Contents

Preface

Clinical Statistics: Introducing Clinical Trials, Survival Analysis, and Longitudinal Data Analysis is written for students in introductory clinical statistics or biostatistics courses. This text is an excellent reference for upper-level undergraduate or graduate degree students who have completed courses in calculus-based introduction to probability theory and mathematical statistics, and who are also familiar with the basics of regression analysis. In particular, students should be comfortable with the concepts of normal, Poisson, and gamma distributions; the Central Limit Theorem; type I and II errors; the maximum likelihood estimation; the likelihood ratio test; the acceptance region; z-tests; and the chi-squared test for independence. In addition, a working knowledge of the multivariate linear and logistic regression models is required.

This book details the underpinnings of clinical trials from the perspective of a clinical statistician. It provides a step-by-step explanation of the role of the statistician, from protocol writing to data monitoring, group randomization, and ultimately writing a final report to the U.S. Food and Drug Administration or its European equivalent. All of the necessary fundamentals of statistical analysis—i.e., the survival and longitudinal data analyses—are included.

The topics covered provide a solid mathematical background for the student and are supplemented by illustrative examples with applications in SAS statistical software. Students learn not only the relevant SAS procedures, but also the theory behind those procedures. This knowledge will allow students to pursue a career as a clinical statistician at biotechnology, pharmaceutical, and biomedical companies.

Clinical Statistics is organized into four chapters. Chapter 1 introduces fundamental concepts of clinical trials and explains in detail the role of trial participants. The statistical aspects of clinical trials are presented in Chapter 2, while Chapter 3 contains the essentials of survival analysis. Chapter 4 provides simple approaches to analyzing longitudinal data from clinical trials. Although it is necessary to read Chapter 1 first, the remaining chapters can be read in any order. The complete solutions manual to the exercises in the text can be found at http://www.jbpub.com/catalog/9780763758509.

I wish to thank the anonymous reviewer for providing valuable comments and suggestions. I would also like to thank the editorial and production teams at Jones and Bartlett Publishers for their hard work, especially Tim Anderson, Katherine Macdonald, Melissa Elmore, and Melissa Potter.

Olga Korosteleva

Chapter 1

Conducting Clinical Trials

This chapter introduces fundamental terminology of clinical trials and explains the duties of their key participants.

1.1 Basic Concepts

Clinical trials are investigations of risk and benefit properties of new therapies proposed for use in humans. For example, a pharmaceutical or a biotechnology company may conduct a clinical trial to determine the effectiveness of a new drug or an innovative biological device.

Physicians at primary care facilities may recommend individuals as enrollees in an appropriate clinical trial. Before joining a trial, qualified candidates go through the *informed consent* process, which is designed to inform them of their rights and of the risks and benefits of the investigated therapy. After achieving an understanding of the facts, the candidate signs an *informed consent form*, a document confirming the patient's consent to take part in the trial. Patients participating in a clinical trial are termed *subjects*.

At enrollment, all subjects receive an *initial treatment*. This treatment may consist of a surgical procedure to implant a bio-device or a doctor's visit to get an initial supply of certain pills, for example. After that, subjects are expected to come to scheduled *follow-up visits*, during which their health condition is checked and necessary measurements are taken and recorded. For a particular clinical trial, the times between the follow-up visits are predetermined and are the same for all subjects. For example, after an initial surgery, the subjects are expected to appear for 1-, 3-, 6-, and 12-month follow-up visits.

After the last scheduled follow-up visit, each subject has a choice of continuing in the study or dropping out. A subject who wishes to remain in the study fills out an *addendum to informed consent form*, a document supporting the subject's willingness to continue in the trial.

Depending on the length of their participation in the trial, subjects maybe be divided into four categories:

- Subjects for whom *adverse events* occur, where an adverse event is defined as contracting a certain disease, developing a certain health condition, or dying, depending on the specialization of the trial.
- Subjects who drop out of the clinical trial prior to the last follow-up visit. They are termed *drop-outs* (or *lost to follow-up subjects*). For example, a subject may move out of a state and can no longer be reached by investigators to schedule a follow-up visit.
- Subjects who voluntarily discontinue participation in the trial after the final follow-up visit.
- Subjects who are still enrolled in the trial at the moment of its completion.

Typically, a clinical trial is stopped when a predetermined number of subjects have been accrued. For example, a study may terminate after 250 subjects have been followed for at least 1 year. Section 2.1 covers this topic in more detail.

Occasionally, a trial is terminated earlier—for example, when collected data strongly support the efficacy of the tested therapy or, conversely, show that the therapy is harmful. More details are presented in Section 2.2.

Clinical trials have certain essential characteristics:

- They are *prospective*—that is, the participants are followed from well-defined points in time called *time zero* or *baseline*. Baselines will not be the same for all subjects because subjects enter the study at different times.
- Ordinarily, clinical trials are either randomized or nonrandomized. A *randomized* trial is used to compare efficacy of two or more tested therapies. Such a trial includes several *treatment groups*, with one treatment assigned to each group. To validate a statistical analysis when comparing results, it should be equally likely for a subject to be assigned to any group. A detailed description of group randomization is given in Section 2.3.
- A special case of a randomized trial is a *randomized controlled* trial that contains a *control group* and a *treatment group*. The treatment group gets the innovative intervention, whereas the control group receives either a *standard* treatment (the treatment that is currently on the market) or a placebo. A *placebo* is a treatment that is administered in the form of

a medication (a pill, a liquid, or a powder) but has no active medicinal ingredients. For example, when testing a new drug that supposedly prevents colds, a placebo might be a sugar pill. However, sick patients in the control group (for example, cancer patients in a clinical trial of a new drug that is expected to cure the cancer) do not receive a placebo if a known beneficial drug is available. There is no such thing as a placebo for the control group in the trials of new biological devices, either. Instead, devices that are widely in use are implanted. For this type of trial, a nonrandomized clinical trial may be a better option.

- In a *nonrandomized* trial, all subjects receive the experimental treatment and, therefore, there is no randomization across groups. In the statistical analysis, the tested group is compared to the *historical control* group—that is, the subjects treated in the past with an available treatment.

- Randomized clinical trials may be *double-blinded*, meaning that neither a subject nor an investigator knows in which group the subject is placed. This restriction is intended to eliminate the possibility that the investigator might be inadvertently biased when assessing the subject's health condition during follow-up visits, if it is known which treatment the subject receives. Likewise, the subject might be biased in self-assessing health condition.

- Clinical trials may involve one clinical center (a *single-center* trial) or multiple centers (a *multicenter* trial). In a multicenter clinical trial, there is a coordinating center and multiple participating sites. The *coordinating center* functions as a research center where the trial is designed, statistical analysis of collected data is performed, and findings are interpreted. A participating *investigative site* is a medical center where subjects are recruited, initially treated, and admitted for follow-up visits. Multicenter trials are more challenging to coordinate, but they recruit subjects faster and the results can be generalized to a larger population because of the broader pool of subjects.

1.2 Phases of Clinical Trials

After testing under laboratory conditions and in animals, a new product (a new drug or treatment) is tested on humans. Trials involving humans—the so-called clinical trials—are divided into four phases:

- In *Phase I* (or *pilot phase*) trials, the new product is tested on a small group of subjects (normally, 30 or fewer people). Usually the tests are done with a group of healthy volunteers or, depending on the specification of the trial, with volunteers in an advanced stage of a disease. For example, a new drug that supposedly prevents the flu would be tested in healthy subjects, whereas a new cancer treatment would be administered to newly diagnosed patients with advanced cancer.

This phase is not blinded. The volunteer participants are aware of the nature and purpose of the product they test. During this pilot phase, the safety of the product is evaluated, an optimal dosage is determined, and dangerous side effects are identified.

- In *Phase II* trials, the initial clinical investigation begins. The product is tested in a larger group of subjects (100–300 people) to determine whether it is effective and to evaluate the rate of adverse events. This phase, like Phase I, is open to volunteers and is not blinded.

- In *Phase III* trials, an extensive scientific clinical investigation of the product takes place. The test is carried out on a very large group of subjects (500–3,000 people) to compare the new product with a placebo or a standard treatment, and to confirm its efficacy and monitor side effects. After this phase is completed, results are reported to the U.S. Food and Drug Administration (FDA) or, if trials are conducted in Europe, to the European Commission (EC) and the European Medicines Agency (EMEA).

- After the product is approved for marketing, clinical trials enter *Phase IV* —the last phase in which post-marketing surveillance takes place. In Phase IV, a test is carried out in the general population after the product is marketed to collect additional information on the product's safety and efficacy over an extended period of time.

This book focuses on Phase III clinical trials, which entail a full-scale evaluation that requires the expertise of a clinical statistician.

1.3 Clinical Trials Management

1.3.1 Key Participants

A *sponsor* of a clinical trial is an individual, organization, or company that initiates and finances the clinical trial. The sponsor is involved in selecting qualified investigators and participating sites, ensuring compliance with all regulations, and monitoring data for safety and efficacy of the tested product.

The *clinical research associate* is an on-site monitor on behalf of the sponsor. This person oversees the trial and ensures that the investigative site meets all regulatory requirements, the personnel are qualified and properly trained, and the constant supply of materials required to conduct the trial is available.

Several external committees, consisting of physicians, statisticians, researchers, and other professionals, monitor the progress and safety of a clinical trial:

- The *institutional review board* (*IRB*) [in Europe, the *independent ethics committee* (*IEC*)] is designated to protect the rights and ensure the safety and well-being of human subjects enrolled in the clinical trial.
- *Data safety monitoring boards* (*DSMBs*) [in Europe, *data monitoring committees* (*DMCs*)] are set up specifically to monitor data continuously to determine that subjects are not exposed to undue risks—for example, highly toxic therapies or highly dangerous medical procedures. This committee is entitled to recommend termination of the trial if there is a safety concern, such as if a very high mortality rate is observed.
- The *U.S. Food and Drug Administration* (*FDA*) [for European sites, *European Commission* (*EC*)/*European Medicines Agency* (*EMEA*)] ensures that the reported data are accurate and that the subjects' rights and confidentiality are protected. It inspects investigation sites, the sponsor, and the IRB, and gives the final approval for marketing of the new product.

  The IRB and DSMBs are usually established at the level of research universities and large hospitals, whereas FDA, EC, and EMEA are governmental bodies.

The key *research personnel* in clinical trials include the principal investigator, the clinical research coordinator(s), the clinical statistician, and the data manager. These people work as a team, staying in constant communication and collaboration with one another. Their qualifications and professional duties are as follows:

- As a rule, the *principal investigator* (PI) is a physician qualified by training and experience. The PI is directly involved in recruitment, evaluation, and treatment of subjects at all investigative sites. He or she supervises the clinical procedures and reviews all clinical and laboratory data. The PI is also responsible for assessing causality of all adverse events and making a decision to close a center if too many adverse events occur. Although this investigator is supposed to be present at all sites at the same time, in practice the PI officially transfers some of the duties to trusted qualified on-site staff.
- The *clinical research coordinator* is the person at the participating clinical center who is responsible for on-site day-to-day operations of the trial. Some of the coordinator's responsibilities include enrolling subjects, scheduling follow-up visits, completing the required data collection forms, and submitting data entries to the coordinating center.

- The *clinical statistician* determines the statistical methodology for the trial; evaluates the trial length (see Section 2.1); for randomized trials, randomizes group allocations for subjects (see Section 2.3); monitors the data (see Section 2.2); analyzes the data; and provides the interim data reports for DSMBs and the final FDA (EC/EMEA) report (see Section 2.4).

- The *data manager* trains the clinical research coordinator at each clinical center on how to fill out and submit data forms. The data manager maintains the database and helps the statistician in data monitoring and analysis. If contradictory data entries or identical entries from a site are submitted to the database, it is the data manager's responsibility to generate a *query*, a request for a correction to the clinical research coordinator at the site. Both parties then make sure the query is resolved.

1.3.2 Employee Documentation

A company conducting clinical trials protects its confidential information about the tested product through a *nondisclosure agreement* (*NDA*) with its employees—that is, a legal document outlining confidential materials that should be restricted from public use at least for the duration of the trials. This document is known in Europe as a *confidentiality agreement* or *confidential disclosure agreement* (*CDA*).

Employees also sign a *statement of economic interest*, a legal form on which a person discloses the amount of his or her assets in the company: ownership of stock, gifts received from the company, outstanding loans from the company, and so on. Usually, this form is filed with the IRB, which is charged with the responsibility of monitoring possible *conflict of interest* cases where the individual's personal financial interest in the success of the trial conflicts with the company's interest in a fair and objective trial.

1.4 Preparation of Protocols

A *clinical trial protocol* is a document that describes every aspect of the proposed trial. It is written by a team of prospective investigators, including a statistician. The protocol is finalized and approved by the IRB prior to the beginning of the trial.

A typical protocol consists of the following parts:

- The title page, containing the title of the trial, the name and complete address of the PI, and the date

- Review of the literature that is related to the clinical problem and justification of the need for the trial

- *Preclinical data analysis*—that is, the results from Phase I and II clinical trials
- Research questions and statistical hypotheses
- Study design: randomized or nonrandomized trial, double-blinded, controlled
- Subjects enrollment procedure: recruitment, screening, and selection (the *inclusion–exclusion criteria*, which are the standards used to determine whether a person may or may not be allowed to participate in a clinical trial)
- Materials and methodology: description of the product, treatment regimen, product preparation, receiving, storage, dispensing, and return
- Data forms: baseline and follow-up data collection forms
- Database management: data collection and clean-up
- Statistical plan: trial length determination (see Sections 2.1 and 2.2), randomization procedure (see Section 2.3), and statistical methods (see Section 2.4)
- Subject safety monitoring plan: reporting of serious adverse events, maintenance of subject privacy and confidentiality.

Once the IRB approves the protocol, it is distributed to all key participants at all investigative sites. The protocol is to be followed precisely. It allows the study to be performed in exactly the same way at all the locations. Investigators use the protocol as a reference for every step of the trial. If a deviation occurs, it is reported to the IRB for a review.

Chapter 2

Implementation of Statistics in Clinical Trials

This chapter explains at which steps the expertise of a clinical statistician is required and how that expertise is implemented.

2.1 Determination of Trial Length

The *sample size* of a clinical trial is the total number of subjects involved in the trial. The sample size should be large enough to deliver statistically meaningful information about the tested product. For most trials, the length of the study is determined entirely by the minimum sample size required to detect the efficacy of the new product. The trial is stopped after the prespecified number of subjects has been enrolled and treated at least for the period of the follow-up visits. For example, a trial may be discontinued after 200 subjects have been followed for at least 12 months.

In the other cases, instead of the sample size, it is more convenient to predetermine the minimum required *number of patient-years*, which is the cumulative time for all subjects in the trial. For example, a trial may be terminated when 800 patient-years has been accrued. The actual number of subjects who should participate in the trial depends on the frequency of patient enrollment, which might not be easily computable.

The minimum required sample size (or the number of patient-years, if applicable) should be estimated before the trial begins and documented in the protocol. This section presents a detailed explanation and examples of this procedure.

The following steps should be completed prior to computing the minimum required sample size (or the number of patient-years) of a trial:

1. The *endpoint* of the clinical trial—a measure of the target outcome—should be defined. There are typically three types of endpoints:

    - A prespecified percentage change from the baseline value in some medical measurement. For example, a trial may continue until an average of a 20% reduction in bad cholesterol level is observed.
    - A prespecified actual change from the baseline value in some medical characteristic. For example, a trial may be terminated when an average of a 1.8-point increase in red blood cell count for anemia patients is observed.
    - A prespecified rate of a certain adverse event (called an *event rate* or *complication rate*), defined as the ratio between the total number of events and the total time in the trial for all subjects. For example, a trial may be discontinued when a 3.2% death rate is observed—that is, when, say, 8 deaths in 250 accrued months are recorded.

    Usually researchers are discouraged from using an actual change endpoint, because subjects have different measurements at the baseline. A percentage change endpoint is used instead.

    Some trials may have multiple endpoints, in which case one of them should be chosen as the *primary endpoint* for computation of the sample size. This endpoint would be the one that in the researchers' estimation requires the largest sample size.

2. A certain family of probability distributions for the endpoint (such as normal, Poisson, or others) should be specified, and its parameters should be estimated based on preclinical or historical data.

    Generally, the mean percentage changes or actual changes are modeled as normally distributed random variables by the Central Limit Theorem (see Example 2.1 later in this chapter). In the case of the event rate endpoint, event occurrences are random and may be modeled by a Poisson distribution (see Example 2.2).

3. The null and the alternative statistical hypotheses for the endpoint should be identified. The nature of the tested product dictates what these hypotheses should be; statistics has no say in this matter. Three situations are typically distinguished:

    - The new product cannot do any worse than the standard one or a placebo, only better. Then researchers should test a one-sided alternative hypothesis that states the superiority of the new product (see Example 2.1).

- The new product is not expected to do better than the marketed one, but has some other desired properties—for example, it is cheaper. Then researchers should test a one-sided alternative stating that the new product does not do much worse than the existing one (see Example 2.2).
- The new product might be more efficient than the standard one, but serious side effects are a possibility. Then a two-sided alternative is in order (see Exercise 2.2).

4. The probability of type I error should be determined. The *probability of type I error* (also known as the *significance level* of the test) is the maximum probability of rejecting the null hypothesis, provided that the null is true. Traditionally, this value is taken as 0.01 or 0.05 (1% or 5%).

5. The probability of type II error and the minimum detectable difference should be decided upon. The *probability of type II error* is the probability of accepting the null hypothesis, given that the null is false and a specific alternative hypothesis holds. In randomized trials, the certain value of the endpoint satisfying this alternative hypothesis is called the *minimum detectable difference* (or the *effect size*) in the endpoints for the treatment and the control groups. For nonrandomized trials with a historical control, this value is usually the value of the historical endpoint. The probability of type II error is usually set at 0.2 or 0.25 (20% or 25%). Occasionally, it is chosen to be 0.1 or 0.15 (10% or 15%).

   Conventionally, in clinical trials the *power* of a test is considered, which is defined as one minus the probability of type II error. That is, the power is the probability of rejecting the null hypothesis, given that a specific alternative is valid.

Whether the length of a trial is determined by a fixed sample size or a number of patient-years depends on the type of the endpoint. For percentage change or actual change, the sample size is to be computed. For the complication rate, the number of patient-years must be calculated.

The rest of this section consists of two concrete examples of trial length determination for the percentage change and the complication rate endpoints.

Example 2.1 A clinical trial is conducted for a new drug to test its efficacy in lowering blood pressure in patients suffering from hypertension. The control subjects receive a marketed drug. The investigators specify the endpoint as the percentage reduction in diastolic blood pressure (the pressure in the blood vessels while the heart is relaxing).

Denote by μ_{tr} and μ_c the true mean percentage reduction in blood pressure for the treatment group and the control group, respectively. The researchers agree that the following quantities are to be used for the sample size computation:

- The hypotheses of interest are $H_0 : \mu_{tr} = \mu_c$ and $H_1 : \mu_{tr} > \mu_c$. The one-sided alternative is taken because researchers are confident that the tested drug cannot do worse than the marketed one.
- The probability of type I error $\alpha = \max \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$ is set at 0.05.
- The minimum detectable difference $\delta = \mu_{tr} - \mu_c$ is considered to be 5%; that is, $\delta = 5$.
- The probability of type II error $\beta = \mathbb{P}(\text{accept } H_0 | H_1 : \mu_{tr} - \mu_c = \delta \text{ holds})$ is fixed at 0.25.
- The data obtained at the Phase II trial suggest that the underlying distribution is approximately normal with a standard deviation of $\sigma = 15$.
- The two-sample z-test is used with an equal number n of subjects in each group.

The objective in this example is to find the value of n, the required group size in the clinical trial. Denote by $\bar{x}_{tr}$ and $\bar{x}_c$ the unknown mean values of the endpoint that will be observed in the Phase III trial in the treatment group and the control group, respectively. The two groups are assumed to be independent. Under H_0, the test statistic

$$Z = \frac{\bar{x}_{tr} - \bar{x}_c}{\sigma\sqrt{2/n}} \sim \mathcal{N}(0,1)$$

The *acceptance region*—the region in which H_0 is accepted—is of the form

$$\{Z < k\} = \left\{ \frac{\bar{x}_{tr} - \bar{x}_c}{\sigma\sqrt{2/n}} < k \right\} = \{\bar{x}_{tr} - \bar{x}_c < k\sigma\sqrt{2/n}\}$$

for some positive real constant k. If a specific alternative $H_1 : \mu_{tr} - \mu_c = \delta$ holds, then

$$\bar{x}_{tr} - \bar{x}_c \sim \mathcal{N}(\delta, 2\sigma^2/n)$$

The probabilities of type I and II errors define two equations for n and k:

$$1 - \alpha = \mathbb{P}(Z < k | Z \sim \mathcal{N}(0,1)) = \Phi(k) \tag{2.1}$$

and

$$\begin{aligned}\beta &= \mathbb{P}(\bar{x}_{tr} - \bar{x}_c < k\sigma\sqrt{2/n} | \bar{x}_{tr} - \bar{x}_c \sim \mathcal{N}(\delta, 2\sigma^2/n)) \\ &= \Phi\left(k - \frac{\delta}{\sigma\sqrt{2/n}}\right) \qquad (2.2)\end{aligned}$$

where Φ denotes the cumulative distribution function of a $\mathcal{N}(0,1)$ random variable.

It can be shown (see Exercise 2.1) that from Equations 2.1 and 2.2,

$$n = 2(\sigma/\delta)^2(\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta))^2 \tag{2.3}$$

In reality, n is taken as the smallest integer exceeding this value, which results in probability of type II error being slightly smaller than the specified value. In this example, plugging into Equation 2.3 the values $\alpha = 0.05$, $\beta = 0.25$, $\sigma = 15$, and $\delta = 5$, results in a sample size of $n \geq 96.83$; that is, $n = 97$ per group is needed. The actual probability of type II error corresponding to this sample size is 0.249. □

Example 2.2 A nonrandomized clinical trial is conducted to test the performance of a new heart valve implant. Investigators would like to monitor rates of several valve-related complications. However, only one of those rates should be chosen as the primary endpoint for estimation of the trial's length. It is explained here how the choice is made.

