CHAPTER

Introduction 1

CHAPTER OUTLINE

1.1 WHAT IS LONGITUDINAL DATA ANALYSIS?

We live in a dynamic world full of change. A person grows, ages, and dies. During that process, we may contract disease, develop functional disability, and lose mental ability. Accompanying this biological life course, social change also occurs. We attend school, develop a career and retire. 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 activities according to our physical and mental conditions. Indeed, change characterizes almost all aspects of our social lives, ranging from the aforementioned social facets to unemployment, drug use recidivism, occupational careers, and other social events. In these biological and social processes, the gradual changes and developments over a life course reflect a pattern of change over time. More formally, such changes and developments may be referred to as an individual's trajectory.

In a wider scope, trajectories are also seen in the pattern of change referring to such phenomena as the decaying quality over time of a commercial product or the collapse of a political system in a country. In the field of business management, change in consumer purchasing behavior is generally linked both with individual characteristics and with competing products. In population studies, demographers

Methods and Applications of Longitudinal Data Analysis. http://dx.doi.org/10.1016/B978-0-12-801342-7.00001-0

are concerned with such longitudinal processes as internal and international migration, and intervals between successive births. In cases such as these events and in others, the pattern of change over time can be influenced and determined by various factors, such as genetic predisposition, illness, violence, environment, medical and social advancements, or the like. Therefore, each trajectory can differ significantly among individuals and other observational units, or by the variables that govern the timing and rate of change in a period of time.

Data available at a single point of time does not suffice to analyze change and its pattern over time. Cross-sectional data, traditionally so popular and so widely used in a wide variety of applied sciences, only designates a snapshot of a course and thus does not possess the capacity to reflect change, growth, or development. Aware of the limitations in cross-sectional studies, many researchers have advanced the analytic perspective by examining data with repeated measurements. By measuring the same variable of interest repeatedly at a number of times, the change is displayed, its pattern over time revealed and constructive findings are derived with regard to the significance of change. Data with repeated measurements are referred to as *longitudinal data*. In many longitudinal data designs, subjects are assigned to the levels of a treatment or of other risk factors over a number of time points that are separated by specified intervals.

Analyzing longitudinal data poses considerable challenges to statisticians and other quantitative methodologists due to several unique features inherent in such data. First, the most troublesome feature of longitudinal analysis is the presence of missing data in repeated measurements. In a longitudinal survey, the loss of observations on the variables of interest frequently occurs. For example, in a clinical trial on the effectiveness of a new medical treatment for disease, patients may be lost to a follow-up investigation due to migration or health problems. In a longitudinal observational survey, some baseline respondents may lose interest in participating at subsequent times. These missing cases may possess unique characteristics and attributes, resulting in the fact that data collected at later time points may bear little resemblance to the sample initially gathered. Second, repeated measurements for the same observational unit are usually related because average responses usually vary randomly between individuals or other observational units, with some being fundamentally high and some being fundamentally low. Consequently, longitudinal data are clustered within observational units. In the meantime, an individual's repeated measurements may be a response to a time-varying, systematic process, resulting in serial correlation. Third, longitudinal data are generally ordered by time either in equal space or by unequal intervals, with each scenario calling for a specific analytic approach. Sometimes, even with an equal-spacing design, some respondents may enter a follow-up investigation after a specified survey date, which, in turn, imposes unequal intervals for different individuals.

Over the years, scientists have developed a variety of statistical models and methods to analyze longitudinal data. Most of these advanced techniques are built upon biomedical and psychological settings, and therefore, these methodologically advanced techniques are relatively unfamiliar to researchers of other disciplines. To

date, many researchers still use incorrect statistical methods to analyze longitudinal data without paying sufficient attention to the unique features of longitudinal data. For these researchers, the advanced models and methods developed specifically for longitudinal data analysis can be readily borrowed for use after careful verification, evaluation, and modification. In health and aging research, for example, the pattern of change in health status is generally the main focus. In analyzing such longitudinal courses, failure to use correct, appropriate methods can result in tremendous bias in parameter estimates and outcome predictions. In these areas, the application of advanced models and methods is essential.

1.2 HISTORY OF LONGITUDINAL ANALYSIS AND ITS PROGRESS

There were some vague, sporadic discussions about the theory of random effects and growth as early as the nineteenth century (Gompertz, 1820; Ware and Liang, 1996). The year 1918 witnessed the advent of the earliest repeated measures analysis when Fisher (1918) published the celebrated article on the analysis of variance (ANOVA). In this historical masterpiece, Fisher introduced variance-component models and the concept of "intraclass correlation." Some later works extended Fisher's approach to the domain of mixed modeling with the developments of such concepts as the split-plot design and the multilevel ANOVA (Yates, 1935; Jackson, 1939). For a long period of time, these variance decomposition methods were the major statistical tool to analyze repeated measurements. Though simplistic in many ways, the advancement of these early works provided a solid foundation for the advancement of the modern mixed modeling techniques. Around the same period, there were also some early mathematical formulations of trajectories to analyze the pattern of change over time in biological and social research (Baker, 1954; Rao, 1958; Wishart, 1938; see the summary in Bollen and Curran, 2006, Chapter 1). Until the early 1980s, however, longitudinal data analysis was largely restricted within the formulation of the classical repeated measures analysis traditionally applied in biomedical settings.

