the discrete event time models for the adolescents' relationship data.

The parameters in the polynomial models are estimated using the same procedure as earlier using maximum likelihood method. The logistic regression model routines in the commonly used software can be used with a little modification of the dataset. First, a new variable need to be formed from the variable period within the person-period dataset. Then,

depending on the degree of the polynomial that we want to fit a series of additional variables should be recomputed from the new variable to represent each polynomial term in the model. For example, in the current data example, a new variable $(period - 12)$ is first computed from the variable period and then for the representation of quadratic term $(period - 12)^2$ is recomputed. In general, the new variable can be obtained by subtracting a constant *c* from period $(period - c)$ and then other series of variables can be obtained by taking the power of this variable to represent each polynomial term in the model. The choice of the constant *c* is arbitrary, however; *c* should be within the range of the observed event time periods. The constant *c* is in fact the time period at which the estimated logit hazard (probability) of event occurrence for the baseline group is the estimate of the intercept in the polynomial model. For example, the estimate of the parameter a_0 for the current example is the sample estimate of the logit hazard of event occurrence for the baseline group at time period (age) 12.

<<<<<< Insert Table 4 about here >>>>>>>>>>>

In general, for the linear model the intercept parameter a_0 represents the value of the logit hazard when the covariate is zero. The covariate in the current linear models is $(period - c)$. Thus, a_0 represents the value of the logit hazard when period equal to *c*. The slope parameter, b_1 is unaffected by the subtraction of constant *c* from the time period and it represents the increase (or decrease) in logit hazard per unit increase in time. For the quadratic model, the intercept still represents the value of the logit hazard in time period *c*. The slope parameter too still measures the increase (or decrease) in logit hazard per unit increase in time but now only at one particular instant, time period *c*. The curvature parameter, b_2 , in quadratic model specifies whether the logit hazard function is convex with a trough (U shape) or concave, with a peak ($\cap$ shape). If b_2 is positive, the hazard function is convex and if it is negative,

the shape is concave. The time period at which the hazard function reaches its peak or trough is given by $[c - \frac{1}{2}(b/b_2)]$.

For example, the estimates of the parameters $a_0, b_1, \text{and } b_2$ for the quadratic model of the adolescents' relationship data are -5.551, 1.191 and -0.075, respectively, i.e., the quadratic model for the baseline group is $\text{logit}(h_0(t_l) = -5.551 + 1.191(period - 12) - 0.075(period - 12)^2$. This implies that the estimate of the logit hazard at age 12 for the baseline is -5.551 (compare with the result in Table 3) and the instantaneous rate of change in logit hazard at age 12 is 1.191. Since the estimate for b_2 , -0.075, is negative the hazard function is concave reaching its peak at time period (age) $[12 - \frac{1}{2}(1.191/(-0.075))] = 19.94$. The peak of the hazard function is after age 19 and close to age 20 implying that the risk of the event, sleeping with someone for the first time, is highest at age 20.

6.5. Time-varying covariates

In previous sections, we considered covariates that have constant values with time. However, the values of some covariates for each person may change over time in practice. With a little modification, it is possible to relate the occurrence of the event of interest to covariates that change their values with time using the discrete event history model. One of the advantages of using person-period dataset is that it naturally allows a time– varying covariate simply to take on its appropriate value for each person in each record or period. In adolescents' relationship example, we focus in this chapter on the event of sleeping with someone for the first time and discussed the effect of time constant covariates youth-centric, gender and education. As mentioned in section 2, the survey had also collected the time at which the adolescents experienced the events 'going out', 'having a steady friend', and 'being very

much in love' for the first time. It may be hypothesized that the occurrence of these events could have effect on the timing of our event of interest, sleeping with someone for the first time. For simplicity and ease of presentation we consider only the effect of going out on the timing of sleeping with someone for the first time. The value on covariate going out (OUT) for each person at period *l* will be 0 if the person did not go out for the first time until period *l*. The value changes to 1 when a person goes out for the first time at period *l*. Thus, the covariate OUT is time-varying binary covariate in this example as its value changes with time. More technically, the covariate is defined as

$$\text{OUT}_l = \begin{cases} 0, & \text{if a person did not go out at time } l \\ 1, & \text{if a person gone out at time } l \text{ or before} \end{cases}$$

The data values can easily be appended in the person-period dataset for example next to the last column of Table 2.

