1

Introduction

1.1 What is survival analysis and how is it applied?

'What is survival analysis?' Before starting discussion on this topic, think about what 'survives.' In the cases considered here, we are talking about things that have a life span, those things that are 'born,' live, change status while they live, and then die. Therefore, 'survival' is the description of a life span or a living process before the occurrence of a status change or, using appropriate jargon, an *event*.

In terms of 'survival,' what we think of first are organisms like various animal species and other life forms. After birth, a living entity grows, goes through an aging process, and then decomposes gradually. All the while, they remain what they are – the same organisms. The gradual changes and developments over a life course reflect the survival process. For human beings in particular, we survive from death, disease, and functional disablement. While biology forms its primary basis, the significance of survival is largely social. At different life stages, we attend school, get married, develop a professional career, and retire when getting old. In the meantime, many of us experience family disruption, become involved in social activities, cultivate personal habits and hobbies, and make adjustments to our daily lives according to physical and mental conditions. These social facets are things that are not organisms but their life span is *like* that of a living being: things that *live*, things that have beginnings, transformations, and then *deaths*. In a larger context, survival can also include such events as an automobile breakdown, the collapse of a political system in a country, or the relocation of a working unit. In cases such as these and in others, existence dictates processes of survival and their status change, indicated by the occurrence of events.

The practice of survival analysis is the use of reason to describe, measure, and analyze features of events for making predictions about not only survival but also 'time-to-event processes' – the length of time until the change of status or the occurrence of an event – such as from living to dead, from single to married, or from healthy to sick. Because a life span, genetically, biologically, or mechanically, can be cut short by illness, violence, environment, or other factors, much research in survival analysis involves making comparisons among

Survival Analysis: Models and Applications, First Edition. Xian Liu.

groups or categories of a population, or examining the variables that influence its survival processes. As they have come to realize the importance of examining the inherent mechanisms, scientists have developed many methods and techniques seeking to capture underlying features of various survival processes. In the academic realm, survival analysis is now widely applied in a long list of applied sciences, owing considerably to the availability of longitudinal data that records histories of various survival processes and the occurrences of various events. At present, the concept of survival no longer simply refers to a biomedical or a demographic event; rather, it expands to indicate a much broader scope of phenomena characterized by time-to-event processes.

In medical research, clinical trials are regularly used to assess the effectiveness of new medicines or treatments of disease. In these settings, researchers apply survival analysis to compare the risk of death or recovery from disease between or among population groups receiving different medications or treatments. The results of such an analysis, in turn, can provide important information with policy implications.

Survival analysis is also applied in biological research. Mathematical biologists have long been interested in evolutionary perspectives of senescence for human populations and other species. By using survival analysis as the underlying means, they delineate the life history for a species' population and link its survival processes to a collection of physical attributes and behavioral characteristics for examining its responses to its environment.

Survival data are commonly collected and analyzed in social science, with topics ranging widely, from unemployment to drug use recidivism, marital disruption, occupational careers, and other social processes. In demography, in addition to the mortality analysis, researchers are concerned with such survival processes as the initiation of contraceptive use, internal and international migration, and the first live birth intervals.

In the field of public health, survival analysis can be applied to the analysis of health care utilization. Such examination is of special importance for both planners and academics because the health services system reflects the political and economic organization of a society and is concerned with fundamental philosophical issues involving life, death, and the quality of life.

Survival analysis has also seen wide applications in some other disciplines such as engineering, political science, business management, and economics. For example, in engineering, scientists apply survival analysis to perform life tests on the durability of mechanical or electric products. Specifically, they might track a sample of products over their life course for assessing characteristics and materials of the product's designed life and for predicting product reliability. Results of such studies can be used for the quality improvement of the products.

1.2 The history of survival analysis and its progress

Originally, survival analysis was used solely for investigations of mortality and morbidity on vital registration statistics. The earliest arithmetical analysis of human survival processes can be traced back to the 17th century, when the English statistician John Graunt published the first life table in 1662 (Graunt, 1939, original edition, 1662). For a long period of time, survival analysis was considered an analytic instrument, particularly in biomedical and demographical studies. At a later stage, it gradually expanded to the domain of engineering to describe/evaluate the course of industrial products. In the past forty years, the scope of

survival analysis has grown tremendously as a consequence of rapid developments in computer science, particularly the advancement of powerful statistical software packages. The convenience of using computer software for creating and utilizing complex statistical models has led scientists of many disciplines to begin using survival models.

As applications of survival analysis have grown rapidly, methodological innovation has accelerated at an unprecedented pace over the past several decades. The advent of the Cox model and the partial likelihood perspective in 1972 triggered the advancement of a large number of statistical methods and techniques characterized by regression modeling in the analysis of survival data. The major contribution of the Cox model, given its capability of generating simplified estimating procedures in analyzing survival data, is the provision of a flexible statistical approach to model the complicated survival processes as associated with measurable covariates. More recently, the emergence of the counting processes theory, a unique counting system for the description of survival dynamics, highlights the dawning of a new era in survival analysis due to its tremendous inferential power and high flexibility for modeling repeated events for the same observation and some other complicated survival processes. In particular, this modern perspective combines elements in the large sample theory, the martingale theory, and the stochastic integration theory, providing a new set of statistical procedures and rules in modeling survival data. To date, the counting process system and the martingale theory have been applied by statisticians to develop new theorems and more refined statistical models, thus bringing a new direction in survival analysis.

1.3 General features of survival data structure

In essence, a survival process describes a life span from a specified starting time to the occurrence of a particular event. Therefore, the primary feature of survival data is the description of a change in status as the underlying outcome measure. More formally, a status change is the occurrence of an *event* designating the end of a life span or the termination of a survival process. For instance, a status change occurs when a person dies, gets married, or when an automobile breaks down. This feature of a status 'jump' makes survival analysis somewhat similar to some more conventional statistical perspectives on qualitative outcome data, such as the logistic or the probit model. Broadly speaking, those traditional models can also be used to examine a status change or the occurrence of a particular event by comparing the status at the beginning and the status at the end of an observation interval. Those statistical approaches, however, ignore the timing of the occurrence of this lifetime event, and thereby do not possess the capability of describing a time-to-event process. A lack of this capability can be detrimental to the quality of analytic results, thereby generating misleading conclusions. The logistic regression, for example, can be applied to estimate the probability of experiencing a particular lifetime event within a limited time period; nevertheless, it does not consider the time when the event occurs and therefore disregards the length of the survival process. Suppose that two population groups have the same rate of experiencing a particular event by the end of an observation period but members in one group are expected to experience the event significantly later than do those in the other. The former population group has an advantaged survival pattern because its average life is extended. Obviously, the logistic regression ignores this timing factor, therefore not providing precise information.

