Cambridge University Press
978-1-107-03003-9 - Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide: Second Edition
Jos W. R. Twisk
Excerpt
More information

1

Introduction

1.1 Introduction

Longitudinal studies are defined as studies in which the outcome variable is repeatedly measured; i.e. the outcome variable is measured in the same subject on several occasions. In longitudinal studies the observations of one subject over time are not independent of each other, and therefore it is necessary to apply special statistical techniques, which take into account the fact that the repeated observations of each subject are correlated. The definition of longitudinal studies (used in this book) implicates that statistical techniques like survival analyses are beyond the scope of this book. Those techniques basically are not longitudinal data analysing techniques because (in general) the outcome variable is an irreversible endpoint and therefore strictly speaking is only measured at one occasion. After the occurrence of an event no more observations are carried out on that particular subject.

Why are longitudinal studies so popular these days? One of the reasons for this popularity is that there is a general belief that with longitudinal studies the problem of causality can be solved. This is, however, a typical misunderstanding and is only partly true. Table 1.1 shows the most important criteria for causality, which can be found in every epidemiological textbook (e.g. Rothman and Greenland, 1998). Only one of them is specific for a longitudinal study: the rule of temporality. There has to be a time-lag between outcome variable Y (effect) and covariate X (cause); in time the cause has to precede the effect. The question of whether or not causality exists can only be (partly) answered in specific longitudinal studies (i.e. experimental studies) and certainly not in all longitudinal studies. What then is the advantage of performing a longitudinal study? A longitudinal study is expensive, time consuming, and difficult to analyze. If there are no advantages over cross-sectional studies why bother? The main advantage of a longitudinal study compared to a cross-sectional study is that the individual development of a certain outcome variable over time can be studied. In addition to this, the individual development

Table 1.1 Criteria for causality

Strength of the relationship
Consistency in different populations and under different circumstances
Specificity (cause leads to a single effect)
Temporality (cause precedes effect in time)
Biological gradient (dose–response relationship)
Biological plausibility
Experimental evidence

of a certain outcome variable can be related to the individual development of other variables.

1.2 General approach

The general approach to explain the statistical techniques covered in this book will be "the research question as basis for analysis." Although it may seem quite obvious, it is important to realize that a statistical analysis has to be carried out in order to obtain an answer to a particular research question. The starting point of each chapter in this book will be a research question, and throughout the book many research questions will be addressed. The book is further divided into chapters regarding the characteristics of the outcome variable. Each chapter contains extensive examples, accompanied by computer output, in which special attention will be paid to interpretation of the results of the statistical analyses.

1.3 Prior knowledge

Although an attempt has been made to keep the complicated statistical techniques as understandable as possible, and although the basis of the explanations will be the underlying epidemiological research question, it will be assumed that the reader has some prior knowledge about (simple) cross-sectional statistical techniques such as linear regression analysis, logistic regression analysis, and analysis of variance.

1.4 Example

In general, the examples used throughout this book will use the same longitudinal dataset. This dataset consists of an outcome variable (Y) that is continuous and is measured six times on the same subjects. Furthermore there are four covariates, which differ in distribution (continuous or dichotomous) and in whether they are time-dependent or time-independent. X_1 is a continuous time-independent

Table 1.2 Descriptive information[a] for an outcome variable Y and covariates X_1 to X_4[b] measured at six occasions

Time-point	Y	X_1	X_2	X_3	X_4
1	4.43 (0.67)	1.98 (0.22)	3.26 (1.24)	143/4	69/78
2	4.32 (0.67)	1.98 (0.22)	3.36 (1.34)	136/11	69/78
3	4.27 (0.71)	1.98 (0.22)	3.57 (1.46)	124/23	69/78
4	4.17 (0.70)	1.98 (0.22)	3.76 (1.50)	119/28	69/78
5	4.67 (0.78)	1.98 (0.22)	4.35 (1.68)	99/48	69/78
6	5.12 (0.92)	1.98 (0.22)	4.16 (1.61)	107/40	69/78

[a] For outcome variable Y and the continuous covariates (X_1 and X_2) mean and standard deviation are given, for the dichotomous covariates (X_3 and X_4) the numbers of subjects in the different categories are given.