Considering the quadratic model specification for the baseline group, the model with both time constant and time-varying covariate is given by:

$$\text{logit}(h_X(t_l)) = a_0 + b_1(period_l - 12) + b_2(period_l - 12)^2 + \beta_1 \text{YC} + \beta_2 \text{G} + \beta_3 \text{E} + \beta_4 \text{OUT}_l \quad (14)$$

Note that the time constant covariates YC, G, and E do not have subscript *l* while the time-varying covariate OUT has a subscript *l* to indicate that the data values for the variable OUT can be different values at different time periods for the same person. The parameter β_4 represents the difference in risk of the event sleeping with someone for the first time among adolescents who recently or previously experienced going out and those who have still not experienced going out controlling for other covariates. Because the covariate OUT is time-

varying, its effect does not contrast static group but adolescents who differ by unit value on the covariate OUT at each point in time, i.e., individuals can switch group membership and the adolescents who constitute the comparison group differ in each time period even if we are comparing two groups those who have experienced going out and who have not. Thus, the interpretation of the time-varying covariate's effect must be attached to each point in time period. In contrast, for the time constant covariates, we need not to attach a time point in the interpretation of the covariate's effect as the group members to be compared and data values of the covariate at each time period are constant. As the last model in equation (14) assumes time invariant effects of both the time constant and time-varying covariates, i.e., the effects on logit hazard in each time period is constant. Although the values of the time-varying covariate and the members of the groups to be compared may vary over time period, the difference between the logit hazard functions for the two groups to be compared in this example is constant and identical in every time period.

By fitting model in equation (14) for the adolescents' relationship data, we got:

$$\begin{aligned}\text{logit}(h_X(t_l)) = &-6.544 + 0.734(period_l - 12) - 0.043(period_l - 12)^2 + 0.508\text{YC} \\ &- 0.549\text{G} + 0.054\text{E} + 2.629\text{OUT}_l\end{aligned}$$

Comparison of the model that exclude the time-varying covariate OUT and this last equation using deviance statistic shows that the time-varying covariate is statistically significant controlling for the effect of the other covariates. The estimates of the parameters for the time constant covariates are interpreted in a similar way as we did in section 6.3. For example, the estimate of the parameter β_2, -0.549, for gender could be interpreted as odds ratio by taking the exponent of the estimate, i.e., controlling for the effect of other covariates in the model, the odds of sleeping with someone for the first time for boys is 0.578 (exp(-0.549)) times that of female adolescents. In another words, the odds of sleeping with someone for the first time

are 1.73 (=1/0.578) times for girls. In similar way, by taking the exponent of the estimate of the parameter β_4, (exp (2.629) = 13.866), at every age from 12 to 24 years the odds of sleeping with someone for the first time are about 14 times higher for adolescents who experienced the event of going out earlier and subsequent times compared to those who remain without the experience of going out controlling for the effect of the other covariates in the model. Note that the risk of sleeping with someone increases only in those time periods concurrent with or subsequent to, the event of going out. Before the event going out occurs, those adolescents who are later at greater risk of sleeping with someone are not different from other adolescents who stay without the event going out.

6.6. Proportionality assumption of the discrete event time models

The models we have considered so far assume that the covariates have an identical effect in every time period under the study which is known as the proportionality assumption. The assumption is crucial for the estimation procedure of most parametric hazard models for continuous time data. However, in the discrete time event history models presented above, apart from simplification of the models, there is no such requirement in the estimation procedure. In some practical situations this assumption is restrictive and can be relaxed by including an interaction term between the covariates and the time period. An inclusion of an interaction term between a covariate and time period allows the effects of the covariate to depend on time instead of being constant at all time periods. The interaction term with the covariate of interest can be made using the dummy variable representation of time period or the alternative polynomial representation. For the current example, the interaction terms are constructed by multiplying each of the covariate YC, G, E and OUT by the variable $(period - 12)$, linear term of the polynomial representation. The non-proportional discrete

event time model is estimated in each case by including the interaction term in model (14). One can test whether an effect of a covariate depends on time by comparing the model with proportionality assumption and the model that includes an interaction term between the covariate and time period. As explained earlier, comparison of the deviance statistics helps to make the comparison of the models. Since none of the differences in deviance statistics between model (14) and the models that include the interaction terms showed a statistical significance for the current example, the detailed results are not shown here. Thus, the data from adolescents' relationship example offer little evidence that the effects of the covariates change over time.