In the statistical analysis:

- A complication rate R for the new heart valve is compared to a historical value R_h.
- The null hypothesis $H_0 : R \geq 2R_h$ is tested against the alternative $H_1 : R < 2R_h$. Note that the null hypothesis indicates that the new valve performs much worse than the historical one; if the null is accepted, the valve should not be marketed.
- The probability of type I error α is 0.05.
- The probability of type II error β, assuming $R = R_h$, is 0.2.
- The number of complications X is modeled as a Poisson random variable with mean $\lambda = RT$ over a fixed time period T.
- The historical mean of the number of complications is $\lambda_h = R_hT$. The null and the alternative hypotheses of interest can be written as $H_0 : \lambda \geq 2\lambda_h$ and $H_1 : \lambda < 2\lambda_h$. The specific value of the alternative for which β is computed is $\lambda = \lambda_h$. The value of λ_h may be computed from the equations for α and β.
- The historical rate $R_h = 0.012$ or 1.2% for *endocarditis* (inflammation of the heart lining and valves) is the smallest among all considered historical complication rates. Therefore, the corresponding number of patient-years $T = \lambda_h/R_h$ is the largest. Because the clinical trial should continue at

least that long, the rate of endocarditis should be chosen as the primary endpoint for the trial. The required number of patient-years T is computed next.

Denote by x the observed number of endocarditis complications. From hypotheses testing theory, the acceptance region is of the form $\{x > x_0\}$, where the critical value x_0 is a positive integer (see Exercise 2.3). For a fixed x_0, $\alpha = \max_{\lambda \geq 2\lambda_h} \mathbb{P}(X \leq x_0)$ corresponds to the case $\lambda = 2\,\lambda_h$ (see Exercise 2.3). Therefore, the equations for α and β are

$$1 - \alpha = \mathbb{P}(X > x_0 | \lambda = 2\lambda_h) \quad \text{and} \quad \beta = \mathbb{P}(X > x_0 \,|\, \lambda = \lambda_h)$$

where $X \sim \text{Poisson}(\lambda)$.

These equations define a system of two nonlinear equations in two unknowns, x_0 and λ_h:

$$1 - \alpha = \sum_{k=x_0+1}^{\infty} \frac{(2\lambda_h)^k}{k!} e^{-2\lambda_h} \tag{2.4}$$

$$\beta = \sum_{k=x_0+1}^{\infty} \frac{\lambda_h^k}{k!} e^{-\lambda_h} \tag{2.5}$$

This system cannot be solved exactly under the restriction that x_0 is an integer. However, an integral version of a Poisson distribution can be utilized. For any $Y \sim \text{Poisson}(\lambda_0)$, and for any positive real y, the following formula holds (see Exercise 2.4):

$$\mathbb{P}(Y > y) = \int_0^{\lambda_0} \frac{u^y}{\Gamma(y+1)} e^{-u} du \tag{2.6}$$

where $\Gamma(y+1) = \int_0^\infty v^y e^{-v} dv$ is the gamma function. Hence, Equations 2.4 and 2.5 can be written as follows:

$$1 - \alpha = \int_0^{2\lambda_h} \frac{u^{x_0}}{\Gamma(x_0+1)} e^{-u} du \tag{2.7}$$

$$\beta = \int_0^{\lambda_h} \frac{u^{x_0}}{\Gamma(x_0+1)} e^{-u} du \tag{2.8}$$

The numerical solution of these equations is $x_0 = 11.296$ and $\lambda_h = 9.287$. For $R_h = 0.012$, the required length of the trial is $T = \lambda_h / R_h = 9.287/0.012 = 774$ patient-years.

The approximate solution to Equations 2.4 and 2.5 is $x_0 = 12$ and $\lambda_h = 9.72$, with the left-hand sides equal to 0.050 and 0.183, respectively. This solution results in $T = 9.72/0.012 = 810$ patient-years.

The quantities 774 and 810 patient-years gave rise to the FDA requirement that a nonrandomized clinical trial of a new heart valve is to be continued for a minimum of 800 patient-years (*FDA Draft Replacement Heart Valve Guidance*, 1994). □

2.2 Interim Data Monitoring

Data monitoring in clinical trials may take the form of an *interim analysis*, a data analysis done while the trial is still in progress to determine whether the trial should be discontinued. A clinical trial may be terminated earlier if it can be shown that the tested product is superior to the standard one, or if the tested product is found to be risky and dangerous.

Clinical investigators should decide a priori whether to conduct a full-length study and make a decision about the product efficacy at the end or to perform interim testings. If researchers have confidence in the tested product, interim data monitoring is a reasonable procedure because it is likely to result in an early termination of the trial.

The number of interim data reports should be prespecified. In addition, the *interim sample sizes* (the number of subjects involved in interim analysis) should be statistically estimated before the trial begins and documented in the protocol.

There are two major statistical methods for calculation of interim sample sizes: classical group sequential testing and the Bayesian sequential procedure.

2.2.1 Classical Group Sequential Testing

In a randomized trial with two treatment groups (possibly, a randomized controlled trial), *classical group sequential testing* is employed in the following manner. When data for n subjects in each group are available, an interim analysis is conducted on the $2n$ subjects. The groups are statistically compared and, if the alternative hypothesis is accepted, the trial is stopped. Otherwise, the trial continues until data for another set of $2n$ subjects, n in each group, become available. Then the statistical test is conducted on the $4n$ subjects. If the alternative is accepted, the trial is discontinued. Otherwise, it continues with periodic evaluations until N sets of $2n$ subjects are available. At this point, the last statistical test is conducted, and the trial terminates.

The same procedure works in a nonrandomized trial. The interim testings are conducted on groups of size n, $2n$, Nn, or, if applicable, at equal intervals of t, $2t$, Nt patient-years (see Exercise 2.9).

The probability of type I error for the N interim statistical tests is a constant α'. For a fixed N, the values of α' and n can be found if α and β—the overall probabilities of type I and type II errors, respectively—are specified.

The *overall probability of type I error* is defined as the probability of at least one interim significant difference given that the null hypothesis is true. The *overall probability of type II error* is the probability of all interim differences being insignificant under a specific alternative hypothesis.

Example 2.3 illustrates how classical group sequential testing can be applied to monitor data in the clinical trial of Example 2.1.

Example 2.3 In Example 2.1, the hypotheses of interest are $H_0 : \mu_{tr} = \mu_c$ and $H_1 : \mu_{tr} > \mu_c$, $\alpha = 0.05$, $\beta = 0.25$, $\delta = 5$, $\sigma = 15$. To conduct this test, a sample size of 97 per group is needed. This is the case of nonsequential testing (or the group sequential test with $N = 1$).

Consider now the case $N = 2$. Let $\bar{x}_{tr}^{(i)}$ and $\bar{x}_c^{(i)}$ be the respective group sample means in the ith set of $2n$ subjects, $i = 1$ or 2. Denote by $\bar{x}_{tr} = (\bar{x}_{tr}^{(1)} + \bar{x}_{tr}^{(2)})/2$ and $\bar{x}_c = (\bar{x}_c^{(1)} + \bar{x}_c^{(2)})/2$ the respective group sample means in the combined set of $4n$ subjects.

The first statistical test of $H_0 : \mu_{tr} = \mu_c$ against $H_1 : \mu_{tr} > \mu_c$ at significance level α' is performed on the initial set of $2n$ subjects. Under H_0, $\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)} \sim \mathcal{N}(0, 2\sigma^2/n)$. The acceptance region is

$$\left\{ \frac{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}}{\sigma\sqrt{2/n}} < k \right\} = \{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)} < k\sigma\sqrt{2/n}\}$$

The relation between the significance level α' and the critical value of the acceptance region k is given by the formula $k = \Phi^{-1}(1 - \alpha')$ or, equivalently, $\alpha' = 1 - \Phi(k)$. Under a specific $H_1 : \mu_{tr} - \mu_c = \delta$, $\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)} \sim \mathcal{N}(\delta, 2\sigma^2/n)$.

If in the first test the null hypothesis is accepted, the second test of $H_0 : \mu_{tr} = \mu_c$ against $H_1 : \mu_{tr} > \mu_c$ at significance level α' is performed on the set of $4n$ subjects. The difference

$$\bar{x}_{tr} - \bar{x}_c = \frac{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}}{2} + \frac{\bar{x}_{tr}^{(2)} - \bar{x}_c^{(2)}}{2}$$

is the sum of two independent random variables that under H_0 have distribution $\mathcal{N}(0, \sigma^2/(2n))$, and under a specific $H_1 : \mu_{tr} - \mu_c = \delta$ have distribution $\mathcal{N}(\delta, \sigma^2/(2n))$. Thus, under H_0, the distribution of $\bar{x}_{tr} - \bar{x}_c$ is $\mathcal{N}(0, \sigma^2/n)$. Therefore, the acceptance region for the second test is

$$\left\{ \frac{\bar{x}_{tr} - \bar{x}_c}{\sigma\sqrt{1/n}} < k \right\} = \left\{ (\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}) + (\bar{x}_{tr}^{(2)} - \bar{x}_c^{(2)}) < 2k\sigma\sqrt{1/n} \right\}$$

Under a specific $H_1 : \mu_{tr} - \mu_c = \delta$, the distribution of $\bar{x}_{tr} - \bar{x}_c$ is $\mathcal{N}(\delta, \sigma^2/n)$.

The definitions of α and β provide two equations for k and n (see Exercise 2.5). The first equation is

$$1-\alpha = \mathbb{P}\left(\frac{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}}{\sigma\sqrt{2/n}} < k,\ \frac{\bar{x}_{tr} - \bar{x}_c}{\sigma\sqrt{1/n}} < k\right)$$

where $\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)} \sim \mathcal{N}(0, 2\sigma^2/n)$ and $\bar{x}_{tr} - \bar{x}_c \sim \mathcal{N}(0, \sigma^2/n)$

$$= \mathbb{P}(Z_1 < k,\ Z_1 + Z_2 < \sqrt{2}k) \qquad (2.9)$$

where

$$Z_1 = \frac{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}}{\sigma\sqrt{2/n}} \quad \text{and} \quad Z_2 = \frac{\bar{x}_{tr}^{(2)} - \bar{x}_c^{(2)}}{\sigma\sqrt{2/n}}$$

are independent $\mathcal{N}(0,1)$ random variables. The second equation is

$$\beta = \mathbb{P}\left(\frac{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)}}{\sigma\sqrt{2/n}} < k,\ \frac{\bar{x}_{tr} - \bar{x}_c}{\sigma\sqrt{1/n}} < k\right)$$

where $\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)} \sim \mathcal{N}(\delta, 2\sigma^2/n)$ and $\bar{x}_{tr} - \bar{x}_c \sim \mathcal{N}(\delta, \sigma^2/n)$

$$= \mathbb{P}\left(Z_3 + \frac{\delta}{\sigma\sqrt{2/n}} < k,\ Z_3 + Z_4 + 2\frac{\delta}{\sigma\sqrt{2/n}} < \sqrt{2}k\right) \qquad (2.10)$$

where $Z_3 = \frac{\bar{x}_{tr}^{(1)} - \bar{x}_c^{(1)} - \delta}{\sigma\sqrt{2/n}}$ and $Z_4 = \frac{\bar{x}_{tr}^{(2)} - \bar{x}_c^{(2)} - \delta}{\sigma\sqrt{2/n}}$ are independent $\mathcal{N}(0,1)$ random variables. To simplify notation, let $n^* = (1/2)(\delta/\sigma)^2 n$. In terms of k and n^*, Equations 2.9 and 2.10 are

$$\begin{aligned} 1-\alpha &= \mathbb{P}(Z_1 < k,\ Z_1 + Z_2 < \sqrt{2}k) \\ \beta &= \mathbb{P}(Z_1 + \sqrt{n^*} < k, Z_1 + Z_2 + 2\sqrt{n^*} < \sqrt{2}k) \end{aligned} \qquad (2.11)$$

where Z_1 and Z_2 are independent $\mathcal{N}(0,1)$ random variables.

The numeric solution of Equation 2.11 for $\alpha = 0.05$ and $\beta = 0.25$ is $k = 1.875$ ($\alpha' = 0.030$) and $n^* = 3.029$ (see Exercise 2.5). Hence, the interim group size is the smallest integer larger than $2(\sigma/\delta)^2 n^* = 54.522$; that is, $n = 55$. The probability of type II error corresponding to this group size is 0.246.

Thus, instead of accruing 97 subjects in each group and testing the hypotheses once at the 5% significance level, the group sequential method with $N = 2$ suggests that investigators test at the 3% significance level with 55 subjects in each group and, if the null is accepted, test a second time at the 3% significance level with a group size of 110 subjects. Researchers who have a very strong belief in the success of the tested product might want to go with the sequential testing plan because there is a good chance of stopping the trial after

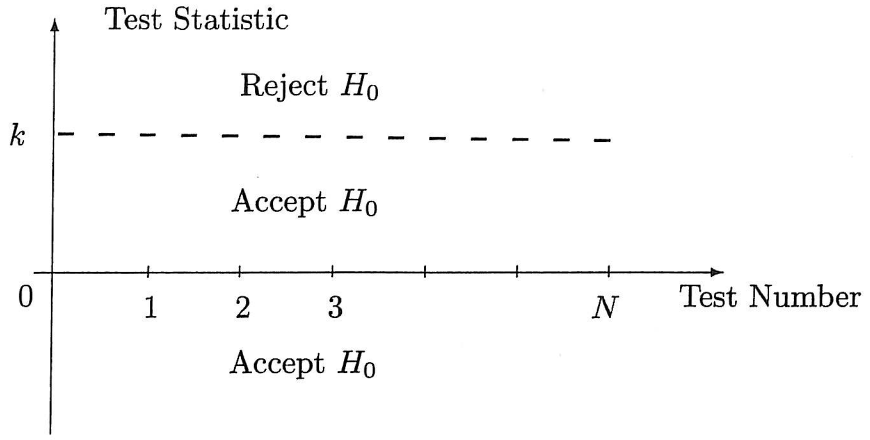

Figure 2.1 The acceptance region for the mth test, $m = 1, \ldots, N$, in the classical group sequential method in Example 2.3.

data have been collected and analyzed for only 55 subjects per group. However, if the product is not doing as well as expected, the trial must continue until 110 subjects are accrued for each group, which is longer than the trial without the interim monitoring (97 subjects per group).

For a general N, the quantities k and n^* can be expressed as follows (see Exercise 2.6):

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(\bigcap_{m=1}^{N}\left\{Z_1+\cdots+Z_m<\sqrt{m}k\right\}\right) \\ \beta &= \mathbb{P}\left(\bigcap_{m=1}^{N}\left\{Z_1+\cdots+Z_m+m\sqrt{n^*}<\sqrt{m}k\right\}\right) \end{aligned} \tag{2.12}$$

where $Z_1, \ldots, Z_N$ are independent $\mathcal{N}(0,1)$ random variables. A schematic plot of the acceptance region for the mth test, $m = 1, \ldots, N$, is given in Figure 2.1. Note that the boundary of this region is a horizontal line. This corresponds to a constant α', the probability of type I error for the interim tests. □

Other group sequential testing procedures are widely used in practice. In these methods, the boundary of the acceptance region is not horizontal, and α' is not the same for all interim tests. One example of a nonclassical group sequential testing can be found in Exercise 2.10.

2.2.2 Bayesian Sequential Procedure

In the *Bayesian sequential procedure*, the clinical endpoint is modeled as a random variable Θ. The *prior* density of Θ, $\pi(\theta) = f_\Theta(\theta)$, can be chosen in many different ways.

Researchers who have a strong belief in the efficacy of the tested product would choose an *enthusiastic* prior (also called an *optimistic* prior), which assumes that the alternative hypothesis $H_1 : \Theta \in \Omega_1$ is more likely to hold than the null hypothesis $H_0 : \Theta \in \Omega_0$, where Ω_0 and Ω_1 are some prespecified sets of possible values of Θ. An alternative choice for the prior distribution is a *skeptical* prior (also called a *pessimistic* prior). It is used by researchers who are cautious about the tested product and assume that the alternative hypothesis has a smaller probability than the null hypothesis or that the probabilities are equal.

Bayesian hypotheses testing is based on $f_\Theta(\theta \,|\, \text{data})$, the *posterior* density of Θ, given the data from trial. The posterior density is computed according to Bayes' formula

$$f_\Theta(\theta \,|\, \text{data}) = \frac{f(\text{data} \,|\, \Theta = \theta)\pi(\theta)}{\int f(\text{data} \,|\, \Theta = \theta)\pi(\theta)d\theta}$$

where $f(\text{data} \,|\, \Theta = \theta)$ denotes the *likelihood density*—that is, the density of the observations given a specific value of the endpoint. Generally, computation of the posterior density is a difficult task that might involve numerical integration. For this reason, it is convenient to choose a *conjugate prior*, defined as a prior density of a certain algebraic form chosen in such a way that the posterior density would be of the same algebraic form.

The decision of accepting or rejecting the null hypothesis is based on the following rule. If the posterior probability of the null hypothesis

$$\mathbb{P}(H_0 \,|\, \text{data}) = \int_{\Omega_0} f_\Theta(\theta \,|\, \text{data})d\theta$$

is small (usually 0.05 or less), then the null is rejected. If the posterior probability of H_0 is large (usually 0.95 or more), the null is accepted. Otherwise, the trial continues.

If a trial is not stopped earlier and reaches its predetermined sample size, the trial should be stopped, and a non-Bayesian statistical test should be performed on the data.

Example 2.4 shows how data monitoring can be performed using the Bayesian approach.

Example 2.4 In Example 2.2, the null hypothesis $H_0 : R \geq 0.024$ is tested against the alternative $H_1 : R < 0.024$. The number of events during a time period T has a Poisson distribution with mean RT. From the Bayesian perspective, R is also a random variable.

The following steps are essential in conducting the Bayesian analysis:

1. The prior density of R should be specified. A computationally convenient choice would be a conjugate prior. The distribution of the data is Poisson. It can be proven that the gamma distribution is conjugate to the Poisson

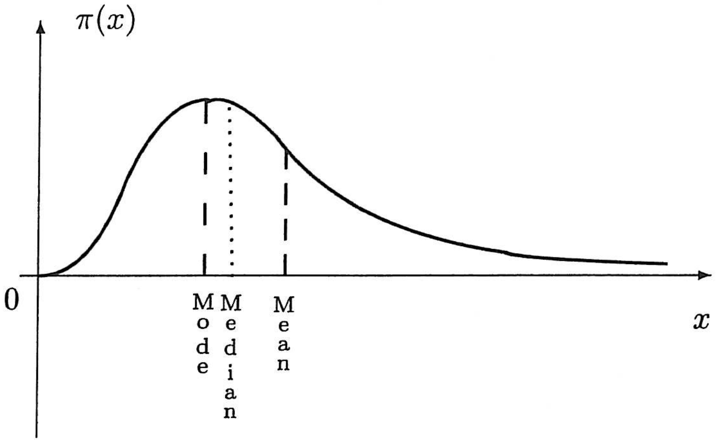

Figure 2.2 The mode, median, and mean of the gamma distribution in Example 2.4.

distribution (Show it!). Thus the prior of R may be taken as Gamma(a, b) with the density

$$\pi(x) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}, \quad x, a, b > 0$$

2. The parameters a and b of this density should be determined. The gamma distribution is unimodal and right-skewed; hence, mode < median < mean. Figure 2.2 illustrates these inequalities. Recall that the *mode* of a continuous distribution is the value that maximizes the density, and the *median* is the value that divides the area under the density curve into halves.

   Consequently,

$$\mathbb{P}(R < \text{mode}) < 0.5 < \mathbb{P}(R < \text{mean}) \tag{2.13}$$

   For a Gamma(a, b) distribution, the mode equals $(a-1)b$ (see Exercise 2.11) and the mean is ab.

   If researchers are inclined toward using an enthusiastic prior, then they should take the mean to be equal to 0.024. This gives the opportunity to specify any desired prior probability of the true H_1 larger than 0.5. Indeed, by Equation 2.13,

$$0.5 < \mathbb{P}(R < \text{mean}) = \mathbb{P}(R < 0.024) = \mathbb{P}(H_1)$$

   For a skeptical prior, the mode should be chosen equal to 0.024. Then, according to Equation 2.13,

$$\mathbb{P}(H_1) = \mathbb{P}(R < 0.024) = \mathbb{P}(R < \text{mode}) < 0.5$$

   and, therefore, the prior probability of H_1 can be fixed at any value less than 0.5. Thus the parameters a and b can be computed numerically from the equations

$$ab = 0.024 \text{ (for an enthusiastic prior)} \tag{2.14}$$

$$(a-1)b = 0.024 \text{ (for a skeptical prior)} \tag{2.15}$$

$$\mathbb{P}(H_1) = \int_0^{0.024} \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}dx \tag{2.16}$$

3. The posterior density of R should be computed. Suppose that t patient-years has been accumulated during which n endocarditis cases were observed. It is not difficult to show (see Exercise 2.12) that the posterior distribution of R is Gamma$(n+a, 1/(1/b+t))$. Under this posterior, the probability that the alternative is correct is

$$\begin{aligned}\mathbb{P}(H_1 \,|\, \text{data}) &= \mathbb{P}(R < 0.024 \,|\, n, t)\\ &= \int_0^{0.024} \frac{x^{a+n-1}(1/b+t)^{a+n}}{\Gamma(a+n)}e^{-x(1/b+t)}dx\\ &= \int_0^{0.024(1/b+t)} \frac{x^{a+n-1}}{\Gamma(a+n)}e^{-x}dx \end{aligned} \tag{2.17}$$

For certain values of n and t, this probability becomes smaller than 0.05 (then the null is accepted) or larger than 0.95 (then the alternative is accepted).

To illustrate these steps with a numerical example, assume that researchers would like to use a skeptical prior with the probability of the true alternative equal to $\mathbb{P}(H_1) = 0.4$. The posterior probability of the alternative is computed according to Equation 2.17, where a and b satisfy Equations 2.15 and 2.16.

As shown in Example 2.2, the minimum required length of the trial without an interim monitoring is 800 patient-years. Suppose that researchers decide a priori to conduct interim Bayesian analyses at $t = 400$ and $t = 600$ patient-years. Table 2.1 summarizes the stopping rules.