[b] Y is serum cholesterol in mmol/l; X_1 is maximal oxygen uptake (in (dl/min)/kg$^{2/3}$)); X_2 is the sum of four skinfolds (in cm); X_3 is smoking (non-smokers vs. smokers); X_4 is gender (males vs. females).

covariate, X_2 is a continuous time-dependent covariate, X_3 is a dichotomous time-dependent covariate, and X_4 is a dichotomous time-independent covariate. All time-dependent independent variables are measured at the same six occasions as the outcome variable Y.

In the chapter dealing with dichotomous outcome variables (i.e. Chapter 7), the continuous outcome variable Y is dichotomized (i.e. the highest tertile versus the other two tertiles) and in the chapter dealing with categorical outcome variables (i.e. Chapter 8), the continuous outcome variable Y is divided into three equal groups (i.e. tertiles).

The dataset used in the examples is taken from the Amsterdam Growth and Health Longitudinal Study, an observational longitudinal study investigating the longitudinal relationship between lifestyle and health in adolescence and young adulthood (Kemper, 1995). The abstract notation of the different variables (Y, X_1 to X_4) is used since it is basically unimportant what these variables actually are. The continuous outcome variable Y could be anything, a certain psychosocial variable (e.g. a score on a depression questionnaire, an indicator of quality of life, etc.) or a biological parameter (e.g. blood pressure, albumin concentration in blood, etc.). In this particular dataset the outcome variable Y was total serum cholesterol expressed in mmol/l, X_1 was fitness level at baseline (measured as maximal oxygen uptake on a treadmill), X_2 was body fatness (estimated by the sum of the thickness of four skinfolds), X_3 was smoking behavior (dichotomized as smoking versus non-smoking), and X_4 was gender. Table 1.2 shows descriptive information for the variables used in the example.

All the example datasets used throughout the book are available from the following website: http://www.jostwisk.nl.

1.5 Software

The relatively simple analyses of the example dataset were performed with SPSS (version 18; SPSS, 1997, 1998). For sophisticated longitudinal data analysis, other software packages were used. Generalized estimating equations (GEE) analysis and mixed model analysis were performed with Stata (version 11; Stata, 2001). Stata is chosen as the main software package for sophisticated longitudinal analysis, because of the simplicity of its output. In Chapter 12, an overview (and comparison) will be given of other software packages such as SAS (version 8; Littel et al., 1991, 1996), R (version 2.13), and MLwiN (version 2.25; Goldstein et al., 1998; Rasbash et al., 1999). In all these packages algorithms to perform sophisticated longitudinal data analysis are implemented in the main software. Both syntax and output will accompany the overview of the different packages. For detailed information about the different software packages, reference is made to the software manuals.

1.6 Data structure

It is important to realize that different statistical software packages need different data structures in order to perform longitudinal analyses. In this respect a distinction must be made between a "long" data structure and a "broad" data structure. In the "long" data structure each subject has as many data records as there are measurements over time, while in a "broad" data structure each subject has one data record, irrespective of the number of measurements over time (Figure 1.1).

1.7 Statistical notation

The statistical notation will be very simple and straightforward. Difficult matrix notation will be avoided as much as possible. Throughout the book the number of subjects will be denoted as $i = 1$ to N, the number of times a particular subject is measured will be denoted as $t = 1$ to T, and the number of covariates will be noted as $j = 1$ to J. Furthermore, the outcome variable will be called Y, and the covariates will be called X. All other notations will be explained below the equations where they are used.

“long” data structure

ID	Y	time	X_4
1	3.5	1	1
1	3.7	2	1
1	3.9	3	1
1	3.0	4	1
1	3.2	5	1
1	3.2	6	1
2	4.1	1	1
2	4.1	2	1
.			
.			
N	5.0	5	2
N	4.7	6	2

“broad” data structure

ID	Y_{t1}	Y_{t2}	Y_{t3}	Y_{t4}	Y_{t5}	Y_{t6}	X_4
1	3.5	3.7	3.9	3.0	3.2	3.2	1
2	4.1	4.1	4.2	4.6	3.9	3.9	1
3	3.8	3.5	3.5	3.4	2.9	2.9	2
4	3.8	3.9	3.8	3.8	3.7	3.7	1
.							
.							
N	4.0	4.6	4.7	4.3	4.7	5.0	2

Figure 1.1 Illustration of two different data structures.