Most survival models account for the timing factor on a status jump. Given this capacity, the second feature of survival data is the description of a time-to-event process. In

the literature of survival analysis, time at the occurrence of a particular event is regarded as a random variable, referred to as event time, failure time, or survival time. Compared to statistical techniques focused on structures, the vast majority of survival models are designed to describe a time course from the beginning of a specific time interval to the occurrence of a particular event. Given this feature, data used for survival analysis are also referred to as time-to-event data, which consist of information both about a discrete 'jump' in status as well as about the time passed until the occurrence of such a jump.

The third primary feature of survival data structure is censoring. Survival data are generally collected for a time interval in which the occurrences of a particular event are observed. As a result, researchers can only observe those events that occur within a surveillance window between two time limits. Consequently, complete survival times for many units under examination are not observed, with information loss taking place either before the onset or beyond the end of the study interval. Some units may be lost to observation in the middle of an investigation due to various reasons. In survival analysis, such missing status on event times is called *censoring*, which can be divided into a variety of types. For most censoring types, a section of survival times for censored observations are observable and can be utilized in calculating the risk of experiencing a particular event. In survival analysis, this portion of observed times is referred to as censored survival times. As censoring frequently occurs, the majority of survival analysis literally deals with incomplete survival data, and accordingly scientists have found ways to use such limited information for correctly analyzing the incomplete survival data based on some restrictive assumptions on the distribution of censored survival times. Given the importance of handling censoring in survival analysis, a variety of censoring types are delineated in Section 1.4.

As survival processes essentially vary massively based on basic characteristics of the observations and environmental conditions, a considerable body of survival analysis is conducted by means of censored data regression modeling involving one or more predictor variables. Given the addition of covariates, survival data structure can be viewed as consisting of information about three primary factors, otherwise referred to as a 'triple:' survival times, censoring status, and covariates. Given a random sample of n units, the data structure for survival analysis actually contains n such triples. Most survival models, as will be described extensively in later chapters, are built upon such a data structure.

Given different emphases on the variety of features, survival analysis is also known as duration analysis, time-to-event analysis, event histories analysis, or reliability data analysis. In this book, these concepts are used interchangeably.

1.4 Censoring

Methodologically, censoring is defined as the loss of observation on the lifetime variable of interest in the process of an investigation. In survival data, censoring frequently occurs for many reasons. In a clinical trial on the effectiveness of a new medical treatment for disease, for example, patients may be lost to follow-up due to migration or health problems. In a longitudinal observational survey, some baseline respondents may lose interest in participating in subsequent investigations because some of the questions in a previous questionnaire are considered too sensitive.

Censoring is generally divided into several specific types. If an individual has entered a study but is lost to follow-up, the actual event time is placed somewhere to the right of the

censored time along the time axis. This type of censoring is called *right censoring*. As right censoring occurs far more frequently than do other types and its information can be included in the estimation of a survival model, the focus of this section is on the description of right censoring. For analytic convenience, descriptions of right censoring are often based on the assumption that an individual's censored time is independent of the actual survival time, thereby making right censoring noninformative. While this assumption does not always hold, the issue of informative censoring and the related estimating approaches are described in Chapter 9. Other types of censoring, including *left censoring* and *interval censoring*, are also described in this section. Additionally, I briefly discuss the impact of left truncation on survival analysis, a type of missing data that is different from censoring.

1.4.1 Mechanisms of right censoring

Right censoring is divided into several categories: Type I censoring, random censoring, and Type II censoring. In *Type I censoring*, each observation has a fixed censoring time. Type I censoring is usually related to a predetermined observation period defined according to the research design. Generally, a specific length of time is designed with a starting calendar date and an ending date. In most cases, only a portion of observations would experience a particular event of interest during this specified study interval and some others would survive to the endpoint. For those who survive the entire observation period, the only information known to the researcher is that the actual survival time is located to the right of the endpoint of the study period along the time axis, mathematically denoted by $T > C$, where T is the event time and C is a fixed censored time. Therefore, lifetimes of those survivors are viewed as right censored, with the length of the censored time equaling the length of the observation period.

Right censoring also occurs randomly at any time during a study period, referred to as *random censoring*. This type of censoring differs essentially from Type I censoring because the censored time is not fixed, but, rather, behaves as a random variable. Some respondents may enter the study after a specified starting date and then are right censored at the end of the study interval. Such observations are also listed in the category of random censoring because their *delayed entry* is random. Statistically, time for random censoring can be described by a random variable C_i (the subscript i indicates variation in C among randomly censored observations), generally assumed to be independent of survival time T_i. Mathematically, for a sample of n observations, case i ($i = 1, 2, \ldots, n$) is considered randomly censored if $C_i < T_i$ and $C_i < C$, where C is the fixed Type I censored time. The censored survival time for random censoring is measured as the time distance from the time of entry into the study to the time when random censoring occurs.

Figure 1.1 graphically displays the occurrences of Type I and random censoring. In this figure, I present data for six individuals who participate in a study of mortality at older ages, noted by, respectively, persons 1, 2, 3, 4, 5, and 6. The study specifies an observation period from 'start of study' to 'end of study.' The sign '×' denotes the occurrence of a death, whereas the sign '+' represents right censoring.