6.7. Competing-risk models

In the models we have considered so far, there is a single destination state from the origin state. In some applications there may be more than one way of (or reason for) exiting an origin state. Such reasons or destination states are referred as competing risks (Chiang, 1991; David & Moeschberger, 1978). For example, in the analysis of mortality or death rates, one may want to distinguish different causes of death; in the analysis of partnership formation, one may transit from single state to either marriage or cohabitation (without formal marriage). The hazard in such cases is defined as for single types of event, but now we have one for each competing risk. Suppose there are D mutually exclusive destination states, then the hazard of event type d at time t_l is:

$$h^{(d)}(t_l) = \Pr(\text{event of type } d \text{ at time } t_l \mid T \geq t_l) \ .$$

The hazard that no event of any type occurs at t_l given survival to time period $t_{l\text{-}1}$ is

$$h^{(0)}(t_l) = 1 - \sum_{d=1}^{D} h^{(d)}(t_l)$$

The survival function that the events occur after time t_l is the same as the probability that no event of any type occurs until and including time t_l is

$$S(t_l) = h^{(0)}(t_1) \times h^{(0)}(t_2) \times \cdots \times h^{(0)}(t_{l-1}) .$$

The model that relates the hazards to the covariates when individuals may leave the origin state to different destination states is the competing-risk model. There are two approaches to model the hazards. One approach is to model the hazards of each competing risk separately using the discrete event time model discussed so far, treating all other events as censored. This approach models the underlying risk of a particular event in the absence of all other risks. The other approach is modelling the hazards of the competing risks simultaneously using a multinomial logistic model.

For the multinomial logistic model, the person-period dataset discussed earlier needs a minor change. A multinomial event indicating categorical variable (response variable) E_{ild} needs to be defined indicating occurrence and type of event d at time period t_l for ith person. The response categories of E_{il} are 0 (no event), 1, 2, . . ., D. The multiple records in the person-period dataset for each person should be defined until one of the events or censoring occurs. The multinomial logistic model that contrasts event type d with no event for a person with covariates $X_1, X_2, \cdots, X_p$ is given by:

$$Log_e\left(\frac{h^d_{X_p}(t_l)}{h^0_{X_p}(t_l)}\right) = \alpha^d_l + \sum_{k=1}^{p} \beta^d_k X_k \qquad (15)$$

Comparison of this last model with the model in equation (8) shows that a separate set of time and covariate effects (α_l^d and β_k^d) are included for each type of event via the index *d*. Note that some of the covariates can be time-varying and may need subscript *l* for such covariates in equation (15). For the multinomial logistic model in equation (15) we estimate *D* equations contrasting each of the competing risks with no event. Further contrasts to compare the competing risks among each other can then be obtained from those *D* equations. For example, for partnership formation, two contrasts (marriage with single ("no event"), cohabitation with single) can be obtained using model in equation (15). The remaining contrast, marriage with cohabitation, may be estimated from the other two contrasts. Using the modified person-period dataset, a multinomial logistic model for discrete event time data in equation (15) can be estimated using routines developed for standard multinomial logistic model in the commonly available software (for example, SPSS, SAS, Stata).

6.8. Unobserved heterogeneity

In the models discussed so far, variability in the hazard of event occurrence is explained using observed covariates and risk factors. However, even after controlling for these observed characteristics, some subjects will be more likely to experience the event than others as a result of unobserved subject-specific risk factors. This unobserved heterogeneity in the hazard is sometimes referred to as frailty (Hougaard,1984,1995). If there are subject-specific unobserved factors that affect the hazard, the estimated form of the hazard function at the population or group level will tend to be different from those at the subject level. For example, if the hazards of all subjects in a population are constant over time, the aggregate population hazard will be decreasing. This can be explained by what is called a selection effect; that is, high risk subjects will tend to have the event first, leaving lower risk subjects

in the population. Therefore, as time goes the risk population is increasingly depleted of those subjects most likely to experience the event, leading to a decrease in the population hazard. Because of this selection, we may see a decrease in the population hazard even if individual hazards are constant (or even increasing). This selection effect not only affects the time dependence, but may, for example, also yield spurious time-covariate effects (Vermunt, 2002, 2009).