According to the values in Table 2.1, researchers should terminate the trial at 400 patient-years if 2 (or fewer) or 17 (or more) endocarditis events are observed. In the former case, the sample complication rate is small, and H_1 is accepted. In the latter case, the observed complication rate is high, and H_0 is accepted. If between 3 and 16 events have occured, then the trial should continue until 600 patient-years is accrued. At this point, if 3 to 6 complications

Table 2.1 Trial-Stopping Rules for 400 and 600 Patient-Years in Example 2.4

t	n	$\mathbb{P}(H_1 \mid n,t)$	t	n	$\mathbb{P}(H_1 \mid n,t)$
400	**2**	**0.9688**	400	16	0.0505
	3	0.9421		**17**	**0.0317**
600	**6**	**0.9643**	600	21	0.0668
	7	0.9399		**22**	**0.0450**

or 22 or more complications are recorded, the trial is stopped. Otherwise, it continues for the prescribed length of 800 patient-years. □

2.3 Randomization of Group Assignments

2.3.1 Principle of Similar-Sized Groups

In a randomized clinical trial, each subject is randomly (equally likely) assigned to any of the groups. The randomization procedure should adhere to the principle that it should generate similarly sized or, even better, equally sized groups. This principle is based on the following proposition.

Proposition 2.1 For normal populations with equal variances, the likelihood ratio test is most powerful if the sizes of the compared groups are equal. The *most powerful* test is the test that has the largest power among all tests with fixed probability of type I error.

Proof: Consider two normal populations with means μ_1 and μ_2 and equal variances σ^2. The test hypotheses are $H_0 : \mu_1 \leq \mu_2$ and $H_1 : \mu_1 > \mu_2$. Suppose two independent random samples of sizes n and $N - n$ are drawn from these populations, where N is a fixed number.

The variance of the difference of the sample means $\mathbb{V}ar(\bar{x}_1 - \bar{x}_2) = \frac{\sigma^2}{n} + \frac{\sigma^2}{N-n}$ is minimized if $n = N/2$ (see Exercise 2.15).

Consider two tests, one with $n = N/2$ and the other with $n \neq N/2$. Denote the corresponding variances of the difference in sample means by $v_{N/2} = 4\sigma^2/N$ and $v = \sigma^2/n + \sigma^2/(N-n)$, respectively. Let $\beta_{N/2}$ and β be the respective probabilities of type II error under the alternative hypothesis $H_1 : \mu_1 - \mu_2 = \delta$ for some $\delta > 0$. It will be shown that $\beta > \beta_{N/2}$.

The equations for α, $\beta_{N/2}$, and β follow (compare them to Equations 2.1 and 2.2, respectively):

$$1 - \alpha = \Phi(k_{N/2}) = \Phi(k)$$

$$\beta_{N/2} = \Phi\left(k_{N/2} - \frac{\delta}{\sqrt{v_{N/2}}}\right)$$

$$\beta = \Phi\left(k - \frac{\delta}{\sqrt{v}}\right)$$

where $k_{N/2}$ and k denote the critical values for the acceptance regions in the two tests, respectively. These equations imply the following relations:

$$k_{N/2} = \Phi^{-1}(1-\alpha) = k$$

$$k_{N/2} - \frac{\delta}{\sqrt{v_{N/2}}} = \Phi^{-1}(\beta_{N/2})$$

$$k - \frac{\delta}{\sqrt{v}} = \Phi^{-1}(\beta)$$

From here,

$$\Phi^{-1}(1-\alpha) - \frac{\delta}{\sqrt{v_{N/2}}} = \Phi^{-1}(\beta_{N/2})$$

$$\Phi^{-1}(1-\alpha) - \frac{\delta}{\sqrt{v}} = \Phi^{-1}(\beta)$$

Subtracting the equations and recalling that $v_{N/2} < v$ yields

$$0 > \frac{\delta}{\sqrt{v}} - \frac{\delta}{\sqrt{v_{N/2}}} = \Phi^{-1}(\beta_{N/2}) - \Phi^{-1}(\beta)$$

Hence,

$$\Phi^{-1}(\beta) > \Phi^{-1}(\beta_{N/2}) \qquad \text{or} \qquad \beta > \beta_{N/2}$$

□

2.3.2 Randomization Methods

Several randomization methods are used. The most common ones are the simple, block, and stratified procedures.

In the *simple* randomization procedure, each subject has equal probabilities of being assigned to any of the groups. For example, if there are two treatment groups and a control group, a new subject can be assigned to any of the three groups with probabilities $\frac{1}{3}$.

This type of randomization is carried out by means of a *table of random digits*, in which any digit $0, \ldots, 9$ appears in any position with probability $\frac{1}{10}$. These tables are often published in statistics books. Random-number-generating software such as Excel or Minitab is used in practice.

In our example, the randomization can be carried out by accepting, say, numbers 1, 2, and 3, and ignoring the other digits: 1 means the subject is allocated to the first treatment group; 2, to the second treatment group; and 3, to the control group. A more time-efficient way is to accept, say, 1, 2, or 3 for the first treatment group; 4, 5, or 6 for the second treatment group; 7, 8, or 9 for the control group; and to ignore 0. Either way, a subject has probability $\frac{1}{3}$ of being allocated to any of the three groups (see Exercise 2.16).

This simple method has one obvious disadvantage. For a small-sized clinical trial, the simple randomization may easily result in seriously unequal group sizes.

In the *block* randomization procedure, subjects are allocated by blocks, with the numbers assigned to each group being equal within each block. For example, subjects are randomized to two treatment groups A and B by blocks of size 4. Four subjects, enrolled sequentially, may be assigned to one of the blocks $AABB$, $ABAB$, $ABBA$, $BABA$, $BBAA$, or $BAAB$, where each block has the probability of $\frac{1}{6}$. The order of the blocks can be determined by the table of random digits, accepting $1, \ldots, 6$ and ignoring the other numbers. The main advantage of this method is that it achieves the balance in group sizes at the end of the trial, as well as the periodic balance at the end of each block.

In the *stratified* randomization procedure, subjects are allocated to groups in a way that achieves balance between groups in regard to certain characteristics, such as gender or age. These characteristics are called *prognostic factors* because they are used for *prognosis* (that is, prediction) of the chance of responding to the new treatment. In this method, several subgroups (called *strata*) are created for each combination of levels of prognostic factors, and block randomization is carried out within each stratum.

For example, suppose two variables, Gender and Age, are believed to have prognostic importance. The variable Gender has two levels (Male and Female), and the variable Age also has two levels ("Under 65" and "65 and Over"). Then there are four strata:

	Prognostic Factors		Sample
Stratum	Gender	Age	Allocation Assignment
1	Male	Under 65	AABB ABAB etc.
2	Male	65 and Over	BABA ABBA etc.
3	Female	Under 65	BBAA BBAA etc.
4	Female	65 and Over	BABA BAAB etc.

2.3.3 Concealment of Group Assignments

An actual group assignment should be kept secret from the subject as well as from the physician responsible for administering therapy, if the trial is double-blinded.

It is recommended that the randomization schedules be made for all future subjects before the trial begins, in an independent central location by a statistician (i.e., a person not involved with the treatment of subjects). After the randomization procedure is completed, the statistician should prepare sequentially numbered, nontransparent, sealed envelopes with the group assignments.

In a single-blinded trial, when a new subject enters the trial, the physician should call the central location where the next envelope in the sequence is opened to reveal the group assignment for the subject. It may be arranged for the physician to keep the box with the envelopes in the office; however, in such a case the assignments are less well protected from tampering.

In a double-blinded trial, the physician should receive a sequentially numbered container with the assigned treatment to be administered to the subject. The containers should be prepared by an independent clinician in some central location to avoid the possibility of fraud.

2.4 Data Reporting

Interim data reports may include the following comparisons and tests. For each type of adverse event, Kaplan–Meier curves (see Sections 3.2 and 3.3) may be constructed for each treatment group. The curves for each group may be compared using the log-rank test (see Section 3.4). Also, using the overall data, point estimators for each endpoint and confidence limits for the estimators may be computed. Alternatively, the results of the group sequential testing (see Subsection 2.2.1) may be presented. In addition, the overall data may be subdivided into several groups according to levels of certain prognostic factors, and the same analysis may be conducted within each group.

Statistical comparison of subjects for each site in a trial with a small number of centers may be done based on a number of demographic and pretrial health-related variables to establish poolability of the data. Sometimes, in a multicenter trial with a large number of centers, this comparison may be done between larger populations—for example, between U.S. and European sites.

Example 2.5 Suppose a new technique for repair of a torn meniscus is being tested in a nonrandomized, 24-month study. Investigators are interested in the complication rate (the total number of complications over the number of patient-years) of meniscus retear, joint pain/tenderness, and knee effusion.

Suppose that the gender of a patient is one of prognostic factors in the trial. Assume that the interim report at 6 months contains estimators of the complication rates for males, females, and total. It gives P-values for the two-sided

Table 2.2 Analysis of Complication Rates in Example 2.5.

Gender	Number of Patient-Years	Number of Complications	Complication Rate	*P*-Value	95% UCL
		Meniscus Retear			
Male	23.43	5	0.2134	0.3595	0.3704
Female	19.82	2	0.1009		0.2183
Total	43.25	7	0.1618		0.2625
		Joint Pain/Tenderness			
Male	23.43	12	0.5122	0.4319	0.7554
Female	19.82	7	0.3532		0.5728
Total	43.25	19	0.4393		0.6051
		Knee Effusion			
Male	23.43	8	0.3414	0.0366*	0.5400
Female	19.82	1	0.0505		0.1335
Total	43.25	9	0.2081		0.3222

*Significantly different complication rates at the 5% significance level.

z-test for equality of complication rates for males and females, and it presents 95% upper confidence limits (UCL) for the estimators. The findings are summarized in Table 2.2.

The theory behind the confidence limits and the test statistic is as follows. Denote by $\lambda = X/T$ the true rate, where X is the number of complications in a population and T is the number of patient-years. It is assumed that T is a large constant and X is a Poisson random variable, the mean of which can be estimated using the maximum likelihood method by the observed number of complications n. Therefore, the maximum likelihood estimators of the mean and variance of λ are $\widehat{\mathbb{E}(\lambda)} = n/T$ and $\widehat{\mathbb{V}ar(\lambda)} = n/T^2$, respectively. Because T is large, the normal approximation is valid and an approximate $100(1-\alpha)\%$ upper confidence limit for λ is

$$n/T + z_\alpha \sqrt{n}/T \quad \text{where } z_\alpha = \Phi^{-1}(1-\alpha)$$

To test $H_0 : \lambda_1 = \lambda_2$ against $H_1 : \lambda_1 \neq \lambda_2$, where λ_1 and λ_2 are the true complication rates in the two populations, compute the z-statistic:

$$z = \frac{\frac{n_1}{T_1} - \frac{n_2}{T_2}}{\sqrt{\frac{n_1+n_2}{T_1+T_2}\left(\frac{1}{T_1} + \frac{1}{T_2}\right)}} \tag{2.18}$$

Table 2.3 Analysis of Proportions of Complications in Example 2.5.

Gender	Number of Subjects	Number of Complications	Proportion of Complications	*P*-Value	95% UCL
		Meniscus Retear			
Male	74	5	0.0676	0.4127	0.1173
Female	58	2	0.0345		0.0746
Total	132	7	0.0530		0.0860
		Joint Pain/Tenderness			
Male	74	12	0.1622	0.5331	0.2392
Female	58	7	0.1207		0.1957
Total	132	19	0.1439		0.1983
		Knee Effusion			
Male	74	8	0.1081	0.0472*	0.1710
Female	58	1	0.0172		0.0456
Total	132	9	0.0682		0.1056

*Significantly different proportions of complications at the 5% significance level.

where n_1, n_2 are the observed number of complications, and T_1, T_2 are the number of patient-years in the two groups, respectively (for derivation of this test statistic refer to Exercise 2.17). Under H_0, the test statistic has an approximately $\mathcal{N}(0,1)$ distribution.

Alternatively, or in addition to the complication rate, it is advisable to report the *proportion of complications*, defined as the ratio between the observed number of complications and the total number of subjects.

Table 2.3 gives the estimated proportions of complications for males, females, and total subjects; *P*-values for the *z*-test; and 95% upper confidence limits (UCL) for the estimators.

The following reasoning is used for the confidence limits and the test statistic. Let N be the number of subjects, and let n be the observed number of complications. It is customary to assume that N is a constant, much larger than n. Denote by X the population number of complications. It is assumed that X has a Poisson distribution, for which the maximum likelihood estimator of the mean equals n. Let $p = X/N$ be the population complication rate. Then $\widehat{\mathbb{E}(p)} = n/N$ and $\widehat{\mathbb{V}ar(p)} = n/N^2$. Because N is very large, the normal approximation is valid, and an approximate $100(1-\alpha)\%$ upper confidence limit for p is

$$n/N + z_\alpha \sqrt{n}/N \quad \text{where } z_\alpha = \Phi^{-1}(1-\alpha)$$

To test $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$, where p_1 and p_2 are the true complication rates in two populations, compute the z-statistic:

$$z = \frac{\frac{n_1}{N_1} - \frac{n_2}{N_2}}{\sqrt{\frac{n_1+n_2}{N_1+N_2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \tag{2.19}$$

where n_1, n_2 are the observed number of complications, and N_1, N_2 are the number of subjects in the two groups, respectively (for derivation of this test statistic, refer to Exercise 2.18). Under H_0, the test statistic has an approximately $\mathcal{N}(0, 1)$ distribution. □

Exercises for Chapter 2

Section 2.1

Exercise 2.1 Derive Equation 2.3 from Equations 2.1 and 2.2. □

Exercise 2.2 Researchers would like to test a new therapy. They are planning to conduct a randomized clinical trial in which group A receives the tested therapy and group B receives a therapy that is currently in use. Investigators are unsure about the efficacy of the therapy. They propose, therefore, to test $H_0 : \mu_A = \mu_B$ against a two-sided alternative $H_1 : \mu_A \neq \mu_B$, where μ_A and μ_B are the mean responses for group A and group B, respectively. The probability of type I error is specified as 0.05, and the power of the test is fixed at 0.85, provided $\mu_A - \mu_B = 7$ units. An estimated population standard deviation is $\sigma = 16$ units. Calculate the required group size for this clinical trial and the actual probability of type II error that corresponds to this group size. □

Exercise 2.3 In Example 2.2, show that (a) the acceptance region for the likelihood ratio test is of the form $\{x > x_0\}$ for some integer constant x_0, and (b) for a fixed x_0, $\alpha = \max_{\lambda \geq 2\lambda_h} \mathbb{P}(X \leq x_0)$ corresponds to the case $\lambda = 2\lambda_h$. Hint: Show that if $X \sim$ Poisson(λ), and $H_0 : \lambda \geq \lambda_0$ is tested against $H_1 : \lambda < \lambda_0$, then (a) the likelihood ratio is

$$\Lambda(x) = \frac{\max\limits_{\lambda \geq \lambda_0} \lambda^x e^{-\lambda}/x!}{\max\limits_{\lambda > 0} \lambda^x e^{-\lambda}/x!} = \begin{cases} 1, & \text{if } x \geq \lambda_0 \\ (\lambda_0/x)^x e^{-(\lambda_0 - x)}, & \text{if } x \leq \lambda_0 \end{cases}$$

(b) $\alpha = \max_{\lambda \geq \lambda_0} \mathbb{P}(X \leq x_0)$ corresponds to $\lambda = \lambda_0$. □

Exercise 2.4 Derive Equation 2.6 probabilistically. Hint: Use the following argument. Suppose N_t is a random number of events in the interval $[0, t]$. Then

$N_t \sim \text{Poisson}(\lambda t)$. Let T_n be the waiting time for the nth event. Show that $T_n \sim \text{Gamma}(n, 1/\lambda)$ with the density

$$f_{T_n}(y) = \frac{\lambda^n y^{n-1}}{\Gamma(n)} e^{-\lambda y}, \quad \lambda > 0,\ y > 0$$

where $\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx$ is the gamma function. Thus $\mathbb{P}(N_t > n) = \mathbb{P}(N_t \geq n + 1) = \mathbb{P}(T_{n+1} < t) = \int_0^t \frac{\lambda^{n+1} y^n}{\Gamma(n+1)} e^{-\lambda y} dy = \int_0^{\lambda t} \frac{u^n}{\Gamma(n+1)} e^{-u} du$. Now finish the argument by noticing that the value of n in the integral can be any real number, not necessarily an integer.

You can also verify that $\mathbb{P}(N_t = n) = \mathbb{P}(N_t > n - 1) - \mathbb{P}(N_t > n) = \frac{1}{n!} \int_0^\lambda (n u^{n-1} - u^n) e^{-u} du = \frac{u^n}{n!} e^{-u} \big|_0^\lambda = \frac{\lambda^n}{n!} e^{-\lambda}$. □

Section 2.2

Subsection 2.2.1

Exercise 2.5 Check that the solution of Equation 2.11 is $k = 1.875$ and $n^* = 3.029$. Use Matlab or similar software. □

Exercise 2.6 Show that, for a general N, the quantities k and n^* solve Equation 2.12. □

Exercise 2.7 Show that, in Equation 2.12 with $N = 3$, the quantities $\alpha' = 0.023$ and $n = 39$. Compute the actual probability of type II error that corresponds to this group size. Explain step-by-step how this sequential testing is carried out. Compare the maximum required group sizes for $N = 1, 2,$ and 3. □

Exercise 2.8 Consider Exercise 2.2. Suppose researchers would like to conduct interim analyses for this trial. For a general N, derive a system of equations similar to Equation 2.12, and draw schematically the acceptance region for the mth test, $m = 1, \ldots, N$. Solve the system numerically for $N = 2$ and $N = 3$. Compute the probabilities of type II error that correspond to the interim group sizes. Compare the maximum required group sizes for $N = 1$, 2, and 3. □

Exercise 2.9 The classical group sequential method is applicable to the clinical trial of Example 2.2. Suppose N interim tests are conducted at times mt, $m = 1, \ldots, N$. The acceptance region for the mth test is

$$\left\{ \frac{X_{mt} - 0.024mt}{\sqrt{0.024mt}} > k \right\} = \left\{ X_{mt} > 0.024mt + k\sqrt{0.024mt} \right\}$$

(a) Show that the equation for α, the overall probability of type I error, is

$$\alpha = \mathbb{P}\left(\bigcap_{m=1}^{N}\left\{\frac{X_{mt} - 0.024mt}{\sqrt{0.024mt}} \leq k\right\}\right), \quad X_{mt} \sim \text{Poisson}(0.024mt)$$

$$= \mathbb{P}\left(\bigcap_{m=1}^{N}\{Z_1 + \cdots + Z_m \leq \sqrt{mk}\}\right)$$

where $Z_1, \ldots, Z_N$ are independent $\mathcal{N}(0,1)$ random variables.

(b) Show that the equation for β, the overall probability of type II error, is

$$1 - \beta = \mathbb{P}\left(\bigcap_{m=1}^{N}\left\{\frac{X_{mt} - 0.024mt}{\sqrt{0.024mt}} \leq k\right\}\right), \quad X_{mt} \sim \text{Poisson}(0.012mt)$$

$$= \mathbb{P}\left(\bigcap_{m=1}^{N}\left\{Z_1 + \cdots + Z_m \leq \sqrt{2mk} + m\sqrt{0.012t}\right\}\right)$$

where $Z_1, \ldots, Z_N$ are independent $\mathcal{N}(0,1)$ random variables.

(c) Compute numerically the values of k and t for $N = 2$, $\alpha = 0.05$, and $\beta = 0.2$. Describe step-by-step how the test is carried out.

(d) Show that α', the interim probability of type I error, is a constant for a fixed N and is computed by $\alpha' = \Phi(k)$. Calculate the value of α' for $N = 2$.

(e) Draw the acceptance region for the mth test, $m = 1, \ldots, N$. □

Exercise 2.10 The objective of this exercise is to show how a nonclassical sequential method may be used for interim data analyses in Example 2.2. Consider only the case $N = 2$. Suppose the first test is conducted at t patient-years, and the second at $2t$ patient-years. In both tests, the null hypothesis is accepted if the observed complication rate is larger than a fixed critical value K.

(a) Show that t and K can be computed from the equations for the overall probabilities of type I and II errors, $\alpha = 0.05$ and $\beta = 0.2$,

$$\alpha = \mathbb{P}(X_t \leq Kt,\ X_t + Y_t \leq 2Kt)$$

where X_t and Y_t are independent Poisson$(0.024t)$ random variables

$$= \left[\sum_{i=0}^{Kt}\frac{(0.024t)^i}{i!}e^{-0.024t}\right]^2 + \sum_{i=0}^{Kt}\sum_{j=Kt+1}^{2Kt-i}\frac{(0.024)^{i+j}}{i!j!}e^{-0.048t}$$

and

$$1 - \beta = \mathbb{P}(X_t \leq Kt,\ X_t + Y_t \leq 2Kt)$$

where X_t and Y_t are independent Poisson$(0.012t)$ random variables

$$= \left[\sum_{i=0}^{Kt} \frac{(0.012t)^i}{i!} e^{-0.012t}\right]^2 + \sum_{i=0}^{Kt} \sum_{j=Kt+1}^{2Kt-i} \frac{(0.012)^{i+j}}{i!j!} e^{-0.024t}$$

(b) Draw the acceptance regions for these two tests.

(c) Check that an approximate numerical solution of these equations is $t = 500$ and $K = 0.016$, which corresponds to the overall probabilities of type I and II errors $\alpha = 0.0401$ and $\beta = 0.1917$.

(d) Using the results of part (c), describe step-by-step how this testing is carried out.

(e) Show that α', the interim probability of type I error, relates to K by the formula $\alpha' = \mathbb{P}(X_t \leq Kt)$, where $X_t \sim$ Poisson$(0.024t)$. Compute α' for the first and the second tests. □

Subsection 2.2.2

Exercise 2.11 Show that the mode of a Gamma(a, b) distribution is $(a-1)b$. □

Exercise 2.12 The number of events has a Poisson(Rt) distribution, and the prior distribution of the random variable R is Gamma(a, b) with the density

$$\pi(x) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}, \quad x, a, b > 0$$

Suppose that n events are observed during a time period t. Show that the posterior distribution of R is Gamma$(n + a, 1/(t + 1/b))$. □

Exercise 2.13 In Example 2.4, assume that researchers are optimistic about the tested valve, and assign a 0.7 prior probability to the alternative hypothesis. Redo the calculations to determine stopping rules similar to the ones given in Table 2.1. Compare the results. □

Exercise 2.14 Suppose researchers want to conduct an interim Bayesian analysis for the trial in Example 2.1. Let $\mu = \mu_{tr} - \mu_c$. The alternative hypothesis of interest is $H_1 : \mu > 0$. The distribution of $\bar{x} = \bar{x}_{tr} - \bar{x}_c$ is $\mathcal{N}(\mu, 2\sigma^2/n)$, where μ is modeled as a random variable. A natural choice of a conjugate prior for μ is a normal distribution (Prove it!). A skeptical prior is used with zero mean and a large variance, which for computational convenience is chosen to be σ^2. At group size $n = 50$, the interim Bayesian test is carried out. H_1 is accepted if its posterior probability is at least 0.95; it is rejected if its probability does not exceed 0.05. Otherwise, the trial continues until the minimum required group size of 97 subjects is accrued.