In Figure 1.1, person 1 enters the study at the beginning of the study and dies within the interval. Therefore, this case is an event, with time-to-event T_1 counted as the time elapsed from the start of the study to the time of death. Person 2 also enters the study at the beginning of the study, but at the end of the study, this person is still alive. Therefore, person 2 is a typical case of Type I right censoring, with the censored survival time equaling the full

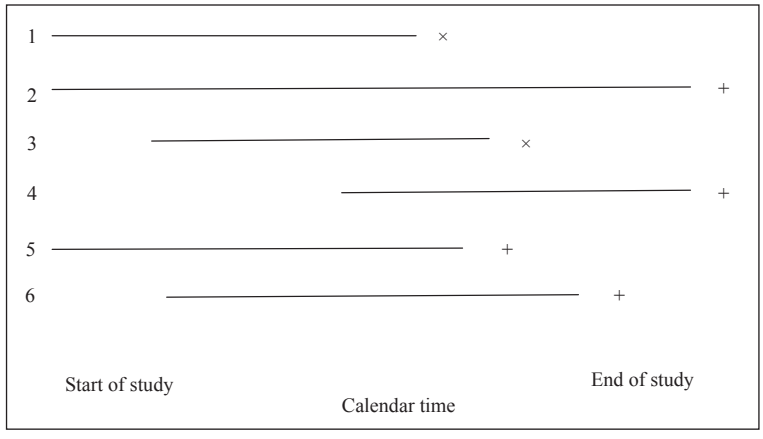

Figure 1.1 Illustration of Type I and random censoring.

length of the study interval. Persons 3 and 4 both enter the study after the start of the study, with person 3 deceased during the interval and person 4 alive throughout the rest of the interval. Consequently, person 3 has an event whose survival time is the distance from the time of the delayed entry to the time of death, whereas person 4 is a case of random censoring with the censored survival time measured as the length of time between the delayed entry and the end of the study. Entering the study later than expected, person 4 can also be considered a left truncated observation, which will be described in Subsection 1.4.2. Finally, persons 5 and 6 are lost to follow-ups before the termination of the study, with person 5 entering the investigation at the start and person 6 entering during the period of investigation. Both persons are randomly censored. Their censored times, denoted by C_5 and C_6, respectively, measured as the time elapsed between the starting date of the study and the censored time for person 5, or between the time of the delayed entry and the censored time for person 6. Unlike person 2, censored times for persons 4, 5, and 6 differ from each other and are smaller than C.

Type II right censoring refers to the situation in which a fixed number of events is targeted for a particular study. When the designed number of events is observed, a study would terminate automatically and all individuals whose survival times are beyond the time of termination are right censored. For those individuals, the censored survival time is measured as the distance from the start of observation to the time at which the study terminates. Type II right censoring is not related to a fixed ending time; rather, it is associated with a time determined by a date when a targeted number of events are observed. Given this restriction, surveys or clinical trials associated with Type II right censoring are much rarer than those with other types of right censoring.

1.4.2 Left censoring, interval censoring, and left truncation

Left censoring refers to a data point, known to be prior to a certain date but unknown about its exact location. This type of censoring frequently occurs in a study design involving two separate study stages. Individuals who enroll in the first selection process but are not eligible for the second process are viewed as left censored. For example, in a study of the initiation of first contraceptive use after marriage, if a couple marries but has already used

contraceptive means prior to marriage, this couple is left censored for further analysis. Another example is a study of first marijuana use among high school students. If a respondent has used marijuana before the study, but does not remember the exact timing of the first use, this observation is left censored. In clinical trials, researchers often specify a recruitment period and a study period. If a patient is recruited into the study but has experienced an event of interest before the study period starts, the case is left censored.

Another type of censoring is *interval censoring*. In some investigations, actual event times are unknown, and the data point is only known to be located between two known time points. Demographers often use aggregate mortality data for a specific calendar year for constructing a life table and, clearly, such mortality data are interval censored. Interval censoring also occurs frequently in clinical trials and large-scale longitudinal surveys in observational studies. For example, a clinical trial on the effectiveness of a new medicine on posttraumatic stress disorder (PTSD) recruits a sample of patients diagnosed with PTSD, proposing a series of periodic follow-up investigations to examine the rate of resolution of this psychiatric disorder. Some patients with PTSD at a starting time point are observed to have recovered at the next follow-up time point. Here, the exact timing of PTSD resolution is unknown and the only information known to the researcher is the time interval in which the event occurred. As a result, the PTSD time span for those patients who have recovered is interval censored. For analytic convenience, interval-censored survival times are often assumed to be located at a fixed time point, either in the middle of a specific interval (Siegel and Swanson, 2004) or immediately prior to an exact follow-up time (Lawless, 2003; Scharfstein, Rotnitzky, and Robins, 1999). In Chapters 4 and 5, this type of censoring is further discussed and illustrated.

1.4.2.1 Left truncation

Time-to-event data are also subject to *left truncation*, a unique type of missing data. A survey respondent who enters the observation process after a given starting date is referred to as a *staggered entry* or a *delayed entry*. Such observations are left truncated, with the truncated time measured as the time distance from the time of entry to the occurrence of an event or of right censoring. Compared to various types of censoring, left truncation is a phenomenon often associated with sample selection that leaves individuals out of observation for some time. In a study of marital disruption, for example, some individuals become married after the investigation starts, so their entry into the study is delayed and their survival times are left truncated at the time of marriage. Left truncation can potentially cause serious selection bias in survival analysis because it underestimates the risk of experiencing a particular event; however, there are standard statistical techniques to handle such bias. In Chapter 5, the impact of left truncation on survival analysis and how to use certain statistical methods for handling it is illustrated.

1.5 Time scale and the origin of time

Survival processes describe the length of time measured from some specified starting time point to the occurrence of a life event. According to this specification, the measurement of an event time should start from a well-defined origin of time and ends at the time when a particular event of interest occurs. Therefore, a metric unit at which time is measured must be specified first.

In proposing a study plan, the specification of the time scale must depend on the nature of the study and the targeted length of an observation period. In observational studies, the occurrence of a behavioral event is usually a gradual process. Examples of such gradual life events are recovery from disability among older persons, changes in marital status, and discontinuation of drinking alcohol among heavy drinkers. In following up those processes, the month may be an appropriate choice as the time scale. Clinical research, on the other hand, can be linked with lifetime events, both with rapid changes in status and also with a relatively slow pace. In cancer research, for example, the survival rate within a fixed time period varies significantly for different types of cancer. A study of surgical treatment on lung cancer may examine the improvement of the survival rate for six months. In such research, a day or week is the appropriate time scale. In studies of more gradual processes such as prostate cancer, the survival rate should be observed for a substantially longer period because these patients typically live much longer than those with lung cancer. Thus, a month is a better option for the second study. In health services research, survival data with different service types can be mixed with a variety of time scales. For example, a patient admitted to a short-stay hospital stays there for only a few days, whereas the average length of stay in a nursing home generally exceeds an entire year (Liu *et al*., 1997). Accordingly, the time scale needs to be specified based on the nature of a particular service type.