The common way to deal with unobserved heterogeneity is to include random effects (or subject-specific effects) in the models discussed so far. This involves the inclusion of a time constant latent covariate in the model and it requires an assumption about the distributional form of the latent variable. Mare (1994) and Vermunt (1997, 2002) presented discrete-time variants of such models. The amount of unobserved heterogeneity is determined by the variance of the latent variable, where the larger the variance the more unobserved heterogeneity. The interpretation of the regression parameters β will also change when random effects are included. In the models discussed so far without random effects, $\exp(\beta)$ is an odds ratio and it compares the odds of an event for two randomly selected individuals with values 1 unit apart on covariate X keeping the same values for other covariates in the model. In a model with random effects, $\exp(\beta)$ is an odds ratio only when the random effects are held constant, i.e. if we are comparing two hypothetical individuals with the same random effect values. Using models with random effects makes sense when it can be expected that important time-constant risk factors are not included in the model. Failure to control for such unobserved factors may bias the estimates of the factors included in the model. Discrete-time models with random effects can be defined using software for multilevel logistic regression analysis. Routines for continuous-time modeling sometimes often contain provisions for specifying models with unobserved heterogeneity (e.g the Stata routines stcox and streg).

7. Final Remarks

This chapter gave a gentle introduction to event history analysis for discrete event times. These methods were introduced to social and behavioral scientists by Allison (1984), Vermunt (1997, 2009), Willett and Singer (1993, 2003), and Yamaguchi (1991) among others. These methods are still relatively unknown and not widely used in psychology, despite their appropriateness for many research questions. We have shown that with an appropriate restructuring of the dataset, the software routines that are familiar to the applied researchers can be used for discrete event history analysis. That is, no specialized software is needed to perform an discrete event time analysis. The methods presented here are technically manageable and could also be used as an introduction to understanding of more advanced methods in continuous event time analysis, as, for example, described by Vermunt (1997, 2009) and Willett and Singer (2003).

Logistic regression model is adopted to relate the hazard of event occurrence to covariates. With an appropriate data restructuring, both the censoring problem and the inclusion of time-varying covariates are managed. The discussion was confined to only right censoring. Left censoring is less common in practice and in general difficult to deal with than right censoring. The method was extended for competing risks using multinomial logistic model. More advanced technique to account for unobserved heterogeneity was briefly discussed. Other more advanced topics that were not discussed in this chapter are models for multivariate events, covariates containing measurement error, , missing data on covariates, and recurrent events.

Models for multivariate events consider distributions of two or more distinct event time variables and jointly model the time variables (Vermunt & Moors, 2005; Wei, et al.,

1989). The objective of simultaneous modeling is to take into account the fact that the occurrence of one life event might directly affect the hazard for another type of event. When the covariates or predictor variables are subject to measurement error the estimates of regression coefficients and their corresponding confidence intervals may be biased and corrections are needed (Rosner, et al., 1990; Nakamura, 1992, Vermunt, 1996). When a covariate is partially missing, excluding the subjects with partially missing covariate values from the analysis leads to biased parameter estimates unless the missing mechanism is missing completely at random (Little & Rubin, 1987). Recurrent events refers to the situation in which subjects may experience the event of interest more than once, for example, repeated divorce or marriage, asthma attacks, child birth, employment, injury, and prison. Different techniques are suggested in literature for the analysis of recurrent event data. Lim, et al. (2007) compared those methods using empirical data of Pediatric Firearm Victim's visit to emergency department Department/Trauma Center at Children's Hospital of Wisconsin and all other hospitals in the Milwaukee metropolitan area between 1990 and 1995.

References

Allison, P. (1982). Discrete-Time Methods for the Analysis of Event Histories. In S. Leinhardt (Ed.), *Sociological Methodology* (pp. 61-98). San Francisco: Jossey-Bass.

Blossfeld, h.-p., & Rohwer, G. (1995). *Techniques of Event History Modeling: New approaches to Causal Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chiang, C.L. (1991). Competing risks in mortality analysis. *Annual Review of Public Health*, 12, 281-307.

Collett, D. (2003). *Modelling Survival Data in Medical Studies, 2nd edition*, Boca Raton, FL: Chapman &Hall/CRC.

Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society, Series B 34*,187-220.

Cutler, S. J. & Ederer, F. (1958). Maximum utilisation of the life table method in analysing survival. *Journal of Chronic Diseases*, 8, 699-712.

David, H. A . & Moeschberger, M. L. (1978). *The Theory of Competing Risks*. London: Griffin.