(a) Show that the prior probability of a true alternative is 0.5.

(b) Show that the posterior distribution of μ given $\bar{x}$ is

$$\mathcal{N}\left(\frac{\bar{x}}{1+2/n}, \frac{2\sigma^2/n}{1+2/n}\right)$$

(c) Find the values of the sample mean $\bar{x}$ for which the interim test accepts or rejects H_1. Describe the stopping rule. □

Section 2.3

Exercise 2.15 Show that in the proof of Proposition 2.1, the variance $\mathbb{V}ar(\bar{x}_1 - \bar{x}_2) = \frac{\sigma^2}{n} + \frac{\sigma^2}{N-n}$ is minimized for $n = N/2$. □

Exercise 2.16 Show that for both randomization methods described under the simple randomization procedure in Section 2.3, the subject is allocated to any group with probability $\frac{1}{3}$. □

Section 2.4

Exercise 2.17 Derive the test statistic given in Equation 2.18, and verify the entries in Table 2.2. □

Exercise 2.18 Derive the test statistic given in Equation 2.19, and verify the entries in Table 2.3. □

Chapter 3

Introduction to Survival Analysis

One of the variables of interest to clinical researchers is the length of time a subject stays in the trial. This chapter focuses on the fundamentals of the statistical analysis of these observed times.

3.1 Basic Definitions

Survival analysis consists of studies of the *survival time* of a subject (usually measured in days, weeks, months, or years), which is the time that elapses between the baseline and the moment an adverse event occurs, or the subject drops out of the trial. Sometimes the survival time is called a *lifetime* or an *event time.*

The survival times for subjects who dropped out of the trial (called *drop-outs* or *lost to follow-up subjects*) are *right-censored* (or, more simply, *censored*). The survival times of the subjects who remain in the trial until it ends are censored as well. This term applies to situations when it is known that the subject survived a certain length of time and was healthy, but the later health condition for this subject is not recorded.

Censored survival times represent very important information and should be kept in the database. Retained censored survival times increase the overall *survival rate* of the subjects—that is, the percentage of people who are alive for a given period of time. For example, if a subject drops out after being in a study for 5 months, the subject is still included in calculation of the survival rate up to 5 months. Naturally, a higher survival rate implies a better treatment efficacy.

In what follows, each uncensored observation is termed "death," regardless of whether a death or a different adverse event has occurred. Denote by T the random variable representing the survival time of a subject. Let $f(t)$, $t \geq 0$, denote the probability density function (pdf) of T, and let $F(t) = \mathbb{P}(T \leq t) = \int_0^t f(x)\,dx$, $t \geq 0$, be the cumulative distribution function (cdf) of T. The distribution of T is called the *survival time distribution* (or the *lifetime distribution*).

The objective of survival analysis is to estimate and model the following functions:

- The *survival function*, $S(t)$, defined as the probability that a subject survives up to time t:

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(x)\,dx = 1 - F(t)\,, \quad t \geq 0 \qquad (3.1)$$

- The *hazard function*, $h(t)$, defined as the following ratio:

$$h(t) = \frac{f(t)}{S(t)}\,, \quad t \geq 0 \qquad (3.2)$$

  It is interpreted as an *instantaneous death rate*, since the probability that a subject dies within an infinitesimally small time interval $[t,\, t+dt)$, given that the subject survived up to time t, $t \geq 0$, is equal to

$$\mathbb{P}(T < t + dt \,|\, T > t) = \frac{\mathbb{P}(t < T < t + dt)}{\mathbb{P}(T > t)} = \frac{f(t)\,dt}{S(t)} = h(t)\,dt$$

- The *cumulative hazard function*, $H(t)$, defined by

$$H(t) = \int_0^t h(x)\,dx, \quad t \geq 0 \qquad (3.3)$$

Example 3.1 Lifetime distributions are commonly modeled by the exponential distribution with the density

$$f(t) = \lambda \exp\{-\lambda t\}, \quad t \geq 0\,,\ \lambda > 0$$

The cdf of this distribution is $F(t) = 1 - \exp\{-\lambda t\}$, $t \geq 0$. Therefore, the survival function is given by $S(t) = 1 - F(t) = \exp\{-\lambda t\}$, $t \geq 0$. By definition, the hazard function is $h(t) = f(t)/S(t) = \lambda$, $t \geq 0$, and the cumulative hazard function is $H(t) = \int_0^t h(x)\,dx = \lambda t$, $t \geq 0$. □

The questions that are addressed in subsequent sections of this chapter include nonparametric and parametric estimations of the survival function as well as the regression modeling of the survival and hazard functions.

The techniques of survival analysis are illustrated by means of SAS, a statistical software package that is widely implemented by clinical researchers and is highly praised for its notable capability in analyzing medical data.

3.2 Estimation of Survival Function by the Kaplan–Meier Method

3.2.1 Definition of the Kaplan–Meier Estimator

A widely used method for estimation of the survival function is the Kaplan–Meier method. This method produces the *Kaplan–Meir estimator*, a nonparametric estimator, which does not assume any known algebraic form of the estimated survival function. The Kaplan–Meier estimator is also referred to as the *KM estimator* or the *product-limit estimator.*

Suppose k distinct survival times are observed. Arranged in increasing order, they are $t_1 < t_2 < \cdots < t_k$. At time t_i, there are n_i subjects who are said to be *at risk*—that is, they survived up to this time (not including it) and were not censored. Denote by d_i the number of subjects who die at time t_i. To simplify notation, let $t_0 = 0$ and $d_0 = 0$. Then the Kaplan–Meier estimator of the survival function $S(t)$ is

$$\hat{S}(t) = \prod_{i\,:\,t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \quad t \geq 0 \tag{3.4}$$

Example 3.2 A biotech company conducted a 2-year clinical trial testing the efficacy of a new heart valve. The survival times (in months) of 10 patients with the heart valve implants were recorded. The plus sign "+" next to the observation signifies that the observation is censored. The data are

24+, 16+, 8, 19, 10, 8+, 5, 17, 20, 10

There are eight distinct survival times, given here in increasing order:

5, 8, 10, 16, 17, 19, 20, 24

Table 3.1 aids in the estimation of the survival function.

In SAS, `lifetest` procedure is used to estimate the survival function by the KM method. In fact, the KM method is the default for this procedure.

The variable `status` used in the definition of the data set in this example is an indicator of a death (or an event) occurring; that is, `status` = 1 for uncensored observations, and 0 for censored ones. The SAS code for this example follows:

```
data valves;
input duration status @@;
datalines;
 24  0    16  0     8  1    19  1    10  1
  8  0     5  1    17  1    20  1    10  1
  ;
```

```
proc lifetest data = valves method = km;
   time duration * status(0);
/* status(0) means status = 0 on censored observations */
run;
```

The SAS output for this example includes the survival times, with censored observations marked by a star, and the KM estimator. A "missing" value in the second column indicates that the estimator of the survival function retains its previous value at this point.

```
Product-Limit Survival Estimates
 duration   Survival
  0.0000     1.0000
  5.0000     0.9000
  8.0000     0.8000
  8.0000*      .
 10.0000       .
 10.0000     0.5714
 16.0000*      .
 17.0000     0.4286
 19.0000     0.2857
 20.0000     0.1429
 24.0000*      .
```

□

Table 3.1 Estimation of $S(t)$ by the Kaplan–Meier Method in Example 3.2

Time t_i	At Risk n_i	Died d_i	Censored at Time t_i	Survival Rate $\left(1-\frac{d_i}{n_i}\right)$	Estimator $\hat{S}(t)$, $t_i < t < t_{i+1}$
0	10	0	0	$1-0=1.00$	1.00
5	10	1	0	$1-\frac{1}{10}=0.90$	$(1.00)(0.90)=0.90$
8	9	1	1	$1-\frac{1}{9}=0.89$	$(0.90)(0.89)=0.80$
10	7	2	0	$1-\frac{2}{7}=0.71$	$(0.80)(0.71)=0.57$
16	5	0	1	$1-0=1.00$	$(0.57)(1.00)=0.57$
17	4	1	0	$1-\frac{1}{4}=0.75$	$(0.57)(0.75)=0.43$
19	3	1	0	$1-\frac{1}{3}=0.67$	$(0.43)(0.67)=0.29$
20	2	1	0	$1-\frac{1}{2}=0.50$	$(0.29)(0.50)=0.15$
24	1	0	1	$1-0=1.00$	$(0.15)(1.00)=0.15$

3.2.2 Derivation of the Kaplan–Meier Estimator

The idea behind the Kaplan–Meier estimator of the survival function, given in Equation 3.4, is the following. Consider the recursive equation

$$\begin{aligned} S(t_i) &= \mathbb{P}\big(T > t_i\big) = \mathbb{P}\big(T > t_i \,|\, T > t_{i-1}\big)\, \mathbb{P}\big(T > t_{i-1}\big) \\ &= \mathbb{P}\big(T > t_i \,|\, T > t_{i-1}\big)\, S(t_{i-1})\,, \; i = 1, \ldots, k \end{aligned} \tag{3.5}$$

where $t_0 = 0$ and $S_0 = 1$.

By nature, T is a continuous random variable. Thus, theoretically speaking, identical observations (commonly termed *tied observations*) are not possible. In reality, however, survival times are measured on a certain scale (e.g., days, months, years), hence allowing tied observations; for instance, the data in Example 3.2 have two pairs of tied observations, at 8 and 10 months, respectively. Therefore, it is convenient to model the survival time T as a discrete random variable taking on values $t_1 < t_2 < \cdots < t_k$. Denote by π_i the conditional probability that a subject survives time t_i, given that the subject survived time t_{i-1}:

$$\pi_i = \mathbb{P}\big(T > t_i \,|\, T > t_{i-1}\big)\,, \quad i = 1, \ldots, k$$

Then, in view of Equation 3.5,

$$S(t_i) = \prod_{j=1}^{i} \pi_j \tag{3.6}$$

The maximum-likelihood method is used to estimate the values of π_i. At time t_i, there are d_i subjects, each of whom dies with probability $1 - \pi_i$ independently of the others, and there are $n_i - d_i$ subjects, each of whom survives with probability π_i, independently of the others. Therefore, the likelihood function is

$$L\big(\pi_1, \ldots, \pi_k\big) = \prod_{i=1}^{k} (1 - \pi_i)^{d_i}\, \pi_i^{n_i - d_i} \tag{3.7}$$

Equivalently, the log-likelihood function equals

$$\ln L\big(\pi_1, \ldots, \pi_k\big) = \sum_{i=1}^{k} \big[\, d_i \ln (1 - \pi_i) + (n_i - d_i) \ln \pi_i \,\big]$$

Equating to zero the derivatives of the log-likelihood function with respect to π_i gives the normal equations

$$\frac{d_i}{1 - \pi_i} = \frac{n_i - d_i}{\pi_i}\,, \quad i = 1, \ldots, k$$

These equations are solved to produce the maximum-likelihood estimators

$$\hat{\pi}_i = 1 - \frac{d_i}{n_i}, \quad i = 1, \ldots, k$$

Plugging these estimators into Equation 3.6 yields

$$\hat{S}(t_i) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j}\right)$$

Now, fix a time t, and suppose $t_i < t < t_{i+1}$ for some $i = 1, \ldots, k$. Because there are no deaths occurring between the survival times t_i and t_{i+1}, the survival function at time $t, S(t)$, is estimated by $\hat{S}(t) = \hat{S}(t_i)$. This gives Equation 3.4.

3.3 The Kaplan–Meier Survival Curve

The *Kaplan–Meier survival curve* is the plot of the Kaplan–Meier estimator of the survival function $\hat{S}(t)$ against time t. This curve is a step-function that decreases at the times of deaths. The censored times are usually marked by a cross (×). If a death and a censoring occur at the same time, a cross for the censored observation is put at the bottom of the step.

Example 3.3 In Example 3.2, the Kaplan–Meier survival curve is a plot of $\hat{S}(t)$ given in the last column of Table 3.1 against time t (see Figure 3.1). For instance, from this plot, the estimated probability of 15-month survival is 0.57.

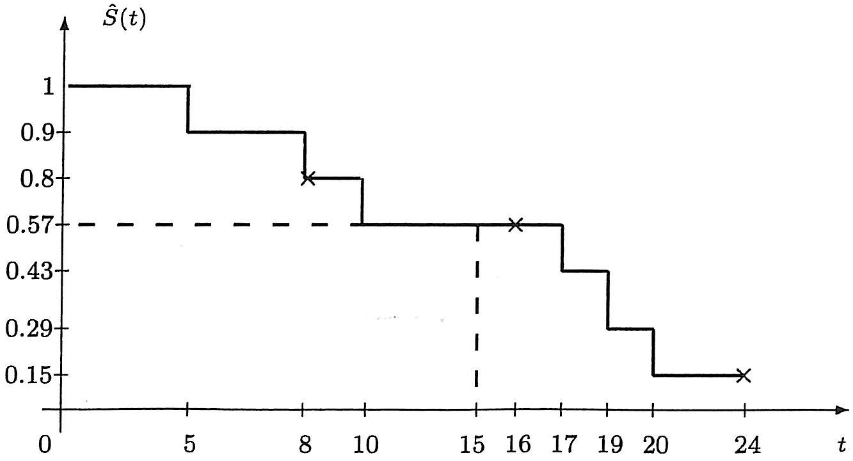

Figure 3.1 The Kaplan–Meier survival curve in Example 3.2

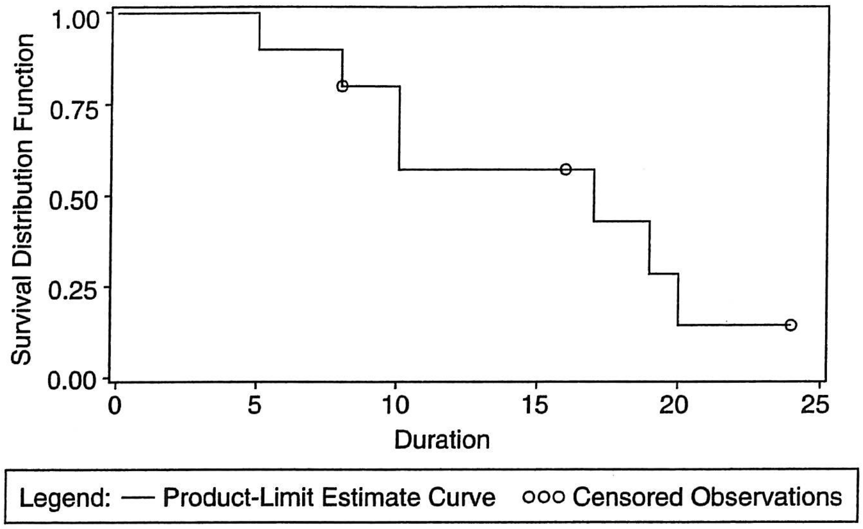

Figure 3.2 The Kaplan–Meier survival curve in Example 3.2 plotted using SAS software

To request the Kaplan–Meier survival curve in SAS, use the following code:

```
proc lifetest data = valves method = km plots = (survival);
  time duration * status(0);
run;
```

The resulting graph is given in Figure 3.2. Note that SAS uses a circle (o) as a default symbol for marking censored observations. □

3.4 Comparison of Two Survival Functions: Log-Rank Test

To compare the efficacy of two treatments, subjects who enter a clinical trial are randomly placed into two treatment groups. The survival data are then recorded for each group. The question of interest is whether the two treatments are equally effective. This translates into testing whether the survival functions for these groups differ significantly. A two-sided test of statistical hypotheses is appropriate. The hypotheses are

$$\begin{aligned} H_0 &: S_1(t) = S_2(t) \quad \text{for all } t \\ H_1 &: S_1(t) \neq S_2(t) \quad \text{for some } t \end{aligned}$$

The most commonly used test for data with censored observations is the *log-rank test*, which derives its name from the fact that it is related to a test that uses logarithms of ranks of observations.

To compute the log-rank statistic, proceed as follows. Denote by $t_1 < t_2 < \cdots < t_k$ the ordered uncensored observations (times of deaths) in both samples combined. At each time t_i, the data can be summarized by a 2×2 table:

	Status of Subject		
Group	Died	Survived	Total
1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Here d_{1i} and d_{2i} are the numbers of subjects who died at time t_i in groups 1 and 2, respectively; $d_i = d_{1i} + d_{2i}$; n_{1i} and n_{2i} are the numbers of subjects at risk at time t_i in groups 1 and 2, respectively; and $n_i = n_{1i} + n_{2i}$.

The null hypothesis is equivalent to independence of the "group" and "status of subject" variables in all 2×2 tables. Under H_0, d_{1i} is a hypergeometric random variable with parameters n_i (the population size), n_{1i} (the size of the group of interest), and d_i (the sample size). The expected value of d_{1i} is

$$\mathbb{E}(d_{1i}) = \frac{n_{1i}\, d_i}{n_i}$$

The variance is

$$\mathbb{V}ar(d_{1i}) = \frac{n_{1i}\, n_{2i}\, (n_i - d_i)\, d_i}{n_i^2\, (n_i - 1)}$$

Summing over all i, $i = 1, \ldots, k$, yields a statistic

$$U = \sum_{i=1}^{k} \left(d_{1i} - \mathbb{E}(d_{1i})\right)$$

where

$$\mathbb{E}(U) = 0 \quad \text{and} \quad \mathbb{V}ar(U) = \sum_{i=1}^{k} \frac{n_{1i}\, n_{2i}\, (n_i - d_i)\, d_i}{n_i^2\, (n_i - 1)}$$

Standardizing leads to the log-rank test statistic

$$z = \frac{U}{\sqrt{\mathbb{V}ar(U)}}$$

which has an approximately $\mathcal{N}(0, 1)$ distribution. Alternatively (in particular, in SAS), the log-rank statistic is z^2, which has an approximately chi-squared distribution with one degree of freedom.

Example 3.4 A clinical trial is conducted to evaluate a new nicotine patch. Subjects are randomly assigned to either the treatment group or the control group. The treatment group receives the nicotine patch under study, while the control group receives the best nicotine patch currently available on the market. The measurement is the length of time (in months) that a subject goes without a cigarette. The data for the two groups are as follows:

Treatment	3.4	3.6+	4.1	4.9+	5.8+
Control	2.0	3.7+	4.3	4.9+	

The researchers would like to know whether the two nicotine patches differ significantly, so a log-rank test is performed. The test hypotheses are

$$\begin{aligned} H_0 &: S_{\text{treatment}}(t) = S_{\text{control}}(t) \quad \text{for all } t \\ H_1 &: S_{\text{treatment}}(t) \neq S_{\text{control}}(t) \quad \text{for some } t \end{aligned}$$

The times of events in both groups combined are 2.0, 3.4, 4.1, and 4.3. The 2×2 tables corresponding to each of these times follow:

$$t_1 = 2.0$$

	Status of Subject		
Group	Died	Survived	Total
Treatment	0	5	5
Control	1	3	4
Total	1	8	9

$$d_{11} = 0,\ \mathbb{E}(d_{11}) = \frac{(5)(1)}{9} = \frac{5}{9},\ \mathbb{V}ar(d_{11}) = \frac{(5)(4)(8)(1)}{(9)^2(8)} = \frac{20}{81}$$

$$t_2 = 3.4$$

	Status of Subject		
Group	Died	Survived	Total
Treatment	1	4	5
Control	0	3	3
Total	1	7	8

$$d_{12} = 1,\ \mathbb{E}(d_{12}) = \frac{(5)(1)}{8} = \frac{5}{8},\ \mathbb{V}ar(d_{12}) = \frac{(5)(3)(7)(1)}{(8)^2(7)} = \frac{15}{64}$$

$$t_3 = 4.1$$

	Status of Subject		
Group	Died	Survived	Total
Treatment	1	2	3
Control	0	2	2
Total	1	4	5

$$d_{13} = 1,\ \mathbb{E}(d_{13}) = \frac{(3)(1)}{5} = \frac{3}{5},\ \mathbb{V}ar(d_{13}) = \frac{(3)(2)(4)(1)}{(5)^2(4)} = \frac{6}{25}$$

$$t_4 = 4.3$$

	Status of Subject		
Group	Died	Survived	Total
Treatment	0	2	2
Control	1	1	2
Total	1	3	4

$$d_{14} = 0,\ \mathbb{E}(d_{14}) = \frac{(2)(1)}{4} = \frac{1}{2},\ \mathbb{V}ar(d_{14}) = \frac{(2)(2)(3)(1)}{(4)^2(3)} = \frac{1}{4}$$

Consequently,

$$\begin{aligned} U &= \left(0 - \frac{5}{9}\right) + \left(1 - \frac{5}{8}\right) + \left(1 - \frac{3}{5}\right) + \left(0 - \frac{1}{2}\right) \\ &= -0.2806 \\ \mathbb{V}ar(U) &= \frac{20}{81} + \frac{15}{64} + \frac{6}{25} + \frac{1}{4} = 0.9713 \end{aligned}$$

The log-rank test statistic is $z = -0.2806/\sqrt{0.9713} = -0.2847$. The approximate P-value for the two-sided test is $2\,\mathbb{P}\big(Z > 0.2847\big) = 0.7759$. Alternatively, the test statistic is $z^2 = 0.081$ and the approximate P-value is $\mathbb{P}\big(\chi^2(1) > 0.081\big) = 0.7759$. Thus the null hypothesis of equal survival functions is not rejected at the 0.05 level of significance, and the conclusion is that the two nicotine patches do not differ significantly.