Once the metric unit is specified, the starting point (or the origin of time) of the event time must be accurately defined. Without a clear and unambiguous definition of the starting time, the event time can be severely misspecified, thereby resulting in erroneous analytic results. In different situations, the starting time can be defined in various ways. As time proceeds with ordinary calendar time, a standard scale needs to be chosen to align individuals at time 0. In general, the ideal scenario is to follow up lifetimes of one or more birth cohorts of individuals from their births to the date when the last survivor dies. This scenario, however, is utterly unrealistic because the researcher launching such a study would definitely pass away or retire long before the study is ended. In demographic and epidemiologic studies, age is often specified as the time scale, but the use of the period-specific data assumes a hypothetic birth cohort. Here, the true origin of time is actually the starting date of a specific calendar period, rather than birth.

1.5.1 Observational studies

In observational studies, survival data are usually collected from large-scale longitudinal surveys. In most cases, researchers would set a calendar date as the beginning time of the study and then draw a random sample of individuals according to a specific study plan. Those individuals' survival status would be followed up for a considerably long period of time (ten years, say). Here, the calendar date for the first interview is used as the origin of time, and all respondents should be aligned on this specific date, with the event time operationally defined as the distance between the date of the first interview and the date of an event. In practice, setting the starting calendar date as the origin of time has some advantages: it is convenient to align respondents for survival analysis and it is a straightforward method to calculate an individual's event time. This procedure, however, can encounter several selection problems. The survival process is incomplete for a targeted population and the observation is relevant only to a truncated chronological period, in which gradual processes of survival from a particular event cannot be entirely captured (Liu, 2000). Some of those limitations can be substantially mitigated by correctly specifying a causal framework. In a

typical longitudinal study, the date of birth itself can be regarded as an explanatory factor, so that the cohort effect on survival processes can be incorporated into a survival model (Clayton, 1978). As age progresses with time, the age at baseline can serve as an important control variable for selection bias from left truncation.

1.5.2 Biomedical studies

In biomedical studies, survival analysis is generally performed to examine the effectiveness of a new medicine or a new treatment on reducing the rate of mortality or of disease. Given this focus, the origin of time in biomedical research is often specified as the starting date of a new treatment/medication or of exposure to disease. Consider, for example, the study of survival from prostate cancer after the surgical treatment. As the event time is defined as the time elapsed from treatment to death, the origin of time in this context should be the date of surgery performed on the patient. As a result, all patients of this study can be aligned by the time origin regardless of when the surgery is performed. Similarly, in a study of asbestos exposure and lung cancer, the date of first exposure to asbestos on a regular basis is an appropriate choice as the origin of time and all study subjects, no matter when they enter the study, should be aligned by the date of first regular exposure to asbestos. Sometimes, clinical trials use the date of randomization as the origin of time, referred to as the study time. Given a period of recruitment, patients would enter the study on different calendar dates, so that such a calendar time, referred to as the patient time, can differ considerably from the study time (Collett, 2003).

1.5.3 Health care utilization

In studies of health care utilization, some mutually exclusive service types – such as nursing home, short-stay hospital, and long-term hospital – are regularly specified for analyzing transitions from one service type to another (Liang *et al*., 1993, 1996; Liu *et al*., 1997). Thus, the admission date should be used as the origin of time and the time elapsed from admission to discharge is the event time. With admission episodes used as the primary unit of analysis, repeated visits within a specific observation period are common and the size of censored cases is relatively small.

There are situations in which the true origin of time is difficult to define. Consider a study of liver cirrhosis and mortality by Liestol and Andersen (2002). As is typical in biomedical research, the origin of time in this study should be the date of diagnosis. This medical condition, however, develops gradually with symptoms vague in the early stage and varying significantly among individuals, thus making the time of diagnosis a questionable time origin. Some patients with liver cirrhosis might be diagnosed with the disease later than others and some are never even diagnosed until the time of death. Consequently, patients with liver cirrhosis cannot be aligned appropriately according to the natural progression of disease, implying the origin of time to be a latent random variable representing the degree of delayed entry. Liestol and Andersen (2002) suggest the use of age or calendar time as a surrogate time 0, because age and calendar time are both well defined and serve as strong determinants of disease severity. Given strong variations in physical characteristics, genetic predisposition, and health behaviors among individuals born in the same calendar year, the use of age or calendar time still makes time a strong random effect; accordingly, more complex procedures need to be designed to account for the impact of this latent factor.

1.6 Basic lifetime functions

Survival analysis begins with a set of propositions on various aspects of a lifetime event: basic concepts, mathematical functions, and specifications generally applied in survival analysis. The focus is placed upon three most basic functions – the survival function, the probability density function, and the hazard function.

1.6.1 Continuous lifetime functions

I start by describing time as a continuous process. Let $f(t)$ be the probability density function (p.d.f.) of event time T. Then, according to probability theory, the cumulative distribution function (c.d.f.) over the time interval $(0, t)$, denoted by $F(t)$, represents the probability the random variable T takes from time 0 to time t ($t = 0, 1, \ldots, \infty$), given by

$$F(t) = \Pr(T \le t) = \int_0^t f(u)\,\mathrm{d}u. \tag{1.1}$$

Defined as the probability that no event occurs from time 0 to time t, the survival function at time t, denoted by $S(t)$, is simply the complement of the c.d.f.:

$$S(t) = \Pr(T > t) = 1 - \Pr(T \le t) = 1 - F(t). \tag{1.2}$$

By definition, given $t \to \infty$, $S(0) = 1$ and $S(\infty) = 0$. For analytic convenience, statisticians and demographers sometimes arbitrarily define a finite ending time, denoted by ω, assuming that no one survives beyond this time point. In this specification, we have $S(0) = 1$ and $S(\omega) = 0$. Empirically, the value of ω can be determined by the maximum life span ever observed, or just by a given very old age beyond which only very few have ever been found to survive, so that the very small value in $S(\omega)$ can be ignored (Liu and Witten, 1995).