Given the substantial limitations and constraints in the traditional approaches in repeated measures analysis, many methodologists expressed grave concerns regarding how to measure and analyze the pattern of change over time correctly (see the summary in Singer and Willett, 2003, Chapter 1). Over the past 30 years, longitudinal data analysis has grown tremendously as a consequence of the rapid developments in mixed-effects modeling, multilevel analysis, and individual growth perspectives. Accompanying these developments in statistical models and methods are the equally important advancements in computer science, particularly the powerful statistical software packages. The convenience of using computer software packages to create and utilize complex statistical models has made it possible for many scientists to analyze longitudinal data by applying complex, efficient statistical methods and techniques, once considered impossible to accomplish (Singer and Willett, 2003).

As applications of various statistical techniques on longitudinal data have grown, methodological innovation has accelerated at an unprecedented pace over the past three decades. The advent of the modern mixed-effects modeling and the various approaches for the analysis of longitudinal data triggered the advancement of a large number of statistical models and methods, characterized by the complex procedures of multivariate regression. The major contribution of mixed-effects models, with their capability of containing both fixed and the random effects, is the provision of a flexible statistical approach to model the autoregressive process involved in the trajectory of individuals, both for average change across time and change for each observational unit. Given such a powerful perspective, both the measurable covariates and the unobservable characteristics can be incorporated in the model simultaneously, thereby deriving more reliable analytic results for the description of a longitudinal process. Being robust to missing data in general circumstances, mixed-effects models also have the added advantage of permitting irregularly spaced measurements across time.

More recently, a variety of Bayes-type approximation methods have been advanced to estimate parameters in the analysis of longitudinal data characterized of nonlinear functions such as proportions and counts. Given their flexibility in modeling nonnormal outcome data, these approximation techniques have enabled researchers to estimate random effects with complex structures and to correctly perform nonlinear predictions. To date, the various approximation methods have been applied by statisticians and some other quantitative methodologists to develop more statistically refined longitudinal models, thus expanding the capacity of mixed-effects modeling in longitudinal data analysis to a new dimension. In a different track, some other methodologists have advanced growth curve modeling by introducing latent factors or/and latent classes within the framework of structural equation modeling (SEM).

1.3 LONGITUDINAL DATA STRUCTURES

Methodologically, longitudinal data can be regarded as a special case of the classical repeated measures data of individuals that are collected and applied in experimental studies. Strictly speaking, there are some conceptual differences between the two data types: repeated measures and longitudinal analysis. The classical repeated measures data represent a wider concept of data type as they sometimes involve a large number of time points and permit changing experimental or observational conditions (West et al., 2007). In contrast, longitudinal data are more specific. They are generally composed of observations for the same subject ordered by a limited number of time points with equally or unequally spaced intervals. Therefore, longitudinal data can be defined as the data with repeated measurements at a limited number of time points with predetermined designs on time scale, time interval, and other related conditions. In statistics and econometrics, longitudinal data is often referred to as *panel data*.

In this section, longitudinal data structures are delineated. I first review the multivariate data format used traditionally in the classical repeated-measures analysis and

applied presently in latent growth modeling. Second, the univariate data structure is introduced, in which an individual or an observational unit (such as a commercial product) has multiple rows of data to record repeated measurements at a limited number of time points. Lastly, balanced and unbalanced longitudinal data are defined and described.

1.3.1 MULTIVARIATE DATA STRUCTURE

The classical repeated measures data are predominantly used in the ANOVA in experimental studies. Traditionally, the data structure for repeated measures ANOVA follows a multivariate format. In this data structure, each subject only has a single row of data, with repeated measurements being recorded horizontally. That is, a column is assigned to the measurement at each time point in the data matrix. To illustrate the multivariate data structure, I provide an example by using the repeated measures data of the Randomized Controlled Clinical Trial on the Effectiveness of Acupuncture Treatments on PTSD, which will be described extensively in Section 1.7 (PTSD is the abbreviation of posttraumatic stress disorder). The PTSD Checklist (PCL) score is the response variable to gauge severity of PTSD symptoms, a 17-item summary scale measured at four time points. The value range of the PCL score is from 17 to 85. In the multivariate data format, the repeated measurements for each subject are specified as four outcome variables lined in the same row, with time points indicated as suffixes attached to the variable name. Additionally, two covariates are included in the dataset: Age and Female (male = 0, female = 1). To identify the subject for further analysis, each individual's ID number is also incorporated. Below is the data matrix for the first five subjects in the multivariate data format.

In Table 1.1, each subject has one row of data with four outcome variables, PCL1–PCL4, the ID number, and the two covariates, Age and Female. Among the five subjects, one person is aged below 30 years, one above 50, and the rest ranging between 38 and 44 years of age. There are four men and one woman. As all observations for the outcome variable are lined horizontally in the same row, the multivariate data structure of repeated measurements contains additional columns, therefore also referred to as the *wide table* format. Clearly, the cross-sectional data format is a special case of the multivariate structure with the outcome variable being observed only at one time. The most distinctive advantage of using the multivariate data structure

Table 1.1 Multivariate Data of Repeated Measurements

ID	PCL1	PCL2	PCL3	PCL4	Age	Female
1	66	31	58	39	27	0
2	48	56	43	43	44	1
3	37	50	53	47	38	0
4	41	23	21	21	53	0
5	51	57	39	46	44	0

is that each subject's empirical growth record can be visually examined (Singer and Willett, 2003). In Table 1.1, for example, it is easy to summarize each subject's trajectory by comparing values of the repeated measurements horizontally. Further examination on the pattern of change over time in the response variable can be performed visually. Perhaps due to this convenience, various latent growth models, which constitute an integral part of the literature on longitudinal data analysis, are designed with such a wide table perspective.