D'Agostino, R.B., Lee, M.L., Belanger, A.J., Cupples, L.A., Anderson, K., & Kannel, W.B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine, 9*, 1501–1515.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American statistical Association, 72*, 557-565.

Fahrmeir, L., & Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association, 91*, 1584-1594.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71, 75-84.

Hougaard, P. (1995). Frailty Models for Survival Data. *Lifetime Data Analysis*, 1, 255-273.

Kalbfleisch, J. and Prentice, R. (2002). The *Statistical Analysis of Failure Time Data. 2nd edition*. Hoboken, NJ: Wiley.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481.

Lim, H.J., Liu, J., and Meltzer-Lange, M. (2007). Comparison of Methods for analyzing recurrent event data: application to Emergency department visits of pediatric firearm victims. *Accident Analysis & Prevention*, 39, 290–299.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York, Wiley.

Mantel, M. H., & Hankey, B. F. (1978). A logistic regression analysis of response time data where the hazard function is time dependent. *Communications in Statistics: Theory and Methods A7*, 333-347.

Mare, R.D. (1994). Discrete-time bivariate hazards with unobserved heterogeneity: a partially observed contingency table approach. In Marsden, P.V. (Ed.) *Sociological Methodology* (pp. 341-383), Oxford: Basil Blackwell.

Merrell, M. (1947) Time-specific life tables contrasted with observed survivorship, *Biometrics*, 3 129-3.

Mould, R.F. (1976). Calculation of survival rates by the life table and other methods. *Clinical Radiology*, 27, 33-38.

Myung, I.J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47, 90-100.

Nakamura, T. (1992). Proportional Hazards Model with Covariates Subject to Measurement Error. *Biometrics*, 48, 829-838.

PASW statistics (2009). Rel. 17.0.2 , Chicago: SPSS inc.

Petersen, T. (1991). The Statistical Analysis of Event Histories. *Sociological Methods and Research, 19*, 270-323.

Rao, C.R. (1973). Linear Statistical Inference and Its Application. 2nd ed. New York: Wiley.

Rosner, B., Spiegelman, D. & Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132, 734-745.

SAS Institute (2009). *Base SAS® 9.2 Procedures Guide*. Cary, NC: SAS Institute Inc.

Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Methods for Studying Change and Event Occurrence*. New York: Oxford University Press.

Singer, J. D., & Willett, J. B. (1993). It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events, *Journal of Educational Statistics, 18*, 155-195.

Thompson, W.A. (1977). On the treatment of grouped observations in life studies. *Biometrics, 33*, 463-470.

Tuma, N. B., & Hannan, M. T. (1979). Approaches to the censoring problem in analysis of event histories. In K. F. Schuessler (Ed.), *sociological methodology* (pp.209-240). San Francisco: Jossey-Bass.

Vermunt, J.K. (1996). *Log-linear event history analysis: a general approach with missing data, unobserved heterogeneity, and latent variables*, Tilburg, NL: Tilburg University Press.

Vermunt, J.K. (1997). *Log-linear models for event histories*. Advanced Quantitative Techniques in the Social Sciences Series, vol 8. London: Sage Publications.

Vermunt, J.K. (2002). A general latent class approach to unobserved heterogeneity in the analysis of event history data. In J. Hagenaars and A. McCutcheon (Eds.), *Applied latent class analysis* (pp.383-407), place: Cambridge University Press.

Vermunt, J.K. (2009). Event history analysis. In: R. Millsap and A. Maydeu-Olivares (eds.), *Handbook of Quantitative Methods in Psychology*, 658-674. London: Sage.

Vermunt, J.K., & Moors, G.B.D. (2005). Event history analysis. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral science* (pp. 568-575). Chichester, UK: Wiley.

Vinken, H. (1998). *Political values and youth centrism. Theoretical and empirical perspectives on the political value distinctiveness of Dutch youth centrists*. Tilburg: Tilburg University Press.

Wei, L.J., Lin, D. Y. & Weissfeld, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, 84, 1065-1073.

Willett, J. B., & Singer, J. D. (1993). Investigating Onset, Cessation, Relapse And Recovery: Why you should, And How You Can, Use Discrete-Time Survival Analysis. *Journal of Consulting and clinical Psychology, 61*, 952-965.

Yamaguchi, K. (1991). *Event History Analysis*. Newbury Park, CA: Sage Publications.

Appendix: Syntaxes for the conversion of person-oriented dataset to person-period dataset, adolescents' relationships data.