The SAS code for this example is as follows:

```
data patches;
input duration status group @@;
datalines;
```

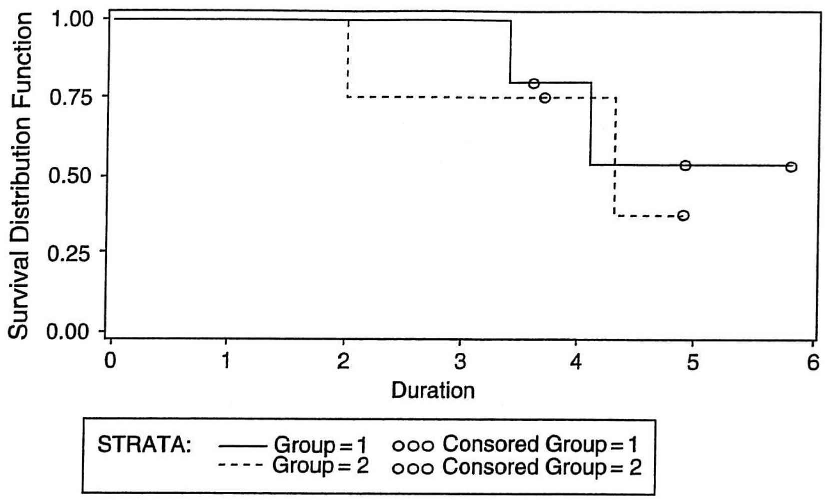

Figure 3.3 The two survival curves in Example 3.4 plotted by SAS

```
   3.4  1  1     3.6  0  1     4.1  1  1
   4.9  0  1     5.8  0  1     2.0  1  2
   3.7  0  2     4.3  1  2     4.9  0  2
;
proc lifetest data = patches method = km plots = (survival);
 time duration * status(0);
      strata group;
 symbol1 value = none color = black line = 1; /*solid line*/
 symbol2 value = none color = black line = 2; /*dashed line*/
run;
```

As part of the SAS output, the log-rank statistic z^2 and the corresponding P-value are computed. The result is shown here:

```
                                 Pr >
 Test       Chi-Square   DF   Chi-Square
Log-Rank      0.0810      1      0.7759
```

For graphical comparison, SAS plots the survival curves for the two groups in the same coordinate plane. The resulting graph is shown in Figure 3.3. Notice that the survival curves lie very close to each other, visually emphasizing the earlier conclusion that the survival rates for the two groups are not significantly different. □

3.5 Estimation of the Survival Function by the Actuarial Method

When the number of observations in a clinical trial is large and survival times are measured precisely, the data collected include many distinct values. As a result, the Kaplan–Meier approach to estimation of the survival function produces a long bulky table, and the survival curve is extremely saw-toothed.

In this case, the *actuarial method* is recommended as an alternative to the Kaplan–Meier estimation method. The corresponding estimator of the survival function is called the *actuarial estimator* (or *life-table estimator*). This section describes the steps for obtaining this estimator.

At the discretion of the researcher, the observed survival times are grouped into intervals, often of equal lengths. Then, for each time interval $[t_i,\ t_{i+1})$, the following quantities are computed:

- d_i, the number of subjects who died within the interval.
- c_i, the number of subjects who were censored within the interval.
- n_i, the number of subjects living at the beginning of the interval.
- $\tilde{n}_i = n_i - c_i/2$, the number of subjects at risk during the interval (in SAS, this number is called the *effective sample size*). Here an assumption is made that the censored observations are at risk for half of the interval.
- $1 - d_i/\tilde{n}_i$, the interval survival rate.

The estimator of the survival function at the beginning of an interval is calculated as the product of interval survival rates for all intervals up to and including the last one. That is, for the interval $[t_i,\ t_{i+1})$,

$$\hat{S}(t_i) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{\tilde{n}_j}\right)$$

The *actuarial survival curve* is a plot of the estimates $\hat{S}(t_i)$ against t_i, with the dots connected by straight lines. This plot depicts the "best guess" regarding the probability of survival as a function of time.

Example 3.5 A new drug is tested on patients with leukemia. The variable recorded is the duration of remission until a relapse occurs (in weeks). The censored observations are for the patients who were still in remission at the time the study terminated. The following data are collected:

1, 1, 2, 2, 3, 4, 6, 8, 8, 9, 9+,
10, 10, 11, 11, 11+, 12, 13+, 14, 14+, 15, 15,
15+, 16, 17, 17, 19, 20+, 23, 25+, 26, 27+, 29

Suppose it is reasonable to group the observations into six time intervals [0, 5), [5, 10), [10, 15), [15, 20), [20, 25), and [25, 30). Then the computations can be summarized by Table 3.2.

The actuarial survival curve is shown in Figure 3.4.

Table 3.2 Estimation of $S(t)$ by the Actuarial Method in Example 3.5

Interval	Died	Censored	At Risk	Interval Survival Rate	Survival Function
$[t_i,\ t_{i+1})$	d_i	c_i	$\tilde{n}_i$	$1 - d_i/\tilde{n}_i$	$\hat{S}(t_i)$
[0, 5)	6	0	33.0	$1 - \frac{6}{33} = 0.82$	1.00
[5, 10)	4	1	26.5	$1 - \frac{4}{26.5} = 0.85$	$(1.00)(0.82) = 0.82$
[10, 15)	6	3	20.5	$1 - \frac{6}{20.5} = 0.71$	$(0.82)(0.85) = 0.70$
[15, 20)	6	1	12.5	$1 - \frac{6}{12.5} = 0.52$	$(0.70)(0.71) = 0.50$
[20, 25)	1	1	5.5	$1 - \frac{1}{5.5} = 0.82$	$(0.50)(0.52) = 0.26$
[25, 30)	2	2	3.0	$1 - \frac{2}{3.0} = 0.33$	$(0.26)(0.82) = 0.21$

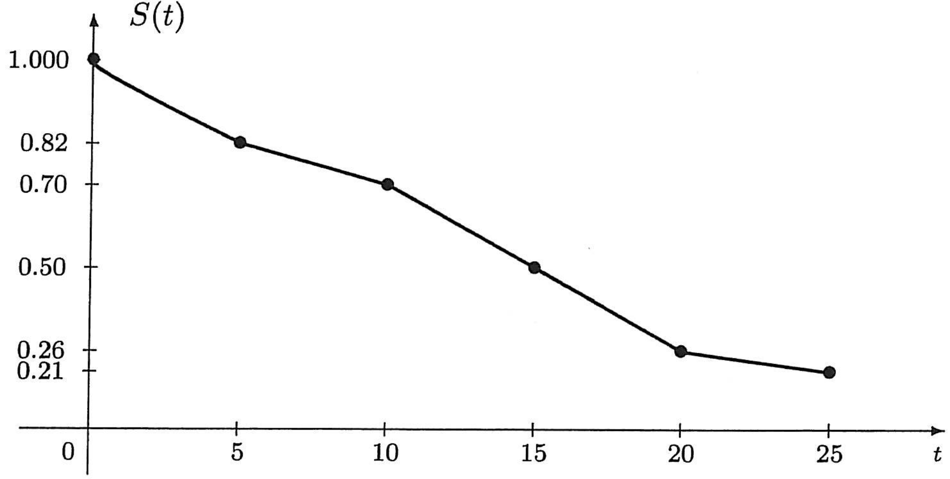

Figure 3.4 The actuarial survival curve in Example 3.5

The SAS code for this example is as follows:

```
data leukemia;
input duration status @@;
datalines;
  1  1     1  1     2  1     2  1     3  1
  4  1     6  1     8  1     8  1     9  1
  9  0    10  1    10  1    11  1    11  1
 11  0    12  1    13  0    14  1    14  0
 15  1    15  1    15  0    16  1    17  1
 17  1    19  1    20  0    23  1    25  0
 26  1    27  0    29  1
  ;
```

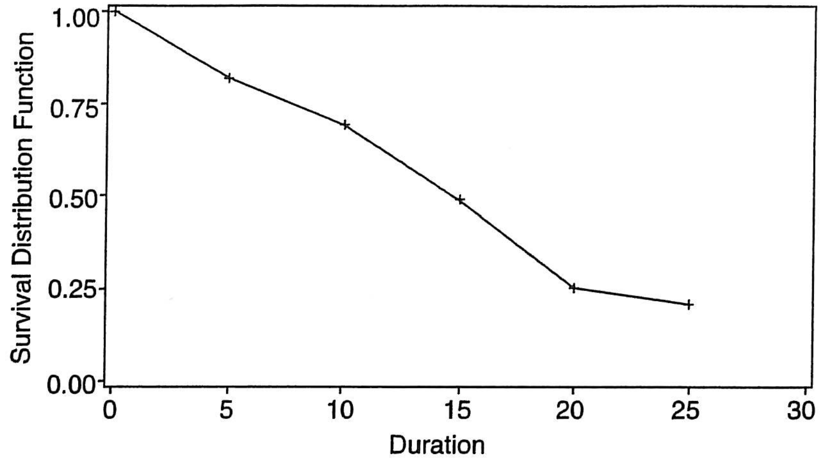

Figure 3.5 The actuarial survival curve in Example 3.5 plotted by SAS

```
proc lifetest data = leukemia method = act /*actuarial method*/
plots =(survival) intervals = 0, 5, 10, 15, 20, 25;
time duration * status(0);
run;
```

The SAS output for this example includes the following columns:

```
                                          Effective
     Interval        Number    Number      Sample
  [Lower,   Upper)   Failed   Censored      Size      Survival
     0        5        6         0          33.0       1.0000
     5       10        4         1          26.5       0.8182
    10       15        6         3          20.5       0.6947
    15       20        6         1          12.5       0.4914
    20       25        1         1           5.5       0.2555
    25        .        2         2           3.0       0.2091
```

The actuarial curve plotted by SAS is given in Figure 3.5. □

3.6 Parametric Estimation of the Survival Function

3.6.1 Definition and Setup

The survival function is estimated by the *parametric method* if a certain algebraic form of this function is assumed and the associated parameters are estimated. The exponential and Weibull distributions are commonly used to model the distribution of the survival time.

- The exponential distribution has density $f(t) = \lambda \exp\{-\lambda t\}$, and the survival function $S(t) = \exp\{-\lambda t\}$, where $t \geq 0$, $\lambda > 0$ (see Example 3.1).
- The density of the Weibull distribution is $f(t) = \alpha \lambda t^{\alpha-1} \exp\{-\lambda t^{\alpha}\}$, where $t \geq 0$, α, $\lambda > 0$. The survival function for this distribution equals to $S(t) = \exp\{-\lambda t^{\alpha}\}$, $t \geq 0$ (see Exercise 3.2). Note that when $\alpha = 1$, the Weibull distribution reduces to the exponential one.

The survival functions for these distributions are depicted schematically in Figure 3.6. The two graphs in this figure exhibit substantially different behavior. The survival probability in the exponential model drops quickly right from the baseline, and then levels off toward the end. In the Weibull model, the survival probability decays slowly for some period of time, then decreases rapidly, and possibly levels off at the end.

The Kaplan–Meier survival curve may give a hint about which model is more appropriate for a particular data set. If the empirical survival curve behaves similarly to either the exponential or Weibull distribution, then this model may have a good fit to the data.

As a rule, the estimation of parameters in these models is done by the maximum-likelihood method. However, specifying the likelihood function is not a straightforward task because the survival data are censored.

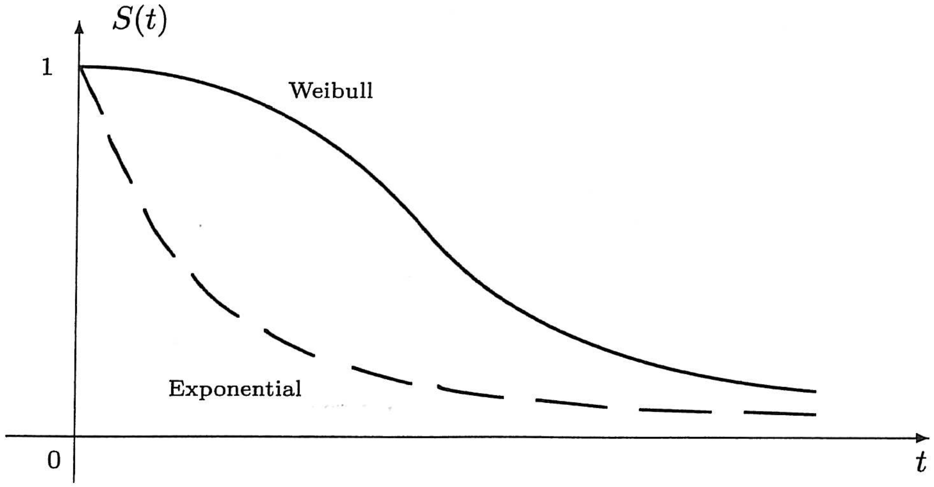

Figure 3.6 Survival functions for the exponential and Weibull distributions

3.6.2 Random Censoring Model

A usual approach is to consider a *random censoring* model. Suppose n pairs $(t_i,\ \delta_i)$ are observed, one for each subject in the trial, where t_i is the survival time, and δ_i is an indicator of a death occurring—that is, $\delta_i = 1$ if the observation is uncensored and $\delta_i = 0$ otherwise. Note that in the SAS code presented throughout this chapter, the variable δ_i is called `status`.

Consider two random variables T_i, the survival time of the ith subject, and C_i, the censoring time of the ith subject. Assume that T_i has the pdf $f_i(t)$ and cdf $F_i(t)$, $t \geq 0$, and that C_i has the pdf $g_i(t)$ and cdf $G_i(t)$, $t \geq 0$. In addition, assume that T_i and C_i are independent. The observed survival time is $\min(T_i,\ C_i)$.

The contribution to the likelihood function of the ith subject with the observed survival time t_i and $\delta_i = 1$ is

$$\lim_{dt \to 0} \frac{1}{dt}\, \mathbb{P}\big(\min(T_i,\ C_i) \in (t_i\,,\ t_i + dt)\,,\ \delta_i = 1\big)$$
$$= \lim_{dt \to 0} \frac{1}{dt}\, \mathbb{P}\big(T_i \in (t_i\,,\ t_i + dt)\,,\ C_i > t\big) = f_i(t)\big(1 - G_i(t)\big)$$

The contribution of the ith subject with the observed survival time t_i and $\delta_i = 0$ is

$$\lim_{dt \to 0} \frac{1}{dt}\, \mathbb{P}\big(\min(T_i,\ C_i) \in (t_i\,,\ t_i + dt)\,,\ \delta_i = 0\big)$$
$$= \lim_{dt \to 0} \frac{1}{dt}\, \mathbb{P}\big(T_i > t_i\,,\ C_i \in (t_i\,,\ t_i + dt)\big) = \big(1 - F_i(t)\big)g_i(t)$$

Thus the likelihood function for the survival time distribution with random censoring is

$$L = \prod_{i=1}^{n} \left[f_i(t_i)(1 - G_i(t_i))\right]^{\delta_i} \left[(1 - F_i(t_i))g_i(t_i)\right]^{1-\delta_i}$$
$$= \prod_{i=1}^{n} (1 - G_i(t_i))^{\delta_i} (g_i(t_i))^{1-\delta_i} \prod_{i=1}^{n} (f_i(t_i))^{\delta_i} (1 - F_i(t_i))^{1-\delta_i}$$

Notice that the first product does not involve parameters of the survival time distribution and, therefore, can be considered constant. Consequently, the likelihood function is proportional to

$$L \propto \prod_{i=1}^{n} (f_i(t_i))^{\delta_i} (1 - F_i(t_i))^{1-\delta_i} \tag{3.8}$$

Equivalently, the log-likelihood function is proportional to

$$\ln L \propto \sum_{i=1}^{n} \delta_i \ln f_i(t_i) + \sum_{i=1}^{n} (1 - \delta_i)\ \ln(1 - F_i(t_i)) \tag{3.9}$$

Finding the maximum-likelihood estimator of parameters of the lifetime distribution is carried out by equating to zero the derivatives of the log-likelihood function with respect to the parameters. This produces the normal equations, which are then solved to obtain the estimators. Next, the survival function is estimated by $\hat{S}(t) = 1 - \hat{F}(t)$, $t \geq 0$, with the maximum-likelihood estimators of the parameters plugged in.

3.6.3 Exponential Distribution Model

Suppose that the survival time distribution is modeled by the exponential distribution with pdf $f(t) = \lambda \exp\{-\lambda t\}$ and cdf $F(t) = 1 - \exp\{-\lambda t\}$, $t \geq 0$, $\lambda > 0$. By Equation 3.9, the log-likelihood function is proportional to

$$\ln L(\lambda) \propto \ln \lambda \sum_{i=1}^{n} \delta_i - \lambda \sum_{i=1}^{n} \delta_i t_i - \lambda \sum_{i=1}^{n} (1 - \delta_i) t_i = \ln \lambda \sum_{i=1}^{n} \delta_i - \lambda \sum_{i=1}^{n} t_i$$

Equating to zero the derivative with respect to λ produces the equation for the maximum-likelihood estimator $\hat{\lambda}$:

$$\frac{d \ln L(\hat{\lambda})}{d \lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\hat{\lambda}} - \sum_{i=1}^{n} t_i = 0$$

The solution is

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} = \frac{\text{number of deaths}}{\text{number of patient-years}} \tag{3.10}$$

where the number of patient-years is the sum of the survival times for all subjects in the trial. Hence, the estimator for λ can be interpreted as the number of deaths per patient-year. The estimator of the survival function is

$$\hat{S}(t) = \exp\left\{-\hat{\lambda} t\right\}, t \geq 0 \tag{3.11}$$

Example 3.6 Assume that in a model with random censoring the ordered observations are

t_i	1	1	2	3	3	5	8	10	16	18
δ_i	1	1	1	0	1	1	1	1	0	1

The Kaplan–Meier survival curve for these data is shown in Figure 3.7. It resembles the survival function for the exponential distribution, so it is reasonable to fit the exponential model.

By Equation 3.10, the maximum-likelihood estimator of λ is

$$\hat{\lambda} = \frac{8}{1 + \cdots + 18} = \frac{8}{67} = 0.1194$$

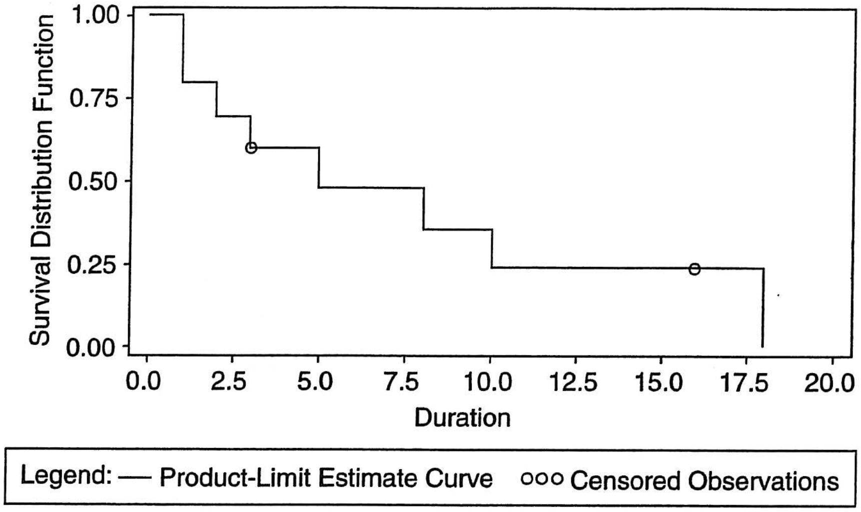

Figure 3.7 Kaplan–Meier survival curve in Example 3.6

Thus, from Equation 3.11, the estimator of the survival function is

$$\hat{S}(t) = \exp\left\{-0.1194\,t\right\}, \quad t \geq 0$$

The SAS code for this example is given in the next section, where a more general model is discussed (see Example 3.9). □

3.6.4 Weibull Distribution Model

Suppose the survival time has the Weibull distribution with the density $f(t) = \alpha\,\lambda\,t^{\alpha-1}\exp\left\{-\lambda\,t^{\alpha}\right\}$, and cdf $F(t) = \exp\left\{-\lambda\,t^{\alpha}\right\}$, where $t \geq 0$, $\alpha, \lambda > 0$. From Equation 3.9, the log-likelihood function is proportional to

$$\begin{aligned}\ln L(\alpha,\lambda) &\propto \ln(\alpha\lambda)\sum_{i=1}^{n}\delta_i + (\alpha-1)\sum_{i=1}^{n}\delta_i \ln t_i - \lambda\sum_{i=1}^{n}\delta_i t_i^{\alpha} - \lambda\sum_{i=1}^{n}(1-\delta_i)t_i^{\alpha}\\ &\propto (\ln\alpha + \ln\lambda)\sum_{i=1}^{n}\delta_i + \alpha\sum_{i=1}^{n}\delta_i\,\ln t_i - \lambda\sum_{i=1}^{n}t_i^{\alpha}\end{aligned}$$

Taking the derivatives with respect to α and λ and setting them equal to zero yields the following system of normal equations:

$$\left\{\begin{aligned}\frac{\partial\,\ln L(\hat{\alpha},\hat{\lambda})}{\partial\alpha} &= \frac{\sum_{i=1}^{n}\delta_i}{\hat{\alpha}} + \sum_{i=1}^{n}\delta_i \ln t_i - \hat{\lambda}\sum_{i=1}^{n}t_i^{\hat{\alpha}}\,\ln t_i = 0\\ \frac{\partial\,\ln L(\hat{\alpha},\hat{\lambda})}{\partial\lambda} &= \frac{\sum_{i=1}^{n}\delta_i}{\hat{\lambda}} - \sum_{i=1}^{n}t_i^{\hat{\alpha}} = 0\end{aligned}\right. \tag{3.12}$$

This system has to be solved numerically. The estimator of the survival function is

$$\hat{S}(t) = \exp\left\{-\,\hat{\lambda}\,t^{\hat{\alpha}}\right\}, \quad t \geq 0 \tag{3.13}$$

Example 3.7 Consider the data in Example 3.2:

t_i	24	16	8	19	10	8	5	17	20	10
δ_i	0	0	1	1	1	0	1	1	1	1

The Kaplan–Meier survival curve is presented in Figures 3.1 and 3.2. The shape of this curve fits the description of the survival function of the Weibull distribution, implying that the Weibull model in this example may be preferable to the exponential one. To find the maximum-likelihood estimators of the parameters α and λ of the Weibull distribution, refer to Equation 3.12. The estimators $\hat{\alpha}$ and $\hat{\lambda}$ solve the system of normal equations

$$\begin{cases} 7/\hat{\alpha} + 17.0674 - \hat{\lambda}\left[24^{\hat{\alpha}}\ln 24 + \cdots + 10^{\hat{\alpha}}\ln 10\right] = 0 \\ 7/\hat{\lambda} - \left[24^{\hat{\alpha}} + \cdots + 10^{\hat{\alpha}}\right] = 0 \end{cases}$$

The numerical solution to these equations is $\hat{\alpha} = 2.2451$ and $\hat{\lambda} = 0.0015$. Thus, according to Equation 3.13, the maximum-likelihood estimator of the survival function is

$$\hat{S}(t) = \exp\left\{-\,0.0015\,t^{2.2451}\right\}, \quad t \geq 0$$

This example will be revisited in the next section (see Example 3.10), where the SAS code and relevant output will be presented and discussed. □

3.7 Regression Model for Survival Time Distribution

Let T be the survival time, and let $x_1, \ldots, x_m$ denote some variables (for example, age, gender, exposure to hazard materials at workplace). A *parametric regression* model for survival time distribution establishes the relationship between the response variable T and the predictor variables $x_1, \ldots, x_m$, which are termed the *covariates*. A parametric regression model is of the form

$$\ln T = \beta_0 + \beta_1\,x_1 + \cdots + \beta_m\,x_m + \sigma\,\varepsilon \tag{3.14}$$

where $\beta_0, \ldots, \beta_m$ are the regression coefficients, σ is a real constant, and ε is the random error.