There are, however, distinctive disadvantages for the multivariate data structure in performing longitudinal data analysis. First, time is the primary covariate in analyzing the pattern of change over time in the response variable. In the wide table format, the time factor is indirectly reflected by the suffix attached to each time point, and therefore, time is not explicitly specified as an independent factor, thereby bringing inconvenience in the analysis of the time effect. Sometimes, intervals between two successive waves are unequally spaced by design or vary across subjects, and the multivariate data structure obviously cannot reflect such variations in spacing. Second, in longitudinal data analysis values of some covariates may vary over time, and failure to address such time-varying nature in the predictor variables can result in bias in analytic results and erroneous predictions of longitudinal processes. There are some complex, cumbersome ways to specify time-varying covariates within the multivariate data framework; these approaches, however, are not user-friendly and are inconvenient to apply.

1.3.2 UNIVARIATE DATA STRUCTURE

Given the aforementioned disadvantages in the multivariate data format, much of the modern longitudinal modeling is performed on the basis of data of univariate structure, in which each subject has multiple rows and time is explicitly specified as a primary predictor on the developmental trajectory of individuals. That is, assuming n time points in a longitudinal data analysis, each subject is assigned n rows of data in the dataset (empirically, some of the rows may be empty due to loss of observation at follow-ups). Table 1.2 displays the same data presented in Table 1.1 but using the univariate data structure.

In Table 1.2, each subject has four rows of data, with each record corresponding to a specific time point. Therefore, in the univariate data format, each subject is assigned a block of records, rather than a single line. Correspondingly, the construction of a univariate longitudinal dataset is said to be based on a *block design*. The repeated measurements of PCL score are now set vertically, with the same name PCL, and suffixes are removed. A new covariate, Time, is added to the data matrix to indicate a specific time point, and a combination of values for the PCL and the time variables designate repeated measurements at four time points. Given the time points, the ID number for each subject is repeated four times in the data matrix. Likewise, as the predictors measured at baseline, the same values of Age and Female for a specific subject are also recorded four times. As subject-specific observations are set vertically, fewer columns but more rows are specified than the multivariate data structure.

Table 1.2 Univariate Longitudinal Data

ID	Time	PCL	Age	Female
1	0	66	27	0
1	1	31	27	0
1	2	58	27	0
1	3	39	27	0
2	0	48	44	1
2	1	56	44	1
2	2	43	44	1
2	3	43	44	1
3	0	37	38	0
3	1	50	38	0
3	2	53	38	0
3	3	47	38	0
4	0	41	53	0
4	1	23	53	0
4	2	21	53	0
4	3	21	53	0
5	0	51	44	0
5	1	57	44	0
5	2	39	44	0
5	3	46	44	0

Therefore, the univariate longitudinal data structure is also referred to as the *long table* format.

Compared to the multivariate data format, the univariate longitudinal data matrix contains the identical information and differs only in structure and the addition of a time factor. Organizing longitudinal data into the univariate structure, however, boasts several distinctive advantages as contrasting to the limitations attached to the multivariate format. First, as time is specified as a direct and explicitly defined predictor variable, the researcher can handle unequally spaced intervals between time points and across individuals effectively by setting specific values of time. Second, the covariate's values for the same individual are recorded vertically in multiple rows, thereby facilitating the specification of time-varying covariates in a straightforward, convenient fashion. Given these strengths, the univariate structure has becomes the most popular data format in longitudinal data analysis.

1.3.3 BALANCED AND UNBALANCED LONGITUDINAL DATA

When conducting longitudinal data analysis, the researcher needs to determine whether the data are "balanced" or "unbalanced." In the classical ANOVA model, balanced repeated-measures data indicate an equal number of observations for all

possible combinations of factor levels (such as multiple treatment factors or marital status groups), whereas "unbalanced" specify an unequal number of observations. In longitudinal data analysis, researchers usually use a looser definition to distinguish between a balanced and an unbalanced data design by considering the number of time points, timing, and spacing of intervals. When a set of time points are so designed as to be common to all individuals, the study design is said to be balanced over time. If data are collected exactly in accordance with such a balanced design, with all subjects having the same set of repeated measurements and there are no missing observations, the resulting longitudinal data are referred to as balanced. Balanced design is regularly applied in clinical experimental studies, in which subjects are randomized into each of the treatments for evaluation with relatively narrow intervals between successive time points. If there is no delayed entry and the follow-up information is complete, such repeated measures data are considered balanced. This definition of balanced data holds for repeated measurements that are both equally and unequally spaced as long as all individuals are followed at the same intervals.