SPSS syntax

```
*first open the data file 'relationships.sav'  which can be obtained from one of the authors.

do repeat D=D12 to D24 /ptime=12,13,14,15,16,17,18,19,20,21,22,23,24.
   if (time_sleeping > ptime) D=0.
   if (time_sleeping=ptime) D=1-censind1.
end repeat.
execute.

VARSTOCASES
  /ID=id
  /MAKE event FROM D12 D13 D14 D15 D16 D17 D18 D19 D20 D21 D22 D23 D24
  /INDEX=period(13)
  /KEEP=boy loweduc youthcen
  /NULL=DROP.
COMPUTE period=period+11.
EXECUTE.

* Making dummy variables for modeling
do repeat D=D12 to D24 /ptime=12,13,14,15,16,17,18,19,20,21,22,23,24.
   if (period >ptime) D=0.
   if (period=ptime) D=1.
   if (period<ptime) D=0.
end repeat.
execute.
```

SAS syntax

```
* Creating a person-period dataset from a person-level dataset ;
* Assuming the person-level dataset exists in drive C;

data relationships_pp1;
 set 'c:\relationships';
  do period= 12 to 24 ;
   if (time_sleeping > period) then event= 0;
   else if (time_sleeping = period) then event=1-censind1;
   else if (time_sleeping< period) then delete;
    output;
          end;
 keep id boy loweduc youthcen period event;
run;

proc print
 data=relationships_pp1;
 run;
```

```
data relationships_pp;
 set relationships_pp1;
 array AD[12:24] D12-D24;
 do dummy =12 to 24;
  if (period eq dummy) then AD[dummy]=1;
  else
   AD[dummy]=0;
  end;
   drop dummy;
  run;

proc print
 data=relationships_pp;
 run;
```

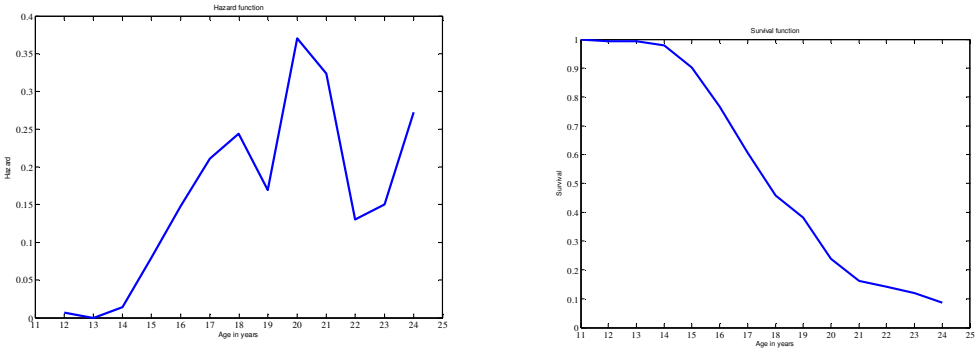


Figure 1. Estimated hazard and survival functions for 142 adolescents experience with sleeping with someone for the first time.

Table 1: Life table describing the ages at ‘sleeping with someone for the first time’ for a sample of 142 adolescents.

	Number			Proportion	
Age interval	Entering Interval	Withdrawing during Interval (censored)	‘slept with someone for the first timer’ during interval	‘slept with someone for the first time’ during interval (Hazard)	who has not slept with someone at the end of the interval (survival function)
[11, 12)	142	0			1.0000
[12, 13)	142	0	1	0.0070	0.9930
[13, 14)	141	0	0	0.0000	0.9930
[14, 15)	141	0	2	0.0142	0.9789
[15, 16)	139	0	11	0.0791	0.9014
[16, 17)	128	0	19	0.1484	0.7676
[17, 18)	109	0	23	0.2110	0.6056
[18, 19)	86	0	21	0.2442	0.4577
[19, 20)	65	0	11	0.1692	0.3803
[20, 21)	54	0	20	0.3704	0.2394
[21, 22)	34	0	11	0.3235	0.1620
[22, 23)	23	0	3	0.1304	0.1408
[23, 24)	20	6	3	0.1500	0.1190
[24, 25)	11	8	3	0.2727	0.0865

Table 2: Conversion of a person-oriented dataset into a person-period dataset for adolescents' relationships data example

Original dataset ('Person-oriented' dataset)