The p.d.f. of T can be expressed in terms of $S(t)$, given by

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}. \tag{1.3}$$

Equation (1.3) indicates that the slope (the derivative) of the survival function determines the p.d.f. of T. As $S(t)$ is a nonincreasing function, this slope must take the negative sign to derive the nonnegative p.d.f. Strictly speaking, the p.d.f. is not a probability, but a probability rate, which can take values greater than one.

The hazard function at time t is defined as the instantaneous rate of failure at time t, generally denoted by $h(t)$ and mathematically defined by

$$h(t) = -\frac{1}{S(t)}\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \frac{f(t)}{S(t)}, \tag{1.4}$$

or

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{T \in (t, t+\Delta t] | T \ge t\}}{\Delta t}. \tag{1.5}$$

Equation (1.4) demonstrates that the hazard rate is conceptually a standardized instantaneous rate of failure relative to the survival rate at time t. From another perspective, Equation (1.5) expresses the hazard rate as the ratio of the conditional probability at t (the probability of experiencing a particular event at time t given the condition $T \geq t$) over an infinitesimal time change. Because Δt tends to 0, the *hazard rate* can be literally understood as the conditional probability of failure with respect to the limit of a time interval. With this instantaneous property, the hazard rate is also referred to as the force of mortality, the intensity rate, or the instantaneous risk (Andersen *et al*., 1993; Kalbfleisch and Prentice, 2002; Liu, 2000; Liu and Witten, 1995). Given standardization and its unique sensitivity to the change in the survival function, the hazard function is considered a preferable indicator for displaying the relative risk of experiencing a particular event in survival analysis.

Given Equation (1.3), the hazard function at time t can also be written by

$$h(t) = \frac{-\mathrm{d}\log S(t)}{\mathrm{d}t}. \tag{1.6}$$

By Equation (1.6), the hazard rate is mathematically defined as the derivative of the log survival probability at time t multiplied by -1. As a survival function is monotonically decreasing, the hazard function is nonnegative but not necessarily smaller than or equal to one. Therefore, as the standardized p.d.f., the hazard rate is a conditional probability rate. It is essential for the reader to comprehend the concept and the underlying properties of the hazard function because most survival models described in later chapters are created on the hazard rate.

The above equations highlight the intimate relationships among $f(t)$, $S(t)$, and $h(t)$. Mathematically, they reflect different profiles of a single lifetime process, with each providing a unique aspect of survival data. Therefore, each of these basic functions can be readily expressed in terms of another. For example, the survival probability $S(t)$ can be expressed as the inverse function of Equation (1.6):

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t h(u)\mathrm{d}u\right) \\ &= \exp[-H(t)], \end{aligned} \tag{1.7}$$

where $H(t)$ is the integration of all hazard rates from time 0 to t, defined as the continuous cumulative hazard function at time t.

Similarly, from Equation (1.7), the cumulative hazard function $H(t)$ can be expressed in terms of $S(t)$, given by

$$H(t) = -\log S(t). \tag{1.8}$$

Furthermore, from Equations (1.4) and (1.7), the probability density function $f(t)$ can be written in terms of the hazard function:

$$f(t) = h(t)\exp\left(-\int_0^t h(u)\mathrm{d}u\right). \tag{1.9}$$

From the above basic functions, the expected life remaining at time t, also referred to as life expectancy at t, can be computed. As it represents the unit-based probability surviving

at time t, $S(t)$ can be considered the intensity of expected life at t. Let $\lim_{t\to\infty} tS(t) = 0$; the expected life remaining at time 0, denoted $\mathrm{E}(T_0)$, can be written by

$$\mathrm{E}(T_0) = \mathrm{E}(T \mid t = 0) = \int_0^\infty S(u)\,\mathrm{d}u. \tag{1.10}$$

Likewise, the expected life remaining at time t, $\mathrm{E}(T_t)$, is

$$\mathrm{E}(T_t) = \mathrm{E}(T \mid T \geq t) = \frac{\int_t^\infty S(u)\,\mathrm{d}u}{S(t)}, \tag{1.11}$$

where $S(t)$ represents exposure for the expected life remaining at time t.

The expected life between time t and time $t + \Delta t$, denoted by $\mathrm{E}({}_{\Delta t}T_t)$, is a component in $\mathrm{E}(T_t)$, given by

$$\mathrm{E}({}_{\Delta t}T_t) = \mathrm{E}\{T \mid T \in (t, t+\Delta t], T \geq t\} = \frac{\int_t^{t+\Delta t} S(u)\,\mathrm{d}u}{S(t)}. \tag{1.12}$$

In later chapters, a large number of nonparametric, parametric, and semi-parametric lifetime functions will be delineated, analyzed, and discussed. All those more complex models build upon the above basic specifications. In other words, more complicated forms of various survival models are just extensions of the basic functions. No matter how difficult an equation looks, from one function other lifetime indicators can be mathematically defined and estimated.

1.6.2 Discrete lifetime functions

If the distribution of event time T is discrete, the length of time axis t can be divided into J time intervals with unit Δt and a discrete time interval $(t,\ t + \Delta t)$. Given this, t becomes a discrete random variable denoted by t_j ($t_j = t_0, t_1, \ldots, t_J$). Accordingly, the discrete probability density function is defined by

$$f(t_j) = \Pr(T \in t_j), \quad i = 1, 2, \ldots, J. \tag{1.13}$$

Given Equations (1.1) and (1.2), the discrete survival function is

$$S(t) = \sum_{j|t_j > t} f(t_j). \tag{1.14}$$

Likewise, the discrete hazard function can be derived from an extension of Equation (1.4):

$$h(t_j) = \frac{f(t_j)}{S(t_j)}, \tag{1.15}$$

where $S(t_j)$ is the expectation of the survival probability with respect to the discrete time interval t_j. Conceptually, $S(t_j)$ differs from $S(t)$ because it represents the average survival

probability with respect to a discrete time interval, rather than at an instantaneous time point. The deviation of $S(t_j)$ from $S(t)$ depends on the interval unit Δt. If $\Delta t \to 0$, $S(t_j) = S(t)$; if Δt does not represent an infinitesimal time unit but is small, $S(t_j) \approx S(t)$, and the difference between the continuous and discrete survival functions is ignorable. If Δt represents a considerable width, such as a week, a month, or even a year, the continuous $S(t)$ is a decreasing function within the interval, so $S(t_j) < S(t)$. Roughly, the discrete time hazard function can be considered the approximate conditional probability of failure in $(t, t + \Delta t)$. There are some conceptual problems in the specification of this approximation because the hazard rate can be greater than 1 in some extreme situations. This issue, however, is not discussed further in this text.