1.8 What's new in the second edition?

Throughout the book changes are made to make some of the explanations clearer, and several chapters are totally rewritten. This holds for Chapter 9 (Analysis of experimental studies) and Chapter 10 (Missing data in longitudinal studies). Furthermore, two new chapters are added to the book: in Chapter 5, the role of the time variable in longitudinal data analysis will be discussed, while in Chapter 13 some new features of longitudinal data analysis will be briefly introduced.

2

Study design

2.1 Introduction

Epidemiological studies can be roughly divided into observational and experimental studies (Figure 2.1). Observational studies can be further divided into case-control studies and cohort studies. Case-control studies are never longitudinal, in the way that longitudinal studies were defined in Chapter 1. The outcome variable Y (a dichotomous outcome variable distinguishing "case" from "control") is measured only once. Furthermore, case-control studies are always retrospective in design. The outcome variable Y is observed at a certain time-point, and the covariates are measured retrospectively.

In general, observational cohort studies can be divided into prospective, retrospective, and cross-sectional cohort studies. A prospective cohort study is the only cohort study that can be characterized as a longitudinal study. Prospective cohort studies are usually designed to analyze the longitudinal development of a certain characteristic over time. It is argued that this longitudinal development concerns growth processes. However, in studies investigating the elderly, the process of deterioration is the focus of the study, whereas in other developmental processes growth and deterioration can alternately follow each other. Moreover, in many epidemiological studies one is interested not only in the actual growth or deterioration over time, but also in the longitudinal relationship between several characteristics over time. Another important aspect of epidemiological observational prospective studies is that sometimes one is not really interested in growth or deterioration, but rather in the "stability" of a certain characteristic over time. In epidemiology this phenomenon is known as tracking (Twisk et al., 1994, 1997, 1998a, 1998b, 2000).

Experimental studies, which in epidemiology are often referred to as (clinical) trials, are by definition prospective, i.e. longitudinal. The outcome variable Y is measured at least twice (the classical "pre-test," "post-test" design), and other intermediate measures are usually also added to the research design (e.g. to evaluate short-term and long-term effects). The aim of an experimental (longitudinal)

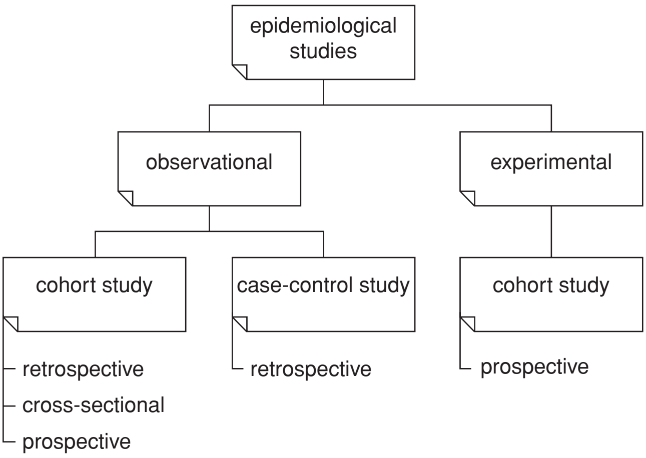

Figure 2.1 Schematic illustration of different epidemiological study designs.

study is to analyze the effect of one or more interventions on a certain outcome variable *Y*.

In Chapter 1, it was mentioned that some misunderstanding exists with regard to causality in longitudinal studies. However, an experimental study is basically the only epidemiological study design in which the issue of causality can be covered. With observational longitudinal studies, on the other hand, the question of probable causality remains unanswered.

Most of the statistical techniques in the examples covered in this book will be illustrated with data from an observational longitudinal study. In a separate chapter (Chapter 9), examples from experimental longitudinal studies will be discussed extensively. Although the distinction between experimental and observational longitudinal studies is obvious, in most situations the statistical techniques discussed for observational longitudinal studies are also suitable for experimental longitudinal studies.