When the number of time points is designed to be uncommon to subjects, the longitudinal data design is said to be unbalanced. Longitudinal data resulting from an unbalanced design are generally unbalanced, with subjects having different sets of time points for repeated measurements. In aging and health research, longitudinal data are sometimes collected with an unbalanced design. In older persons, for example, longitudinal attrition is usually expected to be high due to mortality and other health-related reasons, and therefore, individuals vary as some exit due to death, illness, out-of-scope residence, and the like. As a result, the sample size tends to dwindle rapidly at later waves, thereby being too small to conduct an efficient analysis, particularly when a random sample of individuals is followed for a long period of time. In these situations, researchers can randomly select a new sample at certain follow-up waves to replenish the dwindling sample size. Consequently, subjects in longitudinal data are expected to have different measurement time points. Given different sets of time points, such longitudinal data are unbalanced. Unbalanced longitudinal design can also arise when the timing of measurements is designed to be associated with certain benchmark events such as body fat prior to and right after menarche (Fitzmaurice et al., 2004).

In longitudinal data analysis, most designs follow the balanced perspective, whereas the vast majority of longitudinal data turn out to be unbalanced. Due to the unavoidable presence of missing data, a balanced design routinely results in unbalanced longitudinal data. Those dropping out from observation have fewer time points than the completers, and consequently, the presence of missing data automatically ushers in an unbalanced data structure. In clinical experimental studies with a balanced design, some subjects may enter the study considerably later than a specified starting date, thereby being listed as a delayed entry with timing of measurements different from other respondents. Longitudinal data then switch to an unbalanced structure. For many observational studies, all respondents at baseline are designed to be reinvestigated at fixed intervals. Such a balanced design, however, can rarely lead to balanced longitudinal data given the regular occurrence of attrition in repeated

measurements. Given incomplete information on change, large-scale missing data have caused serious concerns in longitudinal data analysis, particularly for generating the pattern of change over time for individuals. Fortunately, statisticians and other quantitative methodologists have developed a variety of statistically efficient and robust methods to address the impact an unbalanced data structure has, as will be described and discussed in the succeeding chapters.

1.4 MISSING DATA PATTERNS AND MECHANISMS

As indicated in Section 1.1, the presence of missing data in repeated measurements is one of the primary problems to contend with in longitudinal data analysis. In longitudinal processes, missing data regularly occur given dropouts or unanswered items. As designed, longitudinal data are collected in a particular observational period of time in which the outcome and other relevant variables are recorded sequentially at a set of previously designed intervals. Therefore, the researcher can only observe responses for those who are available at each follow-up time point. There are different types of missing data. Some missing data represent a random sample of all cases, and more often, missing observations are not random but the information inherent in them can be addressed by certain observed variables such as age, gender, and health status. In these situations, missing data do not pose serious threats to the quality of a longitudinal data analysis. In some circumstances, however, missing data are related to missing values of the outcome variable, and ignoring such systematic missing data can be detrimental to the estimation and prediction of the pattern of change over time in an outcome variable of interest. Given these missing data types, it is essential for the researcher to understand various missing data patterns and mechanisms before proceeding with a formal longitudinal data analysis.

Missing data can be grouped according to the missing data pattern, which describes which values are observed and which values are missing in the data matrix. In general terms, missing data patterns can be roughly classified into a variety of groups, such as univariate, multivariate, monotone, nonmonotone, and file matching (Little and Rubin, 2002). A univariate missing pattern indicates the situation where missing data occur only in a single variable. As an extension of the univariate case, the multivariate missing pattern refers to missing data in a set of variables, either for the entire unit or for particular items in a questionnaire. If a variable is missing for a particular subject not only at a specific time point but also at all subsequent time occasions, the missing data pattern for this individual is said to be a *monotone missing pattern*. In contrast, if a case is missing at a given time point and then returns at a later follow-up investigation, then the missing data pattern for this subject is referred to as the *nonmonotone missing data*. In longitudinal data analysis, the nonmonotone missing pattern can cause more problems than does the monotone pattern, and thus deserves close attention. Sometimes, variables are not observed together, and such missing data are referred to as a file-matching pattern. There are more missing data patterns such as the latent-factor patterns with variables that are never observed. For

each missing data patterns, there are corresponding statistical techniques to handle its impact on the quality of a longitudinal data analysis.

Missing data mechanisms concern the relationship between missing data and the values of variables in the data matrix. Given this focus, missing data mechanisms can be categorized into three classes: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). If missing data are unrelated to both the missing responses and the set of observed responses, the observed values are representative of the entire sample without missing values. This missing data mechanism is referred to as MCAR. If missing data depend on the set of observed responses but unrelated to the missing values, the missing data are said to be MAR. In longitudinal data analysis, the majority of missing data are categorized as MAR. Correspondingly, most longitudinal models are based on the MAR hypothesis in handling missing values.

In some special situations, missing data are related to specific missing values, referred to as MNAR. In longitudinal data analysis, ignoring the impact of this missing data mechanism can result in serious bias in analytic results and erroneous predictions. Over the years, scientists have developed a variety of empirically sound models and methods to account for MNAR in longitudinal data analysis. Compared to the methods for MAR, however, the statistical models handling MNAR are considered less mature, and research in this regard is ongoing.

With the importance of missing data analysis on longitudinal data, in this book an entire chapter (Chapter 14) is devoted to this area. In that chapter, a section is provided on the mathematical definitions and conditions associated with MCAR, MAR, and MNAR, respectively. Various statistical models handling different missing data mechanisms are described extensively in Chapter 14.

1.5 SOURCES OF CORRELATION IN LONGITUDINAL PROCESSES

Another primary feature in longitudinal data is the presence of correlation in repeated measurements for the same subject, referred to as *intraindividual correlation*. Such correlation highlights the violation of the conditional independence hypothesis regularly applied in multivariate regression modeling. Such a lack of independence in repeated-measures data has driven statisticians and other quantitative methodologists to find ways for correctly performing longitudinal data analysis. There are two perspectives to address intraindividual correlation, each linked to a specific source of variability in longitudinal data. Statistically, variability in longitudinal processes can be summarized into three components: between-subjects variability, within-subjects variations, and the remaining random errors. In the literature of longitudinal data analysis, the first two components of variability, between-subjects and within-subject, address the systematic part inherent in variations in longitudinal processes, and therefore, intraindividual correlation can be modeled by means of including the pattern of variability from those two components. The third component is the random

term for uncertainty as regularly specified in general linear and generalized linear regression models, and therefore, it can be estimated as regression residuals.

Between-subjects variability reflects individual heterogeneity in unspecified and unrecognized characteristics on the trajectory of the response variable. While much of the overall variability can be statistically addressed by specifying observable individual and contextual covariates, some between-subjects heterogeneity may derive from unobservable biological and environmental factors, such as genetic predisposition and physiological parameters. A sample of individuals may have different reactions to the same dose of a new medication on a specific disease; a good behavioral prototype fostered in childhood can help decelerate functional disablement process as age progresses; the social environment is likely to affect drug use recidivism among adolescents released from a rehabilitative center; and so on. As well documented in the literature of longitudinal data analysis, ignoring such unobserved heterogeneity can result in considerable bias in parameter estimates, especially the estimator of standard errors (Diggle et al., 2002; Fitzmaurice et al., 2004; Verbeke and Molenberghs, 2000). In empirical research, a popular approach to handle between-subjects variability is to specify the individual-specific random effects with the assumption of a known distribution. By accounting for individual differences in the analysis, intraindividual correlation in longitudinal data is implicitly manipulated thereby resulting in conditional independence in repeated measurements of the response for the same observational unit. As will be described in the succeeding chapters, many statistical models and methods applied in longitudinal data analysis are designed to handle intraindividual correlation by the specification of such random effects.

The second component of variability in longitudinal processes is within-subject variations. Given various biological, genetic, and environmental predispositions, the values of repeated measurements for the same individual tend to be more similar than those obtained from several randomly selected individuals. Therefore, intraindividual changes over time in the response variable tend to be confined within a person's physical and environmental mechanisms, resulting in positive correlation in longitudinal data. Another distinctive characteristic in such subject-specific dependence is that intraindividual correlation has frequently been observed to decay over time. For example, the repeated measurements of blood sugar for the same person are very likely to be more similar than those from a number of randomly selected persons. A person's blood sugar measurements also tend to be more similar between two successive time occasions than between two nonadjacent time points. In the literature of repeated-measures analyses, this type of change in correlation is referred to as *serial correlation*. The presence of recognizable patterns of correlation has enabled researchers to account for intraindividual correlation by specifying within-subject covariance structures on repeated measurements.

It must be emphasized that between-subjects and within-subject variations are closely interrelated. Clustering among repeated measurements of the response for the same observational unit can be understood as reflecting differences in the pattern of change over time among individuals. Likewise, individual differences reflect

homogeneity of the repeated measurements within an observational unit. Therefore, in longitudinal data analysis, the researcher often needs to consider only one source of systematic variability in statistical modeling, within-subject or between-subjects, to make the longitudinal data conditionally independent. Given the close interaction between the two sources of variability, very often longitudinal data does not support modeling of all possible components simultaneously, particularly when the outcome variable of interest is qualitative rather than quantitative. Only in some special situations do both between-subjects and within-subject variability need to be specified in the same statistical model, and some cases of this sort will be described in the succeeding chapters.

If the specification of the subject-specific random effects or/and a specific covariance structure primarily account for intraindividual correlation in longitudinal data, the remaining within-subject variability is random and noninformative. Correspondingly, this residual component can be specified as random errors in the same fashion as traditionally assumed in general linear models and generalized linear models. For qualitative response variables, however, specification of within-subject components of variability cannot be easily specified, and some complex statistical procedures are needed for the estimation of random errors.

1.6 TIME SCALE AND THE NUMBER OF TIME POINTS

In designing a longitudinal data analysis, the first question in the researcher's mind is perhaps the scale of time. Time can be measured in a variety of metric units – week, month, year, age, session, and the like. The scale that is selected should depend on the nature of the study and the targeted length of an observation period. In clinical experimental studies, the outcome measurement can be linked with events with both rapid changes in status and with a relatively slow pace. In cancer research, for example, the tempo of status change within a fixed time period varies significantly over different types of tumor. A study of surgical treatment of lung cancer may examine the improvement of the survival rate for 6 months. In this research, week is an 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 patients with this type of tumor typically live much longer than those with lung cancer. Therefore, month is a better option for the second case. In the behavioral and social sciences, the occurrence of a change in status is usually a long, gradual process. Examples of such gradual life course events are recovery from disability among older persons, changes in marital status, and discontinuation of drinking alcohol among heavy drinkers. To follow those processes, month or year may be an appropriate choice of time scale. In health services research, different service types are associated with various cycles of turnover. A patient admitted to a short-stay hospital, for example, stays there only for a few days, whereas the average length of stay in a nursing home may be years (Liu et al., 1997). Accordingly, the time scale in the health services research needs to be specified by the nature of the service type.

The number of time points is another primary concern in designing a longitudinal study. While there are no clear-cut rules about the optimum number of time points, the minimum number should not be two, or even three. From a two-point longitudinal dataset, the only pattern of change over time that can be generalized is linear, no matter how different the values of the response differ between the two time points. Important information about the trajectory of individuals can be missed. Longitudinal data with a design of three time points provide the capacity to describe a possible exponential pattern of change over time, but not a more complex function. Four time points can yield a cubic trajectory of individuals. Therefore, for thoroughly reflecting the possible pattern of change in the response variable, the minimum number of time points is four. Analytically, there does not seem to be a definite maximum number of time points that can be determined for longitudinal data analysis. The desired number of time points should be evaluated on a case-by-case base. A high number of time points can probably improve the precision of estimation on the pattern of change over time if the sample size is sufficiently large. In the meantime, specifying a large number of time points can significantly increase the financial burden to conduct a longitudinal analysis and complicate statistical modeling. I personally recommend that for clinical experimental studies, four to six time points, consisting of the baseline and three to five follow-up waves, should be good enough to display the effectiveness of a new medication or treatment on disease. With respect to observational studies with a large sample size, 6 to 10 time points are ideal for the description of a time trend in the response variable and its group differences.

1.7 BASIC EXPRESSIONS OF LONGITUDINAL MODELING

As introduced in Section 1.3, in a typical longitudinal dataset, each subject has multiple rows of data, with each record corresponding to a specific time point. Suppose that a random sample is composed of N subjects with n predesigned time points. For subject i ($i = 1, 2, \ldots, N$), the observed number of time points is usually denoted n_i that does not necessarily equal n due to missing observations. The response measurement for subject i at time point j ($j = 1, 2, \ldots, n_i$) is written as Y_{ij}. The repeated measurements of the response variable Y for subject i can be expressed in terms of an $n_i \times 1$ column vector, denoted by $\boldsymbol{Y}_i$ and given by

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}.$$

Given N subjects, there are N such vectors in a longitudinal dataset. Repeated measurements of the response variable are generally specified as a function of the time factor and some other theoretically relevant covariates. In the analysis of cross-sectional data, various regression models are generally performed by assuming conditional independence of observations in the presence of specified model parameters.

If the response variable Y has a continuous scale, it seems that a linear regression model on the response variable Y can be written as

$$Y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}, \tag{1.1}$$

where $\boldsymbol{X}_{ij}$ is a $1 \times M$ vector of covariates for subject i at time point j, $\boldsymbol{\beta}$ is a vector of the regression coefficients of M covariates with values fixed for all subjects, and ε_{ij} is the random error term conventionally assumed to be normally distributed. In the context of longitudinal data, however, this classical specification does not derive efficient, robust, and consistent parameter estimates, particularly since the error is related to the random errors at other time points, even in the presence of specified observed covariates. Therefore, some additional parameters addressing intraindividual correlation need to be specified to secure the conditional independence hypothesis in random errors.

The importance of addressing intraindividual correlation has triggered the development of many advanced methods in longitudinal data analysis. As summarized by some leading statisticians in this area (Diggle, 1988; Diggle et al., 2002), modeling a longitudinal process should at least consider three different sources of variability. First, average responses usually vary randomly between subjects, with some subjects being fundamentally high and some being low. Second, a subject's observed measurement profile may be a response to time-varying processes. Third, as the individual measurements involve subsampling within subjects, the measurement process adds a component of variation to the data. Such decomposition of stochastic variations in repeated measurements facilitates the formulation of correlation between pairs of measurements on the same subject. Based on the linear regression model specified in Equation (1.1), the multivariate linear model, including all the aforementioned three sources of variability, can be written as

$$Y_{ij} = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + b_i + \tilde{W}_i(T_{ij}) + \varepsilon_{ij}, \tag{1.2}$$

where b_i indicates the variation in the average response between subjects, the term $\tilde{W}_i(T_{ij})$ reflects independent stationary processes given a special form of serial correlation, T_{ij} is the value of time for subject i at time point j, and ε_{ij} represents the subsampling uncertainty within subject. For analytic convenience, each of the three terms, b_i, $\tilde{W}_i(T_{ij})$, and ε_{ij}, are usually assumed to follow a normal distribution with a well-defined variance or a covariance function. With the specification of b_i, $\tilde{W}_i(T_{ij})$, the random component ε_{ij} is conditionally independent of random errors at other time points. Thus, the estimate of $\boldsymbol{\beta}$ tends to be unbiased, and the term $\boldsymbol{X}'_{ij}\boldsymbol{\beta}$ represents the mean response for subject i at time point j.