In the counting process theory (Andersen *et al*., 1993; Fleming and Harrington, 1991), which will be described in Chapter 6, the continuous course of survival is restricted within a limited time interval $(0, \tau)$ where $\tau < \infty$ for a population of N where $N < \infty$ (Aalen, 1975; Andersen and Gill, 1982; Andersen *et al*., 1993; Fleming and Harrington, 1991). Given such restrictions, it is reasonable to view the hazard rate as the conditional probability if N is large. Consequently, the hazard function and the conditional probability are used interchangeably in verifying the validity of counting processes and the martingale theory.

In demographic and epidemiologic studies, researchers often calculate the death rate within a time interval of a considerable width (one year or five years) for measuring the force of mortality for a population of interest (Keyfitz, 1985; Schoen, 1988; Siegel and Swanson, 2004). If time t is expressed as a starting exact age and Δt as the unit of an age interval, the discrete death rate in the interval $(t, t + \Delta t)$, defined as ${}_{\Delta t}M_t$, is written by

$$ {}_{\Delta t}M_t = \frac{N_t - N_{t+\Delta t}}{\pi N_t + (1-\pi)N_{t+\Delta t}} = \frac{S(t) - S(t+\Delta t)}{\pi S(t) + (1-\pi)S(t+\Delta t)}, \tag{1.16} $$

where N_t is the population at t, $N_{t+\Delta t}$ is the population at $(t + \Delta t)$, and π is some weight assigned to derive an unbiased estimate of exposure for the risk of death. Here, $S(t_j)$ is calculated as a weighted average of $S(t)$ and $S(t + \Delta t)$ because, within a wide time interval, not all individuals surviving to t are at the risk for the entire interval (Teachman, 1983b). As a result, the continuous survival probability is a decreasing function within the interval, thereby leading to the condition $S(t_j) < S(t)$. This interval-specific measure for the force of mortality can be conveniently viewed as the discrete realization of the following ratio of two integrals:

$$ {}_{\Delta t}M_t = \frac{\int_t^{t+\Delta t} S(u)h(u)\,\mathrm{d}u}{\int_t^{t+\Delta t} S(u)\,\mathrm{d}u}, \tag{1.17} $$

where the numerator is the cumulative probability densities within the interval between t and $(t + \Delta t)$ and the denominator is the exposure to the risk of dying. Then, when Δt tends to 0, ${}_{\Delta t}M_t = f(t)/S(t) = h(t)$.

From Equations (1.3) and (1.12), the interval-specific force of mortality can be written by

$$ {}_{\Delta t}M_t = \frac{{}_{\Delta t}F_t}{{}_{\Delta t}T_t \times S(t)}, \tag{1.18} $$

where ${}_{\Delta t}F_t$ is the cumulative densities in $(t, t + \Delta t)$ and ${}_{\Delta t}T_t$ is the expected life lived within this specific interval. Here, the hazard rate serves as a step function inherent in the interval with $S(u)$ decreasing due to the elimination of deaths, so that ${}_{\Delta t}M_t$ can be regarded as an average hazard rate with respect to a specific time interval (Siegel and Swanson, 2004). If $h(u)$, where $u \in (t, t + \Delta t)$, is constant throughout the entire interval, ${}_{\Delta t}M_t$ can be regarded as an estimate of $h(u)$.

1.6.3 Basic likelihood functions for right, left, and interval censoring

A likelihood function with survival data describes the probability of a set of parameter values given observed lifetime outcomes. Mathematically, it is either equal to or approximately proportional to the probability of survival data. In this subsection are several simple likelihood functions for right, left, or interval censoring. These likelihoods are basic functions that will serve as a basis for more complicated likelihood functions described in later chapters.

When right censoring occurs, the only information known to the researcher is survival time at the occurrence of censoring. Statisticians utilize this partial information of right censoring when developing a survival model. Specifically, the information of right censored survival times can be well integrated in a likelihood function of survival data.

For a specific observation i, the lifetime process can be described by three random variables: (1) a random variable of event time T_i, (2) a random variable of time t_i, given by

$$t_i = \min(T_i, C_i), \tag{1.19}$$

and (3) a random variable indicating status of surviving or right censoring for t_i, specified by

$$\begin{cases} \tilde{r}_i = 0 \text{ if } T_i = t_i \text{ or } T_i \le C_i \\ \tilde{r}_i = 1 \text{ if } T_i > t_i \text{ or } T_i > C_i, \end{cases} \tag{1.20}$$

where $\tilde{r}_i$ designates whether t_i is a lifetime ($\tilde{r}_i = 0$) or a right censored time ($\tilde{r}_i = 1$). Given these three random variables, the likelihood function for a Type I right censored sample, in which C is fixed as the time distance between the date of entry and the end of study, can be written as the probability distribution of $(t_i, \tilde{r}_i)$. The joint probability density function is given by

$$f(t_i)^{1-\tilde{r}_i} \Pr(T_i > C_i)^{\tilde{r}_i}, \tag{1.21}$$

where $f(\cdot)$ is the probability density function. It follows that, when $\tilde{r}_i = 1\,(t_i = C_i)$, Equation (1.21) reduces to $\Pr(T_i > C_i)$, which is the survival probability at time t, because the first term is 1. Likewise, when $\tilde{r}_i = 0$, Equation (1.21) yields the probability density function $f(t_i)$ because the second term is 1. Assuming the lifetimes $T_1, \ldots, T_n$ for a sample of n are statistically independent and continuous at t_i, the likelihood function for the sample is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i)^{1-\tilde{r}_i} S(t_i)^{\tilde{r}_i}, \tag{1.22}$$

where $S(\cdot)$ is the survival function and $\boldsymbol{\theta}$ is the parameter vector to be estimated in the presence of right censoring.

For a sample of random right censoring, C is no longer fixed but behaves as a continuous random variable. As a result, there are actually two survival functions, from failure and from censoring, and two corresponding densities in the probability distribution. In this case, Equation (1.22) still applies because the survival and density functions for random right censoring are not associated with parameters in $f(t)$, so that they can basically be neglected (see Lawless, 2003, pp. 54–55).