Two widely used parametric regression models, exponential and Weibull, are discussed here.

3.7.1 Exponential Regression Model

Definition

A parametric regression model for the survival time distribution given in Equation 3.14 is called the *exponential* model, if $\sigma = 1$ and ε has the *extreme-value distribution* with the following density:

$$f_\varepsilon(x) = e^{x-e^x}, \qquad -\infty < x < \infty \tag{3.15}$$

Equivalently, T has the exponential distribution with the following density (see Exercise 3.9):

$$f(t) = \lambda \exp\{-\lambda\, t\}, \quad t \geq 0 \tag{3.16}$$

where $\lambda = \exp\big\{-\left(\beta_0 + \beta_1\, x_1 + \cdots + \beta_m\, x_m\right)\big\}$. The survival function for this distribution is $S(t) = \exp\{-\lambda\, t\}$, $t \geq 0$.

Note that if the covariates are absent in this model, it reduces to the exponential distribution model explored in Subsection 3.6.3.

Estimation of Regression Coefficients

The usual approach to estimation of regression coefficients $\beta_0, \ldots, \beta_m$ is the method of maximum likelihood in the random censoring model defined in Subsection 3.6.2. Suppose that the observations are $(t_i\,, \delta_i\,, x_{i1}\,, \ldots, x_{im})$, $i = 1, \ldots, n$, where t_i has the exponential distribution with the density given by Equation 3.16. Then, in view of Equation 3.9, the log-likelihood function is proportional to

$$\begin{aligned} \ln L(\beta_0, \ldots, \beta_m) \propto & -\sum_{i=1}^{n} \delta_i \left(\beta_0 + \beta_1\, x_{i1} + \cdots + \beta_m\, x_{im}\right) \\ & -\sum_{i=1}^{n} t_i \exp\big\{-\left(\beta_0 + \beta_1\, x_{i1} + \cdots + \beta_m\, x_{im}\right)\big\} \end{aligned} \tag{3.17}$$

To simplify the notation, let $x_{i0} = 1$. Then the normal equations for finding $\hat\beta_0, \ldots, \hat\beta_m$ are

$$\sum_{i=1}^{n} \delta_i\, x_{ij} - \sum_{i=1}^{n} x_{ij}\, t_i \exp\big\{-\left(\hat\beta_0\, x_{i0} + \cdots + \hat\beta_m\, x_{im}\right)\big\} = 0\,, \; j = 0, \ldots, m \tag{3.18}$$

The solution of this system can be found numerically. The maximum-likelihood estimator of λ is $\hat\lambda = \exp\big\{-\left(\hat\beta_0 + \hat\beta_1\, x_1 + \cdots + \hat\beta_m\, x_m\right)\big\}$, and the corresponding estimator of the survival function is $\hat S(t) = \exp\{-\hat\lambda\, t\}$ for $t \geq 0$.

Interpretation of Regression Coefficients

In the exponential model, the regression coefficients $\beta_0, \dots, \beta_m$ have a straightforward meaningful interpretation. According to Equation 3.16, the mean survival time is equal to $\mathbb{E}(T) = 1/\lambda = \exp\left\{ \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \right\}$. Hence, the exponential model yields the following interpretation of the regression coefficients:

- For a numerical covariate x_i, the quantity $\exp\{\beta_i\}$ is the *relative change* [or $100\big(\exp\{\beta_i\} - 1\big)$ % is the *percentage change*] in the mean survival time for each unit increase in the covariate, provided the other covariates are fixed. Indeed, assume that all covariates except x_i are held constant, and index the survival time T by x_i, indicating the dependence of T on this covariate. Then
$$\frac{\mathbb{E}(T_{x_i+1})}{\mathbb{E}(T_{x_i})} = \frac{\exp\left\{ \beta_0 + \cdots + \beta_i\,(x_i + 1) + \cdots \right\}}{\exp\left\{ \beta_0 + \cdots + \beta_i\, x_i + \cdots \right\}} = \exp\left\{ \beta_i \right\} \qquad (3.19)$$

- For a categorical covariate x with l levels, the regression model contains $l - 1$ dummy variables defined as $y_i = 1$, if $x = i$, and 0 otherwise, $i = 1, \dots, l - 1$. Thus the mean survival time is of the form
$$\mathbb{E}(T) = \exp\left\{ \beta_0 + \beta_1\, y_1 + \cdots + \beta_{l-1}\, y_{l-1} + \text{other terms} \right\}$$
Then the ratio of the mean survival times of two subjects, one with x at level i and the other with x at level j $(i,\, j = 1, \dots, l - 1)$, provided the values of all other covariates for these subjects are the same, equals
$$\frac{\mathbb{E}(T_{y_i})}{\mathbb{E}(T_{y_j})} = \frac{\exp\left\{ \beta_0 + \beta_i + \text{constant} \right\}}{\exp\left\{ \beta_0 + \beta_j + \text{constant} \right\}} = \exp\left\{\beta_i - \beta_j\right\} \qquad (3.20)$$
Consequently, $100 \exp\{\beta_i - \beta_j\}$ % is the ratio (expressed as a percentage) of the mean survival times for subjects with the covariate x at level i and at level j, controlling for the other covariates. The ratio of the mean survival times for subjects with x at level l, denoted by $\mathbb{E}(T_0)$, and at level i $(i = 1, \dots, l - 1)$ is found as follows:
$$\frac{\mathbb{E}(T_0)}{\mathbb{E}(T_{y_i})} = \frac{\exp\left\{ \beta_0 + \text{constant} \right\}}{\exp\left\{ \beta_0 + \beta_i + \text{constant} \right\}} = \exp\{-\beta_i\} \qquad (3.21)$$
Hence, $100 \exp\{-\beta_i\}$ % is the ratio of the corresponding mean survival times, expressed as a percentage.

3.7.2 Weibull Regression Model

Definition

A parametric model given by Equation 3.14 is termed the *Weibull regression model* if σ is a constant that has to be estimated from the data, and ε has the extreme-value distribution defined in Equation 3.15. Equivalently, T has the Weibull distribution with the following density (see Exercise 3.10):

$$f(t) = \alpha\,\lambda\,t^{\alpha-1}\,\exp\left\{-\,\lambda\,t^{\alpha}\right\},\; t \geq 0 \tag{3.22}$$

where $\alpha = 1/\sigma$ and $\lambda = \exp\left\{-\,(\beta_0 + \beta_1\,x_1 + \cdots + \beta_m\,x_m)/\sigma\right\}$. The survival function for this distribution is $S(t) = \exp\left\{-\,\lambda\,t^{\alpha}\right\},\; t \geq 0$ (see Exercise 3.2).

A special case of this model when $\sigma = 1$ is the exponential regression model introduced in the previous subsection. Also, the version of Equation 3.22 without the covariates is the Weibull distribution model discussed in Subsection 3.6.4.

Estimation of Regression Coefficients

Assume that $(t_i,\, \delta_i, x_{i1}, \ldots, x_{im})$, $i = 1, \ldots, n$, are the observations in the random censoring model, where t_i has the Weibull distribution with density given in Equation 3.22. From Equation 3.9, the log-likelihood function is proportional to

$$\begin{aligned}\ln L(\beta_0, \ldots, \beta_m, \sigma) \propto &-\ln\sigma \sum_{i=1}^{n} \delta_i - \frac{1}{\sigma}\sum_{i=1}^{n} \delta_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im})\\ &+ \left(\frac{1}{\sigma} - 1\right)\sum_{i=1}^{n} \delta_i \ln t_i\\ &- \sum_{i=1}^{n} \exp\left\{\frac{1}{\sigma}\big(\ln t_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im})\big)\right\}\end{aligned}$$

Note that when $\sigma = 1$, this expression takes a simpler form (Equation 3.17), as it should.

Setting to zero the partial derivatives of the log-likelihood function with respect to the parameters produces a set of normal equations, which should be solved numerically. The maximum-likelihood estimator of λ is $\hat{\lambda} = \exp\left\{-\,(\hat{\beta}_0 + \hat{\beta}_1\,x_1 + \cdots + \hat{\beta}_m\,x_m)/\hat{\sigma}\right\}$, and the survival function is estimated as $\hat{S}(t) = \exp\left\{-\,\hat{\lambda}\,t^{1/\hat{\sigma}}\right\},\; t \geq 0$.

Contrary to the coefficients in the exponential model, the parameters of the Weibull model do not lend themselves to an easy interpretation without requiring additional background. Therefore, the question of the parameter interpretation in the Weibull regression model will be addressed in the next section (see Subsection 3.8.3).

3.7.3 Model Fit Evaluation

A formal likelihood ratio goodness-of-fit test may be conducted with the exponential model as the null hypothesis, and the Weibull model as the alternative hypothesis. Denote by $\ln L(\beta_0, \ldots, \beta_m)$ and $\ln L(\beta_0, \ldots, \beta_m, \sigma)$ the respective log-likelihood functions for the exponential and Weibull models. Then the test statistic equals

$$-2\left(\ln L(\beta_0,\ldots,\beta_m) - \ln L(\beta_0,\ldots,\beta_m,\sigma)\right) \tag{3.23}$$

Under the null hypothesis, this expression has approximately the chi-squared distribution with one degree of freedom.

Unlike what is done in the parametric modeling of survival function (see Section 3.6), plotting the Kaplan–Meier survival curve as a graphical diagnostic tool is not appropriate for the parametric regression model, because the Kaplan–Meier estimator of the survival function ignores the presence of covariates.

3.7.4 Data Examples

Example 3.8 In a 5-year study on subjects who undergo a heart valve replacement surgery, the covariates are x_1, age (in years) of a subject at the time of surgery, and x_2, the preoperational New York Heart Association (NYHA) functional class.* The response variable T measures the time (in years) between the surgery and an event. The event in this case is defined as a death caused by a heart failure or a complication not resulting in expiration—for example, a stroke, thromboembolism (a blood clot in the circulation system), or endocarditis (inflammation of the heart lining and valves). Censoring occurs when a subject drops out of the study or when a subject dies of non-heart-related causes (for example, in an automobile accident). At the end of the study, the surviving subjects are censored as well. The observations taken on 20 subjects appear in Table 3.3.

The covariate x_2 is a categorical variable with four levels; hence three dummy variables must be introduced into the regression model. For convenience, put $y_1 = x_1$, and let

$$y_2 = \begin{cases} 1, & \text{if } x_2 = \text{I} \\ 0, & \text{otherwise} \end{cases}$$

$$y_3 = \begin{cases} 1, & \text{if } x_2 = \text{ II} \\ 0, & \text{otherwise} \end{cases}$$

$$y_4 = \begin{cases} 1, & \text{if } x_2 = \text{ III} \\ 0, & \text{otherwise} \end{cases}$$

*There are four NYHA functional classes (I, II, III, and IV), with the least deteriorated heart being categorized as class I, and the most deteriorated as class IV. This classification is used for prescription of physical activity for cardiac patients.

- Class I: Patients have no limitation of activities.
- Class II: Patients are recommended only mild exertion.
- Class III: Patients are recommended rest.
- Class IV: Patients are recommended complete bed rest.

Table 3.3 Data for Example 3.8

Subject	Age, x_1	NYHA Class, x_2	Survival Time, T	Death Indicator, δ
1	56	II	2.7	0
2	28	III	2.8	1
3	68	I	0.7	0
4	54	III	2.9	1
5	65	IV	0.2	1
6	68	III	0.9	0
7	59	II	3.0	0
8	67	III	0.6	1
9	38	I	3.5	0
10	52	I	2.0	1
11	67	III	0.8	1
12	62	II	0.1	1
13	52	II	4.0	1
14	72	II	0.3	0
15	59	I	1.3	0
16	54	III	4.7	0
17	46	II	1.8	1
18	62	IV	0.2	1
19	73	II	0.3	0
20	47	I	4.6	0

Then the model is

$$\ln T = \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \beta_4 y_4 + \sigma \varepsilon$$

where ε has the extreme-value distribution defined in Equation 3.15.

To see whether the exponential or Weibull distribution model fits the data better, and to find the parameter estimates, resort to procedure `lifereg` in SAS. The code for this example is as follows:

```
data valve_replacement;
input age nyha $ duration status @@;
datalines;
 56   II  2.7  0     28  III  2.8  1     68    I  0.7  0
 54  III  2.9  1     65   IV  0.2  1     68  III  0.9  0
 59   II  3.0  0     67  III  0.6  1     38    I  3.5  0
 52    I  2.0  1     67  III  0.8  1     62   II  0.1  1
 52   II  4.0  1     72   II  0.3  0     59    I  1.3  0
```

```
    54   III   4.7   0      46   II   1.8   1      62   IV   0.2   1
    73    II   0.3   0      47    I   4.6   0
;

proc lifereg data = valve_replacement;
   class nyha; /* list of categorical variables */
   model duration * status(0) = age nyha / dist = exponential;
run;
proc lifereg data = valve_replacement;
   class nyha;
   model duration * status(0) = age nyha / dist = weibull;
run;
```

From SAS computations, the respective values of the log-likelihood function for the two models are $\ln L(\beta_0, \dots, \beta_m) = -19.4946$ and $\ln L(\beta_0, \dots, \beta_m, \sigma) = -18.6317$. Therefore, the goodness-of-fit test statistic, as defined in Equation 3.23, equals

$$-2\left(\ln L(\beta_0, \dots, \beta_m) - \ln L(\beta_0, \dots, \beta_m, \sigma)\right) = 1.7258$$

which has an approximate P-value of $\mathbb{P}\big(\chi^2(1) > 1.7258\big) = 0.1889 > 0.05$. Hence, the exponential model is more appropriate for the given data.

The estimates of the regression coefficients $\beta_0, \dots, \beta_m$ in the exponential model ($\sigma = 1$), along with P-values for testing their equality to zero, are part of SAS output.

	Parameter	Estimate	Pr>ChiSq
$\hat{\beta}_0 \rightarrow$	Intercept	-0.5846	0.8082
$\hat{\beta}_1 \rightarrow$	age	-0.0161	0.6562
$\hat{\beta}_2 \rightarrow$	nyha I	3.8570	0.0039
$\hat{\beta}_3 \rightarrow$	nyha II	2.8764	0.0027
$\hat{\beta}_4 \rightarrow$	nyha III	2.5788	0.0063

The P-value for testing $H_0 : \beta_1 = 0$ is larger than 0.05, so including the variable age in the model is not statistically justified. In fact, it may be wise to rerun the model without this covariate. Replacing the existing line in the SAS code with

```
model duration * status(0) = nyha / dist = exponential;
```

produces the following estimates and the corresponding P-values (all less than 0.05):

	Parameter	Estimate	Pr>ChiSq
$\hat{\beta}_0 \rightarrow$	Intercept	-1.6094	0.0228
$\hat{\beta}_2 \rightarrow$	nyha I	4.1026	0.0008

$\hat{\beta}_3 \rightarrow$ nyha II	3.0123	0.0010
$\hat{\beta}_4 \rightarrow$ nyha III	2.7647	0.0014

Thus, by Equation 3.16, the final model for the survival time distribution is the exponential one with density

$$f(t) = \hat{\lambda}\, \exp\{-\hat{\lambda} t\}\,,\ t \geq 0 \tag{3.24}$$

where $\hat{\lambda} = \exp\big\{\, 1.6094 - 4.1026\, y_2 - 3.0123\, y_3 - 2.7647\, y_4 \,\big\}$. The estimated survival function in this model is $\hat{S}(t) = \exp\big\{ -\, \hat{\lambda}\, t \big\}\,,\ t \geq 0$.

From Equation 3.20 and 3.21, the estimated relative percentage in the mean survival times for subjects in different NYHA classes is computed as follows:

- NYHA class II to class I: $100 \exp\{\, 3.0123 - 4.1026 \,\}\,\% = 33.61\,\%$. That is, the average survival time for subjects in NYHA class II is only 33.61% of that for the subjects in NYHA class I.
- NYHA class III to class I: $100 \exp\{\, 2.7647 - 4.1026 \,\}\,\% = 26.24\,\%$.
- NYHA class IV to class I: $100 \exp\{-4.1026\}\,\% = 1.65\,\%$.
- NYHA class III to class II: $100 \exp\{\, 2.7647 - 3.0123 \,\}\,\% = 76.07\,\%$.
- NYHA class IV to class II: $100 \exp\{-3.0123 \,\}\,\% = 4.92\,\%$.
- NYHA class IV to class III: $100 \exp\{-2.7647 \,\}\,\% = 6.30\,\%$. □

Example 3.9 Consider the data in Example 3.6. The Kaplan–Meier survival curve suggested that the exponential distribution model would have a better fit to the data. To confirm this choice, the goodness-of-fit likelihood ratio test statistic (Equation 3.23) may be computed. The `lifereg` procedure in SAS produces the test statistic as well as the fitted model parameter estimates. Following is the SAS code for this example. Note that because no regression modeling is involved, the list of covariates is empty.

```
data fromExample3_6;
input duration status @@;
datalines;
 1  1    1  1     2  1     3  0     3  1
 5  1    8  1    10  1    16  0    18  1
;

proc lifereg;
   model duration * status(0) = / dist = exponential;
run;

proc lifereg;
   model duration * status(0) = / dist = weibull;
run;
```

SAS outputs the values of the log-likelihood functions for the exponential model, $\ln L(\lambda) = -14.3284$, and for the Weibull model, $\ln L(\alpha, \lambda) = -14.3175$. Consequently, the test statistic is $-2\big(\ln L(\lambda) - \ln L(\alpha, \lambda)\big) = 0.0218$, with the approximate P-value$= \mathbb{P}\big(\chi^2(1) > 0.0218\big) = 0.8826 > 0.05$. Thus, for these data, the exponential model is preferred to the Weibull one.

For the exponential model, SAS estimates the `Intercept` $\hat{\beta}_0 = 2.1253$, which, by Equation 3.16, leads to the estimator $\hat{\lambda} = \exp\{-\hat{\beta}_0\} = 0.1194$. Note that the same value was computed in Example 3.6. □

Example 3.10 In Example 3.7, the Weibull distribution model was selected after studying the Kaplan–Meier survival curve. To confirm the correctness of the model, the `lifereg` procedure may be run. The SAS code is similar to that given in Example 3.9.

From SAS output, $\ln L(\lambda) = -10.7511$ and $\ln L(\alpha, \lambda) = -8.3090$. Thus an approximate P-value for the test statistic is $\mathbb{P}\big(\chi^2(1) > 4.8842\big) = 0.0271 < 0.05$, implying that the Weibull model should be used for these data.

For the Weibull model, SAS gives the estimates of the `Intercept` $\hat{\beta}_0 = 2.8817$ and `Scale` $\hat{\sigma} = 0.4454$. Therefore, by Equation 3.22, the estimators of the distribution parameters are $\hat{\alpha} = 1/0.4454 = 2.2452$ and $\hat{\lambda} = \exp\{-\hat{\beta}_0/\hat{\sigma}\} = 0.0015$. Note that the discrepancy between this estimate of α and the one computed in Example 3.7 is due to the round-off error. □

3.8 Cox Proportional Hazards Model

3.8.1 Standard Definition

In a general regression model, the hazard function h depends on time t and time-dependent covariates $x_1(t), \dots, x_m(t)$. In a simpler model, called the *Cox proportional hazards model* (or *Cox model*, for short), where the covariates $x_1, \dots, x_m$ do not depend on time, the hazard function has the following form:

$$h(t, x_1, \dots, x_m, \beta_1, \dots, \beta_m) = h_0(t) \exp\{\beta_1 x_1 + \cdots + \beta_m x_m\} \qquad (3.25)$$

The function $h_0(t)$ is called the *baseline hazard function.* It is the hazard function of a (usually hypothetical) subject whose covariate values are all zeros (called a *baseline subject*). The quantity $\exp\{\beta_1 x_1 + \cdots + \beta_m x_m\}$ is termed the *relative risk* of a subject with covariates $x_1, \dots, x_m$.

The name of this model—the proportional hazards model—arises from the fact that for any two subjects, the ratio of their hazard functions is a constant not depending on time. Indeed, consider two subjects for which the values of the covariates are $x_{i1}, \dots, x_{im}$ and $x_{j1}, \dots, x_{jm}$, respectively. The ratio of the hazard functions is

$$\frac{h(t, x_{i1}, \dots, x_{im}, \beta_1, \dots, \beta_m)}{h(t, x_{j1}, \dots, x_{jm}, \beta_1, \dots, \beta_m)} = \exp\{\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_m(x_{im} - x_{jm})\}$$

3.8.2 Estimation of Regression Coefficients

The baseline hazard function $h_0(t)$ and the regression coefficients $\beta_1, \ldots, \beta_m$ are the unknowns of the Cox model. An alternative formula for the proportional hazards model will be presented later (see Subsection 3.8.4). It avoids the knowledge of the functional form of $h_0(t)$; thus the only question that should be addressed here is the estimation of betas.

The most commonly used method for estimation of the regression coefficients is the *partial-likelihood estimation* method. In this method, the time-dependent factor of the likelihood function is discarded, and the maximization is carried out on the remaining factor, called the *partial-likelihood function*, producing the *maximum partial-likelihood estimators* of the regression coefficients $\beta_1, \ldots, \beta_m$.