The three sources of variability in longitudinal data are usually intertwined, as indicated earlier. Therefore, in much of the empirical research, only one source of systematic variability needs to be considered. The most popular approach in longitudinal data analysis is perhaps the mixed-effects modeling in which the between-subjects random effects are specified to address intraindividual correlation. Let n_i be the number of repeated measurements for subject i in a sample of N individuals, and $\boldsymbol{Y}_i$ be the

n_i-dimensional vector of repeated measurements for the subject. If only a single term of the random effects is considered, a typical linear mixed model is given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i'\boldsymbol{\beta} + b_i + \boldsymbol{e}_i, \tag{1.3}$$

where $\boldsymbol{X}_i$ is a known $n_i \times M$ matrix of covariates with the first column taking constant 1, $\boldsymbol{\beta}$ is an $M \times 1$ vector of unknown population parameters, and b_i is the unknown subject's effect. In this specification, the primary predictor *time* is usually specified as a single continuous variable or a set of polynomials that are included in $\boldsymbol{X}$. As the specification of a single random effect does not confound the fixed regression coefficients, b_i is also referred to as the random effect for the intercept. Given the inclusion of b_i, linear mixed models are thought to yield more efficient and robust regression coefficients than do general linear models in which residuals are potentially dependent. Sometimes, multiple terms of the random effects (e.g., the random effects for the intercept and for the time factor) need to be specified, and then the random effects for subject i are expressed in terms of $\boldsymbol{b}_i$ in mixed-effects modeling given the specification of a design matrix. These extended mixed-effects models will be described and discussed in many of the succeeding chapters.

Given the specification of b_i or $\boldsymbol{b}_i$, the elements in $\boldsymbol{e}_i$ are assumed to be conditionally independent. Specifically, the term $\boldsymbol{e}_i$ is an $n_i \times 1$ column vector of random errors for subject i, given by

$$\boldsymbol{e}_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix}.$$

Another popular perspective in longitudinal modeling is to specify the covariance structure of within-subject random errors in linear regression models while leaving the between-subject random effects unspecified. The use of such a design in modeling a longitudinal process becomes necessary when the application of the random-effects approach does not yield reliable analytic results or when the within-subject variability is sizable in comparison with the between-subjects variability. In this approach, the vector $\boldsymbol{e}_i$ is assumed to follow a multivariate normal distribution with mean 0 and an $n_i \times n_i$ covariance matrix. To pattern the error covariance structure in the linear regression, the time factor must be specified as a classification factor containing n discrete levels for reflecting the repeated effects. In empirical analyses, there are a variety of covariance pattern models that have been designed over the years by statisticians. Although it bears tremendous resemblance to the classical repeated-measures ANOVA-type models, this linear regression model on repeated measurements is generally viewed as a family of linear mixed models in view of its capacity to account for between-subjects variability.

The aforementioned two perspectives on the continuous response variable can be readily extended to statistical modeling on a variety of nonnormal outcome variables, such as rate and proportion, multinomial, and count data. There are many

statistically complex models and methods for those longitudinal data types; the fundamental rationale in these advanced models, however, is based on the classification of variability in longitudinal data. In the succeeding chapters, detailed specifications and statistical inferences for various longitudinal models, linear or nonlinear, will gradually unfold with empirical illustrations presented for their applications.

1.8 ORGANIZATION OF THE BOOK AND DATA USED FOR ILLUSTRATIONS

The remainder of the book is organized as follows. Chapter 2 is devoted to detailing some descriptive techniques that were applied in the early stage of longitudinal data analysis. These classical methods include the empirical time plots of trends, the paired two-time *t*-test, the effect size and its confidence interval, the repeated-measures ANOVA, and the repeated-measures multivariate ANOVA, also referred to as MANOVA. The limitations of the descriptive methods in longitudinal data analysis are discussed. Chapter 3 describes the general specifications, basic inferences, and the estimating procedures of linear mixed models, with a focus on the fixed effects in the presence of the specified random effects. In Chapter 4, the restricted maximum likelihood estimator is introduced, and two computational procedures to estimate parameters in linear mixed models, the Newton–Raphson and the expectation–maximization algorithms, are described. Also delineated in this chapter are the statistical techniques for approximating the subject-specific random effects. Chapter 5 introduces a variety of residual variance–covariance pattern models for both equally and irregularly spaced longitudinal data. Chapter 6 presents the statistical methods on residual and influence diagnostics for checking the adequacy of linear mixed models. In Chapter 7, several special topics in linear mixed models are described and discussed, such as the adjustment of baseline response, misspecification of the normality hypothesis, and the application of pattern-mixture modeling in longitudinal data analysis.

Chapters 8–12 are devoted to statistical models and methods for modeling nonnormal longitudinal data. In particular, Chapter 8 introduces the general inference of generalized linear mixed models given the specification of the random effects. A variety of statistical methods are delineated for the estimation of fixed and random effects in generalized linear mixed models, followed with the description of various approaches to perform nonlinear predictions of nonnormal longitudinal data. Chapter 9 is focused on the delineation of generalized estimating equations modeling the marginal mean parameters by using a quasi-likelihood method. Chapter 10 concerns the general specifications of the mixed-effects logistic regression model, including the statistical inference of the regression, a brief discussion on the interpretability of the conventional odds ratio in longitudinal data analysis, and the approximation of longitudinal trajectories of the response probability and the corresponding standard errors. Chapter 11 presents model specifications and statistical inference of the mixed-effects multinomial logit model. In this chapter, a retransformation method is developed to derive unbiased nonlinear predictions of a set of response probabilities.

Based on the development of the mixed-effects multinomial logit model, Chapter 12 specifies a number of multidimensional transition models, which link categorical outcome data to the time factor, the value of one or more prior states, and other theoretically relevant covariates. In each of these chapters, a step-by-step presentation with empirical illustrations is provided, and the merits and limitations in these statistical models are discussed.

In Chapter 13, techniques of latent growth modeling are described and discussed. The latent growth models covered in this chapter consist of the latent growth model, the latent growth mixture model, and the group-based model, with each method following the SEM perspective. Substantial discussions are provided on the advantages and the existing problems in these SEM-type models with latent factors or latent classes. The last chapter, Chapter 14, is devoted to missing data analysis. Based on the mathematical definitions on MCAR, MAR, and MNAR, a variety of statistical methods are described and discussed in this chapter, dealing with ignorable (MCAR and MAR) and nonignorable missing data (MNAR), respectively. Due to consideration of coherence and conciseness of the text, some supplementary procedures, datasets, and computer programs are presented as Appendices.

As indicated in Section 1.2, longitudinal data analysis was initiated and extensively applied in clinical experimental studies, and the methods have later been extended and advanced to the realm of social and behavioral sciences. While the statistical methods applied in these two scientific domains share tremendous similarities, there are some distinct differences in the data features between them. The clinical experimental research is often based on data obtained from randomized controlled clinical trials that are generally characterized by small sample sizes and with narrowly spaced intervals between successive waves of investigation. In contrast, social and behavioral studies usually rely on large-scale survey data often with widely spaced intervals. Accordingly, in this book two longitudinal datasets are used for empirical illustrations, one from a typical randomized controlled clinical trial and one from a large-scale longitudinal survey. These two datasets represent two data types. The clinical trial dataset is of small scale, with follow-up investigation conducted at every 4 weeks, whereas the second dataset comes from a nationally representative investigation of older Americans with a large sample size at baseline and an interwave interval of every 2 years. The following paragraphs provide detailed descriptions concerning the two datasets.

1.8.1 RANDOMIZED CONTROLLED CLINICAL TRIAL ON THE EFFECTIVENESS OF ACUPUNCTURE TREATMENT ON PTSD

The randomized controlled clinical trial was conducted by the Department of Defense (DoD), Deployment Health Clinical Center (DHCC), Walter Reed National Military Medical Center (WRNMMC) between February 2006 and October 2007. The operational objective of the trial is to test the hypothesis that a brief, 4-week course of acupuncture (89-min treatments) in combination with patients' usual PTSD care is significantly more effective than usual PTSD care alone in active-duty military personnel diagnosed with PTSD. Given such a focus, the design was a two

parallel arm randomized controlled trial comparing the effectiveness of semistructured, brief adjunctive acupuncture for patients' receiving usual PTSD care to the effectiveness of usual PTSD care alone. Participants were recruited from primary care clinics at WRNMMC (68%), while self-referrals from advertisements at WRNMMC (19%) and referrals from providers and patients (13%) made up the remainder of the sample. A total of 55 subjects, 28 who were assigned to acupuncture treatment and 27 to the usual care control group, participated in the study. Once randomized to the acupuncture treatment group, the participants were randomly assigned to one of the three licensed acupuncturists. Participants were assessed at baseline and at 4, 8, and 12 weeks postrandomization, respectively. Therefore, intervals are equally spaced. After completion of study follow-ups, participants assigned to the control group were offered the study acupuncture intervention and a list of local mental health services.

1.8.2 ASSET AND HEALTH DYNAMICS AMONG THE OLDEST OLD (AHEAD)

The second dataset comes from a large-scale longitudinal study on older Americans, the Survey of AHEAD. 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 Household and Retirement Survey (HRS). As a supplemental survey attached to HRS, Wave I of the AHEAD survey was conducted between October 1993 and April 1994. Specifically, a sample of individuals aged 70 years 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 9,473 households and 11,965 individuals in the target area range. The Wave I respondents have been followed by telephone every second or third year, with proxy interviews designed for those deceased between two successive waves. At present, AHEAD survey registers 10 waves of investigation in 1993, 1995, 1998, 2000, 2002, 2004, 2006, 2008, 2010, and 2012. As a longitudinal, multidisciplinary, and US population-based study, AHEAD provides a highly representative and reliable data base for longitudinal data analysis of older Americans aged 70 years 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 National Death Index (NDI). To provide empirical illustrations, this book uses AHEAD data of six waves from 1998 to 2008, viewing the 1998 panel as baseline. Given the illustrative nature of using this dataset, I randomly select 2000 persons from the baseline AHEAD sample to conduct empirical analyses in the book. Additionally, due to the same reason for providing examples, the weight factor, included in the AHEAD dataset for adjusting oversampling of certain subpopulations, is not considered in the empirical illustrations. Therefore, readers interested in following the examples of this book for an actual analysis should use the full AHAED sample and incorporate the weight variable for deriving unbiased analytic results.