In terms of left censored observations, the likelihood function is associated with different censoring mechanisms. As left censoring occurs before the time of observation, the random variable indicating status of surviving or left censoring at t_i is defined as

$$\begin{cases} \tilde{l}_i = 0 \text{ if } T_i > t_i \\ \tilde{l}_i = 1 \text{ if } T_i \le t_i, \end{cases} \tag{1.23}$$

where $\tilde{l}_i$ denotes whether t_i is a lifetime ($\tilde{l}_i = 0$) or a left censored time ($\tilde{l}_i = 1$). Given this, the likelihood function for left censored observations can be written as another joint probability distribution linked with $(t_i, \tilde{l}_i)$:

$$F(t_i)^{\tilde{l}_i} \Pr(T_i > t_i)^{1-\tilde{l}_i}, \tag{1.24}$$

where $F(t_i)$ is the cumulative distribution function (c.d.f.). It follows that, when $\tilde{l}_i = 0$ $(T_i > t_i)$, Equation (1.24) represents the survival probability S_i at time t_i because the first term is 1; when $\tilde{l}_i = 1$, Equation (1.24) becomes the c.d.f. $F(t_i)$ because the second term is 1. Consequently, given a series of lifetimes $T_1, \ldots, T_n$, the likelihood function for a left censored sample is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} F(t_i)^{\tilde{l}_i} S(t_i)^{1-\tilde{l}_i}. \tag{1.25}$$

As interval censoring is associated with a time range within which a particular event occurs, we have $t_{i-1} < T \le t_i$, and the contribution to the likelihood is simply $F(t_i) - F(t_{i-1})$ or $S(t_{i-1}) - S(t_i)$. Accordingly, the overall likelihood function for interval censoring is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [F(t_i) - F(t_{i-1})]^{\tilde{i}_i} S(t_i)^{1-\tilde{i}_i}, \tag{1.26}$$

where $\tilde{i}_i$ is the status indicator for interval censoring (1 = interval censored; 0 = else).

If the survival data are mixed with right, left, and interval censored observations, the total likelihood function is written by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} [1 - F(t_i)]^{\tilde{r}_i} F(t_i)^{\tilde{l}_i} [F(t_i) - F(t_{i-1})]^{\tilde{i}_i} \\ &= \prod_{i=1}^{n} S(t_i)^{\tilde{r}_i} F(t_i)^{\tilde{l}_i} [F(t_i) - F(t_{i-1})]^{\tilde{i}_i}. \end{aligned} \tag{1.27}$$

Mathematically, maximizing the above likelihood function yields the maximum likelihood estimate of $F(t)$. In Chapters 4 and 5, more complex likelihood functions will be described for specifying more parameters in $\boldsymbol{\theta}$.

1.7 Organization of the book and data used for illustrations

The remainder of the book is organized as follows. Chapter 2 is devoted to some descriptive approaches that are largely applied in survival analysis, including the Kaplan–Meier (product-limit) and the Nelson–Aalen estimators, calculation of the variance, the confidence interval, and the confidence bands for the survival function, the life table methods, and several testing techniques for comparing two or more group-specific survival functions. The applicability of these descriptive methods is discussed. Chapter 3 describes some popular parametric distributions of survival times with mathematical details. Chapter 4 focuses on the description of parametric regression models, with covariates involved in the analysis of survival data. General parametric regression modeling and the corresponding statistical inference are presented as a unique statistical approach combining a known parametric distribution of survival times with multivariate regression procedures. Several widely used parametric models are delineated extensively with empirical illustrations. Given its widespread applicability and flexibility, the Weibull regression model is particularly heeded to and discussed.

Chapters 5 through 7 are devoted mainly to the Cox model and its advancements. In particular, Chapter 5 describes basic specifications of the Cox model and partial likelihood. Some advances in estimating the Cox model are also presented and discussed in this chapter, such as the statistical techniques handling tied observations, the creation of a survival function without specifying an underlying hazard function, the hazard model with time-dependent covariates, the stratified proportional hazard model, modeling of left truncated survival data, and the specification of several popular coding schemes for qualitative factors and the statistical inference of local tests in the Cox model. Chapter 6 first introduces basic specifications of counting processes and the martingale theory, with particular relevance to the Cox model. Then I present, in order, five types of residuals used in the Cox model, techniques for the assessment of the proportional hazards assumption, methods of evaluating the functional form for a covariate, and approaches on identification of influential observations in the Cox model. Each of these sections is supplemented by an empirical illustration. Chapter 7 displays statistical techniques for analyzing survival data with competing risks and with repeated events. In addition to a step-by-step presentation on empirical data, the merits and limitations in various statistical techniques in those areas are discussed in this chapter.

Chapter 8 briefly discusses advantages and existing problems in the application of the structural hazard rate models. A simplified structural hazard model is specified with restrictive assumptions, with a detailed example provided for illustrating step-by-step procedures. Chapter 9 concerns several special topics in survival analysis, including informative censoring, bivariate and multivariate survival models, the frailty theory, mortality crossovers and maximum life span, survival convergence and the preceding mortality crossover, and the calculation of sample size required for survival analysis. The strengths and limitations of each of those advanced techniques are discussed. Due to consideration of coherence and

conciseness of the text, some supplementary procedures, datasets, and computer programs are presented as appendices.

In this book, data from a large-scale longitudinal study is the major source for illustrating empirical examples. These data come from the Survey of Asset and Health Dynamics among the Oldest Old (AHEAD), a nationally representative investigation of older Americans. This survey, conducted by the Institute for Social Research (ISR), University of Michigan, is funded by the National Institute on Aging as a supplement to the Health and Retirement Study (HRS). The HRS Wave I survey was conducted in 1992, with a sample size of 12 652 persons including selected respondents aged between 40 and 70 years and their spouses, if married, regardless of age. Those respondents have been followed up by telephone every other year, with proxy interviews for those who have deceased prior to follow-up.