2.2 Observational longitudinal studies

In observational longitudinal studies investigating individual development, each measurement taken on a subject at a particular time-point is influenced by three factors: (1) age (time from date of birth to date of measurement); (2) period (time or moment at which the measurement is taken); and (3) birth cohort (group of subjects born in the same year). When studying individual development, one is mainly interested in the age effect. One of the problems of most of the designs used in studies of development is that the main age effect cannot be distinguished from the two other "confounding" effects (i.e. period and cohort effects).

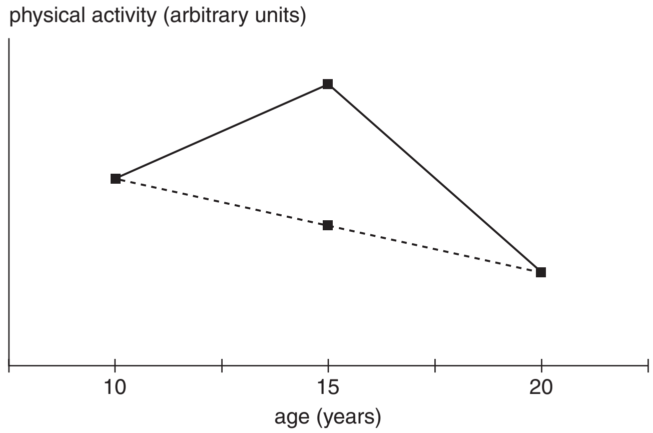

Figure 2.2 Illustration of a possible time of measurement effect (– – – "real" age trend, —— observed age trend).

2.2.1 Period and cohort effects

There is an extensive amount of literature describing age, period and cohort effects (e.g. Lebowitz, 1996; Robertson et al., 1999; Holford et al., 2005). However, most of the literature deals with classical age–period–cohort models, which are used to describe and analyze trends in (disease-specific) morbidity and mortality (e.g. Kupper et al., 1985; Mayer and Huinink, 1990; Holford, 1992; McNally et al., 1997; Robertson and Boyle, 1998; Rosenberg and Anderson, 2010). In this book, the main interests are the individual development over time, and the longitudinal relationship between different variables. In this respect, period effects or time of measurement effects are often related to a change in measurement method over time, or to specific environmental conditions at a particular time of measurement. A hypothetical example is given in Figure 2.2. This figure shows the longitudinal development of physical activity with age. Physical activity patterns were measured with a five-year interval, and were measured during the summer in order to minimize seasonal influences. The first measurement was taken during a summer with normal weather conditions. During the summer when the second measurement was taken, the weather conditions were extremely good, resulting in activity levels that were very high. At the time of the third measurement the weather conditions were comparable to the weather conditions at the first measurement, and therefore the physical activity levels were much lower than those recorded at the second measurement. When all the results are presented in a graph, it is obvious that the observed age trend is highly biased by the "period" effect at the second measurement.

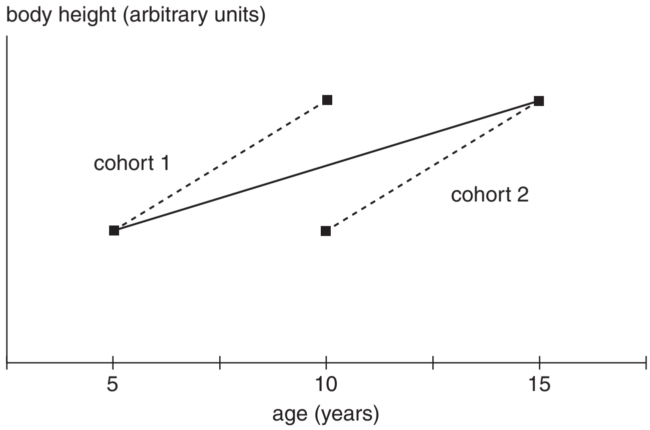

Figure 2.3 Illustration of a possible cohort effect (– – – cohort specific, —— observed).