Intuitive Derivation of Partial-Likelihood Function

To see on an intuitive level how this method works, consider a random censoring model (introduced in Subsection 3.6.2) that does not allow tied uncensored observations (times of deaths), but allows ties between an uncensored observation and one or more censored observations. Consider the ith subject with the values of the covariates $x_{i1}, \ldots, x_{im}$. The conditional probability that the subject dies at time t_i, given that t_i is one of the ordered observed survival times, is

$$\frac{\mathbb{P}(\, i\text{th subject dies at } t_i \,)}{\mathbb{P}(\, \text{one death at } t_i \,)} = \frac{\mathbb{P}(\, i\text{th subject dies at } t_i \,)}{\sum_{j \in R(t_i)} \mathbb{P}(\, j\text{th subject dies at } t_i \,)}$$

where $R(t_i)$ denotes the set of all subjects who are at risk at time t_i. Replacing the probability of death at time t_i with the probability of death in the interval $[t_i, t_i + \Delta)$ and passing to the limit as $\Delta \to 0$, yields the following expressions:

$$\begin{aligned}
&\lim_{\Delta \to 0} \frac{\mathbb{P}(\, i\text{th subject dies in } [t_i, t_i + \Delta)\,)/(\Delta\, S(t_i))}{\sum_{j \in R(t_i)} \mathbb{P}(\, j\text{th subject dies in } [t_i, t_i + \Delta)\,)/(\Delta\, S(t_i))} \\
&= \frac{h(\, t_i, x_{i1}, \ldots, x_{im}, \beta_1, \ldots, \beta_m \,)}{\sum_{j \in R(t_i)} h(\, t_i, x_{j1}, \ldots, x_{jm}, \beta_1, \ldots, \beta_m \,)} && \text{[by Equation 3.2]} \\
&= \frac{h_0(t_i) \exp\{\, \beta_1 x_{i1} + \cdots + \beta_m x_{im} \,\}}{\sum_{j \in R(t_i)} h_0(t_i) \exp\{\, \beta_1 x_{j1} + \cdots + \beta_m x_{jm} \,\}} && \text{[by Equation 3.25]} \\
&= \frac{\exp\{\, \beta_1 x_{i1} + \cdots + \beta_m x_{im} \,\}}{\sum_{j \in R(t_i)} \exp\{\, \beta_1 x_{j1} + \cdots + \beta_m x_{jm} \,\}} && (3.26)
\end{aligned}$$

The *partial-likelihood* function $L_p(\, \beta_1, \ldots, \beta_m \,)$ is then defined as the product of the individual conditional probabilities (Equation 3.26), each raised to the power δ that indicates whether the death occurred for this individual:

$$L_p\,(\beta_1,\,\ldots,\,\beta_m\,)=\prod_{i=1}^{n}\left[\frac{\exp\{\,\beta_1\,x_{i1}+\cdots+\beta_m\,x_{im}\,\}}{\sum_{j\,\in\,R(\,t_i)}\,\exp\{\,\beta_1\,x_{j1}+\cdots+\beta_m\,x_{jm}\,\}}\right]^{\delta_i} \tag{3.27}$$

Rigorous Derivation of Partial-Likelihood Function

A rigorous derivation of the expression for the partial-likelihood function (Equation 3.27) is now presented. Consider a random censoring model (see Subsection 3.6.2) without tied uncensored observations, and let the lifetime distribution for the ith subject have pdf $f_i(t_i)$, survival function $S_i(t_i)$, and hazard function $h_i(t_i)=h_0(t)\,\exp\{\beta_1\,x_{i1}+\beta_m\,x_{im}\}$, $i=1,\,\ldots,\,n$. According to Equation 3.8, and the definition of the survival function (Equation 3.1), the likelihood function in this model is proportional to

$$\begin{aligned}
&\prod_{i=1}^{n}(f_i(t_i))^{\delta_i}\,(1-F_i(t_i))^{1-\delta_i}=\prod_{i=1}^{n}(f_i(t_i))^{\delta_i}\,(S_i(t_i))^{1-\delta_i}\\
&=\prod_{i=1}^{n}\,(h_i(t_i)\,S_i(t_i))^{\delta_i}\,(S_i(t_i))^{1-\delta_i}\quad\left[\text{by Exercise 3.1, part (d)}\right]\\
&=\prod_{i=1}^{n}(h_i(t_i))^{\delta_i}\,S_i(t_i)=\prod_{i=1}^{n}\left[\frac{h_i(t_i)}{\sum_{j\,\in\,R(\,t_i)}\,h_j(t_i)}\right]^{\delta_i}\times\Big[\sum_{j\,\in\,R(\,t_i)}\,h_j(t_i)\Big]^{\delta_i}\,S_i(t_i)\\
&=\prod_{i=1}^{n}\left[\frac{\exp\{\beta_1\,x_{i1}+\cdots+\beta_m\,x_{im}\}}{\sum_{j\in R(t_i)}\exp\{\beta_1\,x_{j1}+\cdots+\beta_m\,x_{jm}\}}\right]^{\delta_i}\times\prod_{i=1}^{n}\Big[\sum_{j\in R(t_i)}\,h_j(t_i)\Big]^{\delta_i}\,S_i(t_i)
\end{aligned}$$

Discarding the second product, which depends on time, once again leads to the partial-likelihood function in Equation 3.27.

Partial-Likelihood Score Equations

To find the maximum partial-likelihood estimators of β's, it is convenient to maximize the log-partial-likelihood function, defined as

$$\begin{aligned}
\ln L_p\,(\beta_1,\,\ldots,\,\beta_m\,)&=\sum_{i=1}^{n}\,\delta_i\,(\,\beta_1\,x_{i1}+\cdots+\beta_m\,x_{im}\,)\\
&\quad-\sum_{i=1}^{n}\,\delta_i\,\ln\,(\,\sum_{j\,\in\,R(\,t_i)}\,\exp\{\,\beta_1\,x_{j1}+\cdots+\beta_m\,x_{jm}\,\}\,)
\end{aligned}$$

The estimators $\hat{\beta}_1, \ldots, \hat{\beta}_m$ are found as the numerical solution to the normal equations, called the *partial-likelihood score equations*:

$$\frac{\partial \ln L_p\left(\hat{\beta}_1, \ldots, \hat{\beta}_m\right)}{\partial \beta_k} = \sum_{i=1}^{n} \delta_i \, x_{i\,k} - \sum_{i=1}^{n} \delta_i \frac{\sum_{j \in R(t_i)} x_{j\,k} \exp\left\{ \hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_m x_{jm} \right\}}{\sum_{j \in R(t_i)} \exp\left\{ \hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_m x_{jm} \right\}} = 0, \quad k = 1, \ldots, m. \tag{3.28}$$

Breslow Approximation

Due to the inability to monitor continuously the survival times of subjects, tied event times are observed quite often. Several methods have been proposed for the handling of ties. The *Breslow approximation*, which is used as the default method in SAS, is described here.

Suppose there are k distinct death times $t_1 < t_2 < \cdots < t_k$. Let $D(t_i)$ be the set of subjects whose death times are t_i . Suppose there are d_i subjects in the set $D(t_i)$. The partial-likelihood function for the Breslow approximation is then

$$L_p\left(\beta_1, \ldots, \beta_m\right) = \prod_{i=1}^{k} \left[\frac{\prod_{l \in D(t_i)} \exp\left\{ \beta_1 x_{l1} + \cdots + \beta_m x_{lm} \right\}}{\left(\sum_{j \in R(t_i)} \exp\left\{ \beta_1 x_{j1} + \cdots + \beta_m x_{jm} \right\} \right)^{d_i}} \right]$$

Example 3.11 Suppose five subjects have ordered survival times t_1, t_1, t_2+, t_3, and t_4+, respectively. Notice that there are two distinct death times t_1 and t_3 . Denote by $r_j = \exp\left\{ \beta_1 x_{j1} + \cdots + \beta_m x_{jm} \right\}$, $j = 1, \ldots, 5$, the relative risk of the jth subject. Then the partial-likelihood function for the Breslow approximation is

$$L_p\left(\beta_1, \ldots, \beta_m\right) = \frac{r_1 \, r_2}{\left(r_1 + r_2 + r_3 + r_4 + r_5\right)^2} \, \frac{r_4}{\left(r_4 + r_5\right)}$$

□

The estimates of the regression coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_m$ solve the system of partial-likelihood score equations

$$\frac{\partial \ln L_p(\hat{\beta}_1, \ldots, \hat{\beta}_m)}{\partial \beta_q} = \sum_{i=1}^{k} \sum_{l \in D(t_i)} x_{lq} - \sum_{i=1}^{k} d_i \frac{\sum_{j \in R(t_i)} x_{jq} \exp\{\hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_m x_{jm}\}}{\sum_{j \in R(t_i)} \exp\{\hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_m x_{jm}\}} = 0, \quad q = 1, \ldots, m \tag{3.29}$$

3.8.3 Interpretation of Regression Coefficients

The Cox proportional hazards model yields the following interpretation of the regression coefficients β's that correspond to numerical covariates. The quantity $\exp\{\beta\}$ is the change [equivalently, $100(\exp\{\beta\}-1)$ % is the percentage change] in the hazard function for each unit increase in the covariate, provided the other covariates stay fixed. To see this, write

$$\begin{aligned}&\frac{h(t,\, x_1,\, \ldots,\, x_i+1,\, \ldots,\, x_m,\, \beta_1, \ldots,\, \beta_m)}{h(t,\, x_1,\, \ldots,\, x_i,\, \ldots,\, x_m,\, \beta_1,\, \ldots,\, \beta_m)}\\ &= \frac{h_0(t)\, \exp\{\beta_1\, x_1 + \cdots + \beta_i\,(x_i+1) + \cdots + \beta_m\, x_m\}}{h_0(t)\, \exp\{\beta_1\, x_1 + \cdots + \beta_i\, x_i + \cdots + \beta_m\, x_m\}}\\ &= \exp\{\beta_i\}\end{aligned} \tag{3.30}$$

An interpretation of regression coefficients introduced into the model by a categorical variable is as follows. Suppose the variable has l levels, and let $\beta_1, \ldots, \beta_{l-1}$ denote the regression coefficients in front of the appropriate dummy variables. If no other covariates are present in the model, a subject at the lth level of the covariate is the baseline subject. The quantity $100\exp\{\beta_i-\beta_j\}$ % signifies the ratio (expressed as a percentage) of hazard functions for subjects at level i and at level j of the covariate $(i,\, j = 1, \ldots, l-1)$, provided the other covariates have equal values. The quantity $100\exp\{\beta_i\}$ % is the relative percentage in hazard functions for subjects with the covariate at level i $(i = 1, \ldots,\, l-1)$, and at level l.

Example 3.12 The Weibull regression model introduced in Subsection 3.7.2 is an example of the Cox proportional hazards model. The hazard function is equal to $h(t) = \alpha\,\lambda\, t^{\alpha-1}$, where $\alpha = 1/\sigma$, and $\lambda = \exp\{-(\beta_0 + \beta_1\, x_1 + \cdots + \beta_m\, x_m)/\sigma\}$. This function can be rewritten in the standard form given in Equation 3.25:

$$h(t) = h(t,\, x_1,\, \ldots,\, x_m\,, \beta_1,\, \ldots,\, \beta_m) = h_0(t)\, \exp\{\beta_1^*\, x_1 + \cdots + \beta_m^*\, x_m\}$$

with $h_0(t) = 1/\sigma\, \exp\{-\beta_0/\sigma\}\, t^{1/\sigma-1}$, and $\beta_i^* = -\beta_i/\sigma\,,\ i = 1\,, \ldots,\, m$. Consequently, the regression coefficients in the Weibull model may be interpreted as described previously.

A special case of the Weibull model when $\sigma = 1$—that is, the exponential regression model defined in Subsection 3.7.1—is also a Cox proportional hazards model. As shown in Example 3.1, the hazard function is equal to λ, which is the reciprocal of the mean for this distribution. Thus the quantity $\exp\{-\beta\}$ may be interpreted as the relative change in the hazard function or, equivalently, $\exp\{\beta\}$ may be interpreted as the relative change in the mean survival time, which has been done in Equation 3.19. □

3.8.4 Alternative Form of the Cox Model

Definition and Notation

An alternative form of the Cox proportional hazards model is derived as follows. By part (c) of Exercise 3.1, and by the standard definition of the proportional hazards model (Equations 3.25), the survival function can be written as

$$\begin{aligned} S(t) &= \exp\Big\{ -\int_0^t h(u, x_1, \ldots, x_m, \beta_1, \ldots, \beta_m)\, du \Big\} \\ &= \exp\Big\{ -\int_0^t h_0(u) \exp\{ \beta_1 x_1 + \cdots + \beta_m x_m \}\, du \Big\} \\ &= \big[S_0(t)\big]^r \end{aligned} \tag{3.31}$$

where $r = \exp\{\beta_1 x_1 + \cdots + \beta_m x_m\}$ is the relative risk, and $S_0(t) = \exp\{ -\int_0^t h_0(u)\, du \}$ is the *baseline survival function*, which is the survival function for the baseline subject.

Example 3.13 Refer to Example 3.12. The survival function for the Weibull regression model is

$$S(t) = \exp\{-\lambda t^{\alpha}\}, \quad t \geq 0$$

where $\alpha = 1/\sigma$ and $\lambda = \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)/\sigma\}$. Therefore, the alternative form of the Cox model (Equation 3.31) in this situation is

$$S(t) = \big[S_0(t)\big]^{\exp\{\beta_1^* x_1 + \cdots + \beta_m^* x_m\}} \tag{3.32}$$

with $S_0(t) = \exp\{-\exp\{-\beta_0/\sigma\}\, t^{1/\sigma}\}$ and $\beta_i^* = -\beta_i/\sigma, i = 1, \ldots, m$. In particular, when $\sigma = 1$, the survival function for the exponential regression model may be written as follows:

$$S(t) = \big[S_0(t)\big]^{\exp\{\beta_1^* x_1 + \cdots + \beta_m^* x_m\}} \tag{3.33}$$

with $S_0(t) = \exp\Big\{ -\exp\{-\beta_0\}\, t\Big\}$ and $\beta_i^* = -\beta_i, i = 1, \ldots, m$. □

Note that for the Weibull regression model (and its special case, the exponential model), an explicit algebraic form of the baseline survival function $S_0(t)$ is known. This is not the case for a general proportional hazards model, where $S_0(t)$ has to be estimated nonparametrically based on the observations.

Estimation of Baseline Survival Function

As a default, SAS uses the maximum likelihood approach to estimate the baseline survival function $S_0(t)$. Denote by $\pi_i = \mathbb{P}(T > t_i \,|\, T > t_{i-1})$, the conditional survival probability at time t_i for a baseline subject. The conditional survival probability of a subject with covariates $x_{j1}, \ldots, x_{jm}$ can be obtained by raising π_i to the power $r_j = \exp\left\{ \beta_1 x_{j1} + \cdots + \beta_m x_{jm} \right\}$. Then the contribution to the likelihood function from the subjects who died at time t_i is $1 - \pi_i^{r_j}$, $j \in D(t_i)$; from those who were at risk but did not die, it is $\pi_i^{r_j}$, $j \in R(t_i) \setminus D(t_i)$. Thus the likelihood function is

$$L(\pi_1, \ldots, \pi_n, \beta_1, \ldots, \beta_m) = \prod_{i=1}^{n} \prod_{j \in D(t_i)} (1 - \pi_i^{r_j}) \prod_{j \in R(t_i) \setminus D(t_i)} \pi_i^{r_j} \tag{3.34}$$

Equivalently, the log-likelihood function is equal to

$$\ln L(\pi_1, \ldots, \pi_n, \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} \left[\sum_{j \in D(t_i)} \ln\left(1 - \pi_i^{r_j}\right) + \sum_{j \in R(t_i) \setminus D(t_i)} r_j \ln \pi_i \right] \tag{3.35}$$

This function is maximized with respect to π_i after the partial-likelihood estimators of β's, satisfying Equation 3.28 (or Equation 3.29 if there are ties), are plugged in. It is not difficult to show (see Exercise 3.14) that the normal equations are

$$\sum_{j \in D(t_i)} \frac{\hat{r}_j}{1 - \hat{\pi}_i^{\hat{r}_j}} = \sum_{j \in R(t_i)} \hat{r}_j, \quad i = 1, \ldots, n \tag{3.36}$$

where $\hat{r}_j = \exp\left\{ \hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_m x_{jm} \right\}$.

In the case of no tied values among the observed death times, the size of $D(t_i)$ is 1, and there exists an explicit solution to Equation 3.36:

$$\hat{\pi}_i = \left(1 - \frac{\hat{r}_{j'(i)}}{\sum_{j \in R(t_i)} \hat{r}_j} \right)^{1/\hat{r}_{j'(i)}}, \quad i = 1, \ldots, n$$

where $\hat{r}_{j'(i)}$ denotes the estimated relative risk for the j'th subject whose time of death is t_i—that is, $j' \in D(t_i)$.

In the case of tied observations among the death times, the normal equations (Equation 3.36) should be solved numerically.

The baseline survival function is estimated by a step function

$$\hat{S}_0(t) = \prod_{i\,:\,t_i \leq t} \hat{\pi}_i, \quad t \geq 0$$

Therefore, by Equation 3.31, the estimate of the survival function $S(t)$ for a subject with the values of covariates $x_1, \ldots, x_m$ is

$$\hat{S}(t) = \left[\prod_{i\,:\,t_i \leq t} \hat{\pi}_i \right]^{\exp\left\{ \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m \right\}}, \quad t \geq 0 \tag{3.37}$$

Remark 3.14 The estimator (Equation 3.37) is a generalization of the Kaplan–Meier estimator of the survival function (Equation 3.4). If there are no covariates, then all the relative risks equal 1, and the likelihood function in Equation 3.34 simplifies to the form given in Equation 3.7. □

Data Example

Example 3.15 Consider the data in Example 3.8. The objective of the present example is to fit the alternative form of the Cox proportional hazards model (Equation 3.31) to these data. Before attending to this matter, however, recall that the final model in Example 3.8 was the exponential regression model (Equation 3.24) with the fitted survival function $\hat{S}(t) = \exp\{-\hat{\lambda}\,t\}$, $t \geq 0$, where $\hat{\lambda} = \exp\{1.6094 - 4.1026\,y_2 - 3.0123\,y_3 - 2.7647\,y_4\}$. This function can be rewritten according to Equation 3.33:

$$\hat{S}(t) = \left[\,\hat{S}_0(t)\right]^{\exp\{-4.1026\,y_2 - 3.0123\,y_3\,-2.7647\,y_4\}} \tag{3.38}$$

where $\hat{S}_0(t) = \exp\{-\,t\,e^{1.6094}\} = \exp\{-\,4.9998\,t\}$. The estimator $\hat{S}(t)$ should not differ too much from the one computed using Equation 3.37. The calculations and comparison of the results are presented next.

To request in SAS the estimates of $S_0(t)$ and β values for the Cox model given by Equation 3.31, use the procedure **phreg**. Unlike `lifereg`, **phreg** cannot handle categorical variables; thus the values for the dummy variables y_2, y_3, and y_4 must be input manually. The required SAS code follows:

```
data fromExample3_8;
input y2 y3 y4 duration status @@;
datalines;
 0  1  0  2.7  0     0  0  1  2.8  1
 1  0  0  0.7  0     0  0  1  2.9  1
 0  0  0  0.2  1     0  0  1  0.9  0
 0  1  0  3.0  0     0  0  1  0.6  1
 1  0  0  3.5  0     1  0  1  2.0  1
 0  0  1  0.8  1     0  1  0  0.1  1
 0  1  0  4.0  1     0  1  0  0.3  0
 1  0  0  1.3  0     0  0  1  4.7  0
 0  1  0  1.8  1     0  0  0  0.2  1
 0  1  0  0.3  0     1  0  0  4.6  0
;
proc phreg outest = betas;
/*data=betas contains estimates of betas*/
   model duration * status(0) = y2 y3 y4;
```

```
    baseline out = outdata survival = s;
run;

proc print data = betas;
run;

proc print data = outdata;
run;
```

There are two uncensored tied observations at time 0.2. SAS uses the Breslow approximation to handle the ties; thus the estimators of β solve Equation 3.29. As part of SAS output, the estimated regression coefficients and their P-values (all less than 0.05) are printed:

```
               Parameter
 Variable       Estimate   Pr>ChiSq
 y2             -3.79237     0.0144
 y3             -2.73611     0.0352
 y4             -2.53165     0.0480
```

Note that these estimates do not drastically differ in absolute value from the ones in the exponential model (Equation 3.38).

Further, in the SAS code, the statement **baseline** produces the estimate s of the survival function for a subject who has as values of the covariates their respective sample means. That is, the following function is evaluated:

$$S_{est}(t) = \left[\hat{S}_0(t)\right]^{\exp\left\{\hat{\beta}_1\,\bar{x}_1+\cdots+\hat{\beta}_m\,\bar{x}_m\right\}}$$

where $\bar{x}_1, \ldots, \bar{x}_m$ are the sample means of the respective covariates.

SAS outputs the following table:

```
 y2    y3    y4   duration       s
0.25  0.35  0.3     0.0      1.00000
0.25  0.35  0.3     0.1      0.97716
0.25  0.35  0.3     0.2      0.90434
0.25  0.35  0.3     0.6      0.82978
0.25  0.35  0.3     0.8      0.75196
0.25  0.35  0.3     1.8      0.65984
0.25  0.35  0.3     2.0      0.57237
0.25  0.35  0.3     2.8      0.47476
0.25  0.35  0.3     2.9      0.37416
0.25  0.35  0.3     4.0      0.22072
```

In this table the values for the covariates are all the same: they are equal to the sample means of the respective variables. The column labeled s contains the following estimator:

$$\begin{aligned} S_{est}(t) &= \left[\hat{S}_0(t)\right]^{\exp\left\{-(3.79237)(0.25)-(2.73611)(0.35)-(2.53165)(0.3)\right\}} \\ &= \left[\hat{S}_0(t)\right]^{\exp\left\{-2.66523\right\}} \end{aligned}$$

From here,

$$\hat{S}_0(t) = \left[\hat{S}_{est}(t)\right]^{\exp\left\{2.66523\right\}}$$

Recall that this is a stepwise function. To see whether its overall decrease resembles that of its counterpart $\exp\left\{-4.9998\,t\right\}$ in the exponential model in Equation 3.38, add the following lines to the SAS code:

```
data new;
set outdata;
s_null = s**exp(2.66523); /*two *’s mean exponentiation*/
/*computes baseline survival function for Cox model*/
s_exp = exp(-4.9998 * duration);
/*computes baseline survival function for exponential model*/
run;

proc print data = new;
run;
```

The output contains the following columns:

```
duration   s_null     s_exp
  0.0      1.00000    1.00000
  0.1      0.71741    0.60654
  0.2      0.23573    0.36789
  0.6      0.06846    0.04979
  0.8      0.01663    0.01832
  1.8      0.00254    0.00012
  2.0      0.00033    0.00005
  2.8      0.00002    0.00000
  2.9      0.00000    0.00000
  4.0      0.00000    0.00000
```

As can be seen from the table, the two functions are not far apart from each other at the observed death times, suggesting that their rates of decrease are similar, and, overall, the exponential and Cox models produce similar estimators for the survival function.

Finally, the estimated survival function for any subject in the Cox model is given by

$$\hat{S}(t) = [S_{est}(t)]^{\exp\{2.66523-3.79237\,y_2-2.73611\,y_3-2.53165\,y_4\}}$$

In the Cox model, the fitted regression coefficients yield the following interpretation (see Subsection 3.8.3). The ratios of hazard functions, expressed as a percentage, for two subjects in different NYHA classes are

- NYHA class I to class II: $100\exp\{-3.79237+2.73611\}\,\% = 34.78\,\%$. Thus the hazard function for subjects in NYHA class I is only 34.78% of that for subjects in NYHA class II.
- NYHA class I to class III: $100\exp\{-3.79237+2.53165\}\,\% = 28.34\,\%$.
- NYHA class I to class IV: $100\exp\{-3.79237\}\,\% = 2.25\,\%$.
- NYHA class II to class III: $100\exp\{-2.73611+2.53165\}\,\% = 81.51\,\%$.
- NYHA class II to class IV: $100\exp\{-2.73611\}\,\% = 6.48\,\%$.
- NYHA class III to class IV: $100\exp\{-2.53165\}\,\% = 7.95\,\%$.

□

Exercises for Chapter 3

Section 3.1

Exercise 3.1 Prove that the following formulas are true for the functions $f(t), S(t), h(t)$, and $H(t)$ for any $t \geq 0$:

(a) $h(t) = -S'(t)/S(t)$

(b) $H(t) = -\ln S(t)$

(c) $S(t) = \exp\{-H(t)\} = \exp\Big\{-\int_0^t h(x)\,dx\Big\}$

(d) $f(t) = h(t)\,S(t) = h(t)\exp\{-H(t)\}$

Exercise 3.2 Often survival times are assumed to follow the Weibull distribution with density

$$f(t) = \alpha\,\lambda\,t^{\alpha-1}\exp\{-\lambda t^{\alpha}\}, \quad t \geq 0, \quad \alpha, \lambda > 0$$

Show that the survival, hazard, and cumulative hazard functions for this distribution are, respectively, $S(t) = \exp\{-\lambda t^{\alpha}\}$, $h(t) = \alpha\lambda t^{\alpha-1}$, and $H(t) = \lambda t^{\alpha}$, $t \geq 0$.

Section 3.2

Exercise 3.3 Suppose the data from a clinical trial consist of deaths at 2.1, 2.9, 3.6, 4.5, 5.6, and 6.9 months, and censored observations at 3.0, 3.6, 6.9, and 9.1 months. Find the Kaplan–Meier estimator of the survival function. Perform this calculation both by hand and by using SAS software.

Section 3.3

Exercise 3.4 For the data given in Exercise 3.3, construct the Kaplan–Meier survival curve. Perform this calculation both by hand and by using SAS.

Section 3.4

Exercise 3.5 A clinical study is conducted that tests a new cancer treatment. Nine subjects with advanced-stage liver cancer are randomized into the treatment and control groups. The study continues until all subjects die. The survival times (in years) are recorded. The data are as follows:

Treatment	2.3	3.1	3.2	3.6	3.6
Control	1.2	1.6	2.3	3.1	

(a) Test by hand at the 5% significance level whether the new treatment is effective.

(b) Plot by hand the Kaplan–Meier survival curves for the two groups and comment on their relative positions.

(c) Repeat parts (a) and (b) using SAS software.

Section 3.5

Exercise 3.6 Clinical researchers recorded the number of days until the recurrence of a kidney infection in 45 subjects who use an innovative portable dialysis machine. The censored observations are for the subjects who still had no recurrence of a kidney infection at the time the study terminated. The data are as follows:

15,	11,	22,	121+,	38,	45,	76,	18,	139+,
105+,	51,	44,	10,	111+,	137+,	11,	132,	43,
10,	271+,	77+,	56,	44,	28,	27,	36,	11,
76,	115+,	148+,	43,	56,	179+,	182,	123,	27,
174,	16,	24,	18,	95,	128+,	40,	36,	13

The researchers decide to group the observations into eight time intervals of nonequal lengths: [0, 20), [20, 30), [30, 40), [40, 50), [50, 100), [100, 150),

$[150,\ 200)$, and $[200, 300)$. Construct a survival curve using the actuarial estimation method. Do the work both by hand and by using SAS.

Section 3.6

Exercise 3.7 A new drug is tested in patients with emphysema (a chronic pulmonary disease). The trial is run until all the subjects die. The observations are survival times in years:

$$0.1,\quad 0.1,\quad 0.3,\quad 0.9,\quad 1.0,\quad 1.1,\quad 1.2,\quad 1.3,\quad 2.1,\quad 3.0,\quad 3.6,\quad 5.8$$

Using SAS, draw the Kaplan–Meier survival curve for these data. Explain why the exponential model for lifetime distribution is appropriate. Find by hand the maximum-likelihood estimator of the survival function.

Exercise 3.8 A clinical trial for a new medication is conducted on 12 subjects with advanced-stage pancreatic cancer. The trial is stopped when all subjects die. The data are the survival times in months:

$$1.1,\quad 1.2,\quad 1.3,\quad 1.3,\quad 1.5,\quad 2.0,\quad 2.1,\quad 2.1,\quad 2.2,\quad 3.1,\quad 3.8,\quad 4.1$$

Using SAS draw the Kaplan–Meier survival curve for these data. Explain why the Weibull model for survival time distribution fits the data well. Compute by hand the maximum-likelihood estimator of the survival function.

Section 3.7

Exercise 3.9 For a parametric regression model defined in Equation 3.14, show that if $\sigma = 1$ and the random term ε has the extreme-value distribution with the density given by Equation 3.15, then the survival time T has the exponential distribution with the density shown in Equation 3.16.

Exercise 3.10 In the parametric regression model introduced in Equation 3.14, show that if the random error ε has the extreme-value distribution with density given by Equation 3.15, then the survival time T has the Weibull distribution with the density shown in Equation 3.22.

Exercise 3.11 Consider the data in Exercise 3.7. Use SAS software to obtain the goodness-of-fit test statistic for testing the exponential model versus the Weibull model. Does your result confirm that the exponential model has a better fit? Assume a 0.05 level of significance. Use SAS to estimate the parameter(s) of the model.

Exercise 3.12 Write SAS code for Exercise 3.8. Obtain the goodness-of-fit test statistic, and draw conclusions. Use a significance level of 0.05. Also, estimate the parameter(s) of the chosen model.

Exercise 3.13 A new product for treating recurring ear infections in babies is tested. The covariates are the age of the baby (in months) and the number of previous infections. Each subject starts using the product and continues until an infection occurs. All babies who reach the age of 15 months without recurrence of an ear infection are automatically dropped from the trial. The observations are censored for these subjects. Babies who are younger than 15 months old at the end of the clinical trial and who have not had a recurrence are censored as well. The data on 20 subjects are as follows:

Subject	Age	Number of Infections	Duration	Censored
1	2.0	1	10.1	0
2	2.1	1	10.9	0
3	3.0	4	1.6	0
4	3.1	1	10.1	1
5	3.8	5	0.3	0
6	4.2	3	7.3	1
7	5.1	3	8.2	0
8	5.4	2	8.0	0
9	6.0	1	5.7	0
10	7.0	1	4.9	0
11	7.6	3	2.5	0
12	7.7	3	1.0	0
13	7.8	3	2.8	0
14	8.1	6	1.4	1
15	8.2	2	6.3	0
16	8.5	2	4.0	0
17	9.4	4	1.8	0
18	11.0	2	1.9	1
19	12.5	2	2.5	1
20	13.1	3	1.9	1

(a) Which parametric regression model for the survival time distribution is more appropriate for these data—exponential or Weibull? Conduct the goodness-of-fit test using SAS. Assume a significance level of 5%. Treat the number of infections as a categorical variable with four levels: 1, 2, 3, and 4 or more.

(b) Use SAS to estimate the parameter(s) of the model you chose in part (a). Rerun the model, if necessary, to obtain a reduced model with all significant covariates.

Section 3.8

Exercise 3.14 Derive the normal equations given in Equation 3.36.

Exercise 3.15 Fit the Cox proportional hazards model to the data in Exercise 3.13. Use SAS. Assume the 5% significance level for the tests involved. Are the estimates of the regression coefficients and the baseline survival function similar to the ones obtained earlier? Interpret the estimated regression coefficients.

Exercise 3.16 Seventeen subjects are recruited to a clinical study of a new treatment for cirrhosis of the liver, a chronic disease characterized by the loss of functional liver cells. It is believed that the new treatment may put cirrhosis into remission, if the disease is caused by alcohol abuse. The covariates are the age of a subject (in years) and the indicator of current alcohol abuse (yes $= 1$, no $= 0$). The survival time is the time until a remission (in weeks). A subject who died during the trial is censored. A subject who was awaiting remission at the end of the trial is censored as well. The observations are as follows:

Subject	Age	Alcohol Abuse	Time to Remission	Censored
1	42	1	0.2	0
2	45	0	1.7	0
3	47	0	1.6	0
4	49	1	1.4	0
5	51	0	2.4	0
6	53	0	3.5	0
7	54	1	2.8	0
8	55	1	2.2	1
9	57	0	4.5	0
10	57	0	3.6	0
11	58	0	5.1	0
12	61	1	3.4	0
13	62	0	2.4	0

(continued)

Subject	Age	Alcohol Abuse	Time to Remission	Censored
14	67	1	5.3	0
15	68	1	2.6	1
16	68	1	3.8	0
17	69	1	5.8	0

Fit the Cox proportional hazards model. Assume a 0.05 level of significance. Use SAS. Interpret the estimated regression coefficients. Do the subjects who abuse alcohol have a larger "hazard" of going into remission?

Chapter 4

Introduction to Longitudinal Data Analysis

Often clinical researchers are interested in tracking changes in the measurements that are taken on subjects during follow-up visits. The focus of the present chapter is to introduce several regression models that may be used to fit the data.

4.1 Basic Definitions

Longitudinal data analysis is the analysis of the data obtained through a longitudinal study. A *longitudinal study* collects measurements repeatedly over time on the same subjects. In medical research, the data obtained in this way are called *longitudinal data.* An alternative method of measurement-taking consists of conducting a *cross-sectional study,* in which the measurements are collected on subjects at a single point in time. The advantage of longitudinal studies is that the changes in the measurements are traceable, whereas in cross-sectional studies they are not.

Clinical trials are a special case of longitudinal studies, in which the response measurements are taken several times during follow-up visits. For example, in a clinical trial of a new drug developed to help decrease levels of low-density lipoprotein (LDL—commonly known as bad cholesterol) in the body, periodic measurements of the blood cholesterol level are recorded. Repeated measurements of some other variables, such as amount of daily exercise, body weight, and observing a proper diet, might be taken during the follow-up visits as well.

4.2 Graphical Presentation

Several plots are frequently constructed for visual display of longitudinal data. Example 4.1 illustrates the use of graphical presentation.

Example 4.1 Twenty-four subjects with recurrent malignant gliomas (cancerous brain tumors) are randomized into two treatment groups. Subjects in group 1 receive intravenous chemotherapy, and subjects in group 2 receive concurrent chemo-radiotherapy. The diameter (in centimeters) of the tumor shown via magnetic resonance imaging (MRI) is recorded for each subject at 0-, 3-, 6-, 12-, 18-, and 24-month visits to the clinic. The data are shown in Table 4.1.

Table 4.1 Data for Example 4.1

		Tumor Size					
Subject	Group	0 Month	3 Months	6 Months	12 Months	18 Months	24 Months
1	1	3.1	3.0	2.7	2.3	2.1	1.8
2	1	3.3	2.9	2.4	1.8	1.7	0.2
3	1	2.9	2.4	2.3	2.1	2.1	1.6
4	1	3.2	2.7	2.7	2.2	2.1	1.3
5	1	3.5	3.2	3.2	3.2	3.1	0.8
6	1	3.6	3.5	1.7	1.6	1.5	1.1
7	1	2.2	2.0	2.7	2.4	2.4	1.6
8	1	3.8	3.7	3.0	2.8	2.6	1.0
9	1	3.4	3.1	3.3	4.3	4.1	3.6
10	1	4.6	4.4	2.6	2.5	2.4	1.5
11	1	2.7	2.6	3.1	3.1	3.0	1.6
12	1	4.9	4.7	2.9	2.9	2.8	2.6
13	2	4.0	3.5	3.1	2.0	1.1	1.0
14	2	3.8	3.3	2.7	1.6	1.2	0.6
15	2	3.7	3.2	2.9	2.4	1.5	1.4
16	2	3.2	3.1	2.9	1.8	0.6	0.0
17	2	2.5	2.1	1.5	0.3	0.1	0.0
18	2	2.9	2.2	3.1	1.9	1.7	1.5
19	2	2.8	2.6	2.8	2.6	2.6	2.3
20	2	3.1	2.7	2.3	2.0	1.9	1.3
21	2	4.3	4.2	2.4	2.3	1.2	0.4
22	2	2.9	2.4	2.1	2.1	1.7	1.7
23	2	4.0	3.4	2.3	1.4	0.9	0.0
24	2	2.5	2.4	2.0	1.0	0.3	0.0

To start, one can plot the response variable for each visit, and join the data points related to the same subjects (see Figure 4.1). The lines obtained in this way are called *individual response profiles.* In Figure 4.1, the solid lines correspond to subjects in group 1 and the dashed ones to subjects in group 2. As seen on the graph, all tumors are shrinking (or, rather, not growing) in size as time progresses, except for one outlying individual. Also, most of the individual profiles for group 2 subjects tend to lie lower than those for most of the group 1 subjects.

To make this statement rigorous, the *mean response profile* may be constructed for each group. On this graph, by-visit group means of the response variable are plotted and the points are connected by a straight line (see

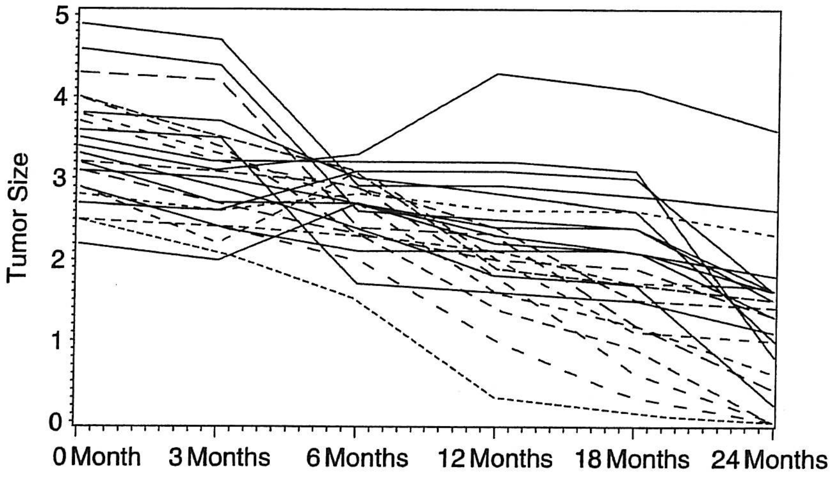

Figure 4.1 Individual response profiles in Example 4.1 (solid lines = group 1, dashed lines = group 2)

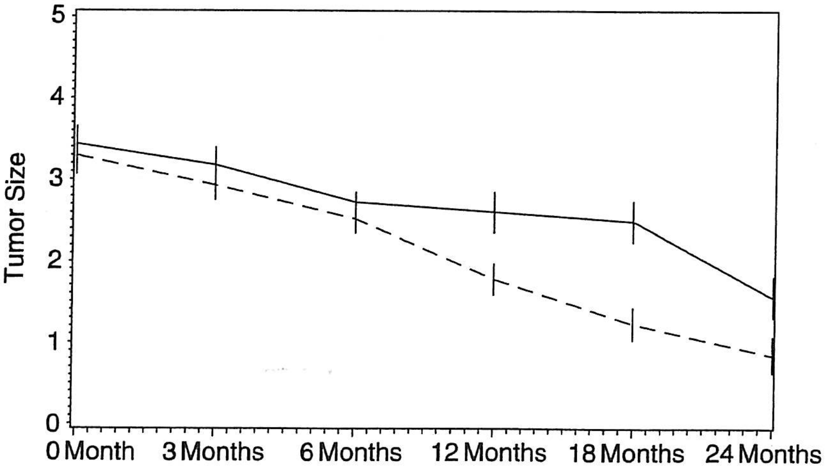

Figure 4.2 Mean response profiles in Example 4.1 (solid line = group 1, dashed line = group 2)

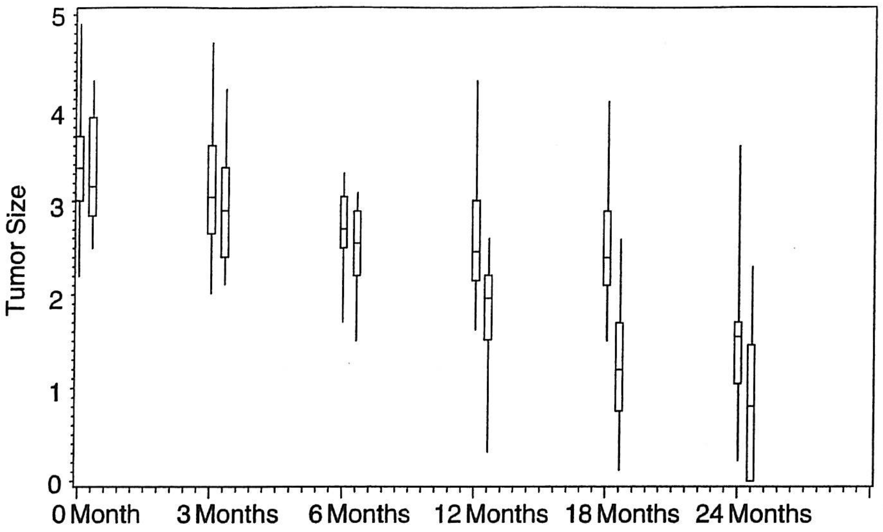

Figure 4.3 Boxplots for two treatment groups in Example 4.1 (left = group 1, right = group 2).

Figure 4.2). As seen in Figure 4.2, the mean tumor size for group 2 subjects at each visit is smaller than that for group 1 subjects. The solid vertical lines depicted on the graph span the interval of sample mean ± one sample standard deviation.

Another option is to draw boxplots for the response variable for each group at each visit. A boxplot is a convenient way to display the five-number summary for a data set: the median, the upper and lower quartiles, and the largest and smallest observations. The graph is given in Figure 4.3. For each visit, the boxplot on the left corresponds to the subjects in group 1, and the one on the right to the subjects in group 2. Note that the boxplots for group 1 tend to lie higher than those for group 2.

Thus, as the visual aids suggest, the second treatment—concurrent chemoradiotherapy—is superior to the first one—chemotherapy.

The SAS code that produces these graphs is as follows:

```
data glioma;
input individual group visit0 visit3
visit6 visit12 visit18 visit24;
cards;
  1   1   3.1   3.0   2.7   2.3   2.1   1.8
  2   1   3.3   2.9   2.4   1.8   1.7   0.2
                      ...
 24   2   2.5   2.4   2.0   1.0   0.3   0.0
;
   data new;
   set glioma;
array x{6} visit0 visit3 visit6 visit12 visit18 visit24;
   do visits = 1 to 6;
   tumorsize = x{visits};  /*creates response vector*/
   if group = 1 then boxposition = visits;
   else boxposition = visits + 0.1;
/*above two lines are needed for boxplots display*/
   output;
   end;
keep individual group visits boxposition tumorsize;
/*now data set = new contains only the kept variables*/
run;

axis1 label = none value = (t=1 ‘0Month’ t=2 ‘3Months’ t=3 ‘6Months’
   t=4 ‘12Months’ t=5 ‘18Months’ t=6 ‘24Months’ t=7 ‘’);

/*individual response profiles*/
proc gplot data = new;
plot tumorsize * visits = individual / nolegend haxis = axis1;
symbol1 interpol = join value = none color = black
line = 1 repeat = 12;
/*these settings are repeated for 12 subjects in group 1*/
symbol2 interpol = join value = none color = black
line = 2 repeat = 12;
/*these settings are repeated for 12 subjects in group 2*/
run;

goptions reset = sym bol;

/*mean response profiles*/
proc gplot data = new;
plot tumorsize * visits = group / nolegend haxis = axis1;
symbol1 interpol = stdm1j color = black line = 1;
symbol2 interpol = stdm1j color = black line = 2;
/*stdm1 = solid vertical bar connects group mean with*/
/*plus/minus one sample standard deviation,*/
/*j = group means are joined across visits*/
run;

goptions reset = sym bol;

/*boxplots for two treatment groups*/
proc gplot data = new;
plot tumorsize * boxposition = group / nolegend haxis = axis1;
```

```
symbol1 interpol = box00 color = black;
/*box00 = vertical lines extend from box to min and max*/
symbol2 interpol = box00 color = black;
run;
```

□

4.3 Random Intercept Model

Longitudinal studies are designed to measure the change in the response variable for individual subjects in relation to a set of *covariates*, the predictor variables. Linear models for longitudinal observations must take into account the *covariance structure* of the data. In the models considered in this section, it is assumed that the data for different subjects are independent—that is, it is assumed that there is a within-subject correlation of the response variable across time, whereas the between-subject correlation is negligible.

A *mixed-effects* model is used to model longitudinal data. In this model, some of the covariates have *random effects* (random levels), while the others have *fixed effects* (fixed levels). In this section, a simple special case of a mixed-effects model called a random intercept model is discussed. In Sections 4.4 and 4.5, some other models with interesting correlation structures of the error terms are considered.

4.3.1 Model Definition and Interpretation of Coefficients

Definition

Suppose longitudinal observations are available on n subjects. The data are collected at fixed times $t_1, \dots, t_k$. Let y_{ij} be the observed response on the ith subject, $i = 1, \dots, n$, at time t_j, $j = 1, \dots, k$, and $x_{1\,ij}, \dots, x_{p\,ij}$ be the observed values of p fixed-effects covariates on the ith subject at time t_j. A *random intercept model* has the form

$$y_{ij} = \