As a supplemental survey to the HRS, Wave I of the AHEAD survey was conducted between October 1993 and April 1994. Specifically, a sample of individuals aged 70 or older (born in 1923 or earlier) was identified throughout the HRS screening of an area probability sample of households in the nation. This procedure identified 9473 households and 11 965 individuals in the target area range. Like the HRS, the Wave I respondents have been followed up by telephone every second or third year, with proxy interviewing designed for those deceased between two successive surveys. At present, the AHEAD survey registers nine waves of investigation in 1993, 1995, 1998, 2000, 2002, 2004, 2006, 2008, and 2010. As a longitudinal, multidisciplinary, and US population-based study, AHEAD provides a highly representative and reliable data base for the survival analysis of older Americans aged 70 or older.

AHEAD acquires detailed information on a number of domains, including demographic characteristics, health status, health care use, housing structure, disability, retirement plans, and health and life insurance. Survival information throughout the follow-up waves has been obtained by a link to the data of the National Death Index (NDI). This book uses survival data of the AHEAD survey throughout the first six waves (1993–2004). Given the illustrative nature of using this dataset, I randomly selected 2000 persons from the baseline AHEAD sample for the analysis in the book.

While AHEAD provides a solid and reliable data source for empirical illustrations, this single dataset cannot cover the entire scope of this book. As a result, in addition to the AHEAD longitudinal data, some clinical and simulated data are used when appropriate. In performing empirical analyses, I used the SAS software package for programming a variety of statistical procedures on data management and data analysis.

1.8 Criteria for performing survival analysis

This book introduces, delineates, and summarizes a large number of survival models and techniques. Given the statistical methods and techniques presented here, the reader, after learning how to apply each one of them, might raise more questions about performing survival analysis. Given a topic of interest and the availability of a dataset, what are the criteria for performing survival analysis? In other words, what is the most appropriate perspective for conducting a survival analysis? There are four underlying criteria – relevance to theory, accurate description of data, computational tractability, and interpretability of analytic results. As the focus of this book is on application and practice, these criteria are established with specific regard to the applied phases of survival analysis.

In performing survival analysis, the first criterion is whether a statistical model builds upon an existing theory on a particular event of interest. A lifetime event usually involves complex mechanisms. Whereas one factor can influence some other factors in shaping the risk and frequency of experiencing a particular event, at the same time it may be affected by others, thus constituting a complicated structure of causal linkages in survival processes. Such complexity of causality makes it extremely difficult, if not impossible, to describe a complete set of interrelationships using a single conceptual model. Additionally, there is often a lack of suitable measures for many conceptual factors in the description of event histories. A practical approach to deal with these issues is to link a lifetime process with an explicit theoretical framework that portrays a portion of the causal effects on a particular event. The underlying conceptual framework must be based on a relevant, well-established theory or previous findings. A good theory gives rise to valid theoretical hypotheses, provides direction for specifying interrelationships of conceptual factors, and facilitates selection of measurable variables for explaining the dynamics of a lifetime phenomenon. This strategy can aid in yielding new constructive findings, in turn further solidifying the connotation of the underlying theory. In some sense, creating a survival model without relevance to theory is just like building an edifice without reference to a blueprint.

The second criterion for performing survival analysis is accurate description of data. Except for some special cases, survival models are created and applied for the generation of unbiased parameter estimates, the prediction of trajectories of survival processes, and the derivation of analytic results. Much of survival analysis is based on empirical data from various sources, such as large-scale sample surveys, clinical trials, and vital registration statistics. Given the usual representativeness of such information, it is essential for the scientist to describe the empirical data accurately, with model-based lifetime trajectories agreeable to observed patterns. Considerable deviations of model-derived trajectories from observed data generally indicate the misspecification of an underlying survival model because it fails to describe data correctly by using parameter estimates. In other words, an incorrect description of empirical data suggests failure to reflect the true set of experiences generated by the stochastic processes. Therefore, an accurate description of data is a required condition for making further inference on model fitness and the quality of various parameter estimators.

The third criterion for performing survival analysis is computational tractability. In survival analysis, a statistical model needs to be developed in such a way that the audience can follow what computational procedures are applied and how analytic results are produced, and thus understanding the rationale of the analysis. In particular, specifications of covariates, mathematical functions, and causal relationships among variables must be explicit and unambiguous, and the estimation of an underlying survival model must be based on statistical procedures and steps that are recognizable and applicable. New computational procedures, if necessary, must be fully presented, with methodological inference self-contained, for allowing the audience to understand justifications of the advancement. Exceptionally complicated link functions and reader-unfriendly mathematical expressions need to be avoided whenever possible, and the addition of model parameters should be well justified analytically. Indeed, if not computationally tractable, a survival model cannot be readily accepted by other researchers, thus failing to be disseminated in a timely fashion.

The fourth criterion is interpretability of analytic results. In survival analysis, most analysts' professional orientations are toward applied areas, and the material described in most survival analysis books is designed to demonstrate the variety of approaches in which

survival models can be applied in various disciplines. This nature of applicability points to the importance of generating interpretable and substantively meaningful analytic results when performing survival analysis. Survival outcomes are generally modeled as a nonlinear function of explanatory variables, so that model parameters are sometimes not directly interpretable. Under such circumstances, the researcher needs to convert unexplainable results into interpretable ones by means of techniques of functional transformation. If hard to interpret, analytic results of a survival model cannot translate into useful explanations, in turn obstructing timely spreading of potentially valuable implications.

In all areas of applied science, statistical modeling, survival analysis being no exception, is by no means an easy task. In constructing a useful survival model, both overfitting and underfitting must be avoided while maintaining relevance to a substantive theme. By abiding by these four criteria, the difficulty in creating a good survival model can be considerably mitigated. As Box (1976) asserts, all models are wrong, and scientists cannot obtain a 'correct' one by excessive elaboration; nevertheless, if a statistical model correctly describes the essence and the trend of a phenomenon while overlooking ignorable noises, it is a useful perspective for the scientist to apply. Given this principle, many hypotheses and assumptions in statistical modeling are established just for analytic convenience and simplicity: there would rarely be a true normal distribution, there never exists an exactly linear association between two factors, and a conceptual framework can look more like an artifact than a reflection of the real world. All these steps are necessary otherwise a statistical model would become a garbage can. For survival analysis in particular, the performance of a statistical technique is a process of abstraction, rather than a course of mirroring what we see in our daily lives. If a parsimonious model with only a few parameters generates the same statistical power and the same substantive implications as does a more complicated one, the former method is the signature of a good model and the latter a mediocre one (Box, 1976).