One of the most striking examples of a cohort effect is the development of body height with age. There is an increase in body height with age, but this increase is highly influenced by the increase in height of the birth cohort. This phenomenon is illustrated in Figure 2.3. In this hypothetical study, two repeated measurements were carried out in two different cohorts. The purpose of the study was to detect the age trend in body height. The first cohort had an initial age of 5 years; the second cohort had an initial age of 10 years. At the age of 5, only the first cohort was measured, at the age of 10, both cohorts were measured, and at the age of 15 only the second cohort was measured. The body height obtained at the age of 10 is the average value of the two cohorts. Combining all measurements in order to detect an age trend will lead to a much flatter age trend than the age trends observed in both cohorts separately.

Both cohort and period effects can have a dramatic influence on interpretation of the results of longitudinal studies. An additional problem is that it is very difficult to disentangle the two types of effects. They can easily occur together. Logical considerations regarding the type of variable of interest can give some insight into the plausibility of either a cohort or a period effect. When there are (confounding) cohort or period effects in a longitudinal study, one should be very careful with the interpretation of age-related results.

It is sometimes argued that the design that is most suitable for studying individual growth/deterioration processes is a so-called "multiple longitudinal design." In such a design the repeated measurements are taken in more than one cohort with overlapping ages (Figure 2.4). With a "multiple longitudinal design" the main age effect can be distinguished from cohort and period effects. Because subjects of the same age are measured at different time-points, the difference in outcome variable

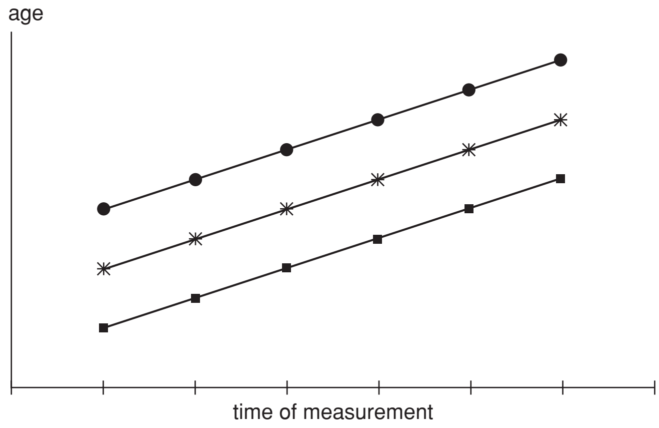

Figure 2.4 Principle of a multiple longitudinal design; repeated measurements of different cohorts with overlapping ages (■ cohort 1, ∗ cohort 2, • cohort 3).

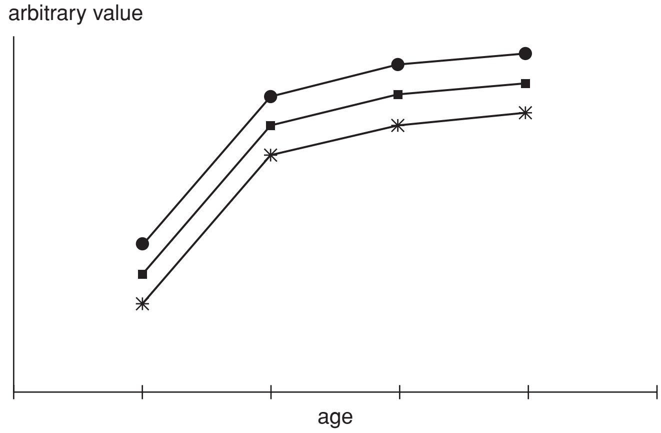

Figure 2.5 Possibility of detecting cohort effects in a "multiple longitudinal design" (∗ cohort 1, ■ cohort 2, • cohort 3).

Y between subjects of the same age, but measured at different time-points, can be investigated in order to detect cohort effects. Figure 2.5 illustrates this possibility: different cohorts have different values at the same age.

Because the different cohorts are measured at the same time-points, it is also possible to detect possible time of measurement effects in a "multiple longitudinal design." Figure 2.6 illustrates this phenomenon. All three cohorts show an increase in the outcome variable at the second measurement, which indicates a possible time of measurement effect.