ID	Time	Censor	Gender
2	15	No	Female
9	24	Yes	Male
12	15	No	Male

Converted 'person-period' dataset

ID	Period	D_{12}	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}	D_{18}	D_{19}	D_{20}	D_{21}	D_{22}	D_{23}	D_{24}	Sleeping with someone (event)	Gender
2	12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Female
2	13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Female
2	14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Female
2	15	0	0	0	1	0	0	0	0	0	0	0	0	0	1	Female
9	12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Male
9	13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Male
9	14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Male
9	15	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Male
9	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Male
9	17	0	0	0	0	0	1	0	0	0	0	0	0	0	0	Male
9	18	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Male
9	19	0	0	0	0	0	0	0	1	0	0	0	0	0	0	Male
9	20	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Male
9	21	0	0	0	0	0	0	0	0	0	1	0	0	0	0	Male
9	22	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Male
9	23	0	0	0	0	0	0	0	0	0	0	0	1	0	0	Male
9	24	0	0	0	0	0	0	0	0	0	0	0	0	1	0	Male
12	12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Male
12	13	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Male
12	14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Male
12	15	0	0	0	1	0	0	0	0	0	0	0	0	0	1	Male

Table 3: Parameter estimates and goodness of fit statistic for the discrete event time models fitted to adolescents' relationships data

		Model 1		Model 2		Model 3		Model 4	
Covariate	Parameter	Log odds (estimates)	baseline hazard	Log odds (estimates)	baseline hazard	Log odds (estimates)	baseline hazard	Log odds (estimates)	Baseline hazard
D_{12}	α_{12}	-4.949	0.0070	-5.062	0.0063	-4.780	0.0083	-4.872	0.0076
D_{13}	α_{13}	-21.203	0.0000	-21.312	0.0000	-21.053	0.0000	-21.135	0.0000
D_{14}	α_{14}	-4.241	0.0142	-4.353	0.0127	-4.071	0.0168	-4.158	0.0157
D_{15}	α_{15}	-2.454	0.0791	-2.566	0.0714	-2.274	0.0933	-2.360	0.0863
D_{16}	α_{16}	-1.747	0.1484	-1.851	0.1358	-1.552	0.1748	-1.625	0.1645
D_{17}	α_{17}	-1.319	0.2110	-1.421	0.1945	-1.165	0.2378	-1.236	0.2251
D_{18}	α_{18}	-1.130	0.2442	-1.232	0.2258	-0.978	0.2733	-1.043	0.2606
D_{19}	α_{19}	-1.591	0.1692	-1.683	0.1567	-1.363	0.2038	-1.418	0.1950
D_{20}	α_{20}	-0.531	0.3703	-0.625	0.3486	-0.318	0.4212	-0.364	0.4100
D_{21}	α_{21}	-0.738	0.3234	-0.823	0.3051	-0.417	0.3972	-0.447	0.3901
D_{22}	α_{22}	-1.897	0.1304	-1.956	0.1239	-1.564	0.1731	-1.543	0.1761
D_{23}	α_{23}	-1.735	0.1500	-1.785	0.1437	-1.389	0.1996	-1.360	0.2042
D_{24}	α_{24}	-0.981	0.2727	-1.011	0.2668	-0.670	0.3385	-0.656	0.3416
youth-centrism (YC, yes=1)	β_1			0.310				0.483	
Gender (G) (male=1)	β_2					-0.542		-0.592	
Education (E) (high=1)	β_3					-0.112		-0.059	
-2LL		645.979		643.933		623.299		619.356	

Table 4: Polynomial representations for time period in a baseline discrete event time model for adolescents' relationships data

Polynomial	model for the baseline logit hazard	Number of parameters	-2LL	Difference in -2LL in comparison to: Previous model	Difference in -2LL in comparison to: General model
Linear	$\text{logit}(h_0(t_l) = a_0 + b_1(period_l - 12)$	2	704.967		58.988
Quadratic	$\text{logit}(h_0(t_l) = a_0 + b_1(period_l - 12) + b_2(period_l - 12)^2$	3	663.291	41.676	17.312
cubic	$\text{logit}(h_0(t_l) = a_0 + b_1(period_l - 12) + b_2(period_l - 12)^2 + b_3(period_l - 12)^3$	4	660.710	2.581	14.731
General	$\text{logit}(h_0(t_l)) = \alpha_{12}T_{12} + \cdots + \alpha_{24}T_{24}$	13	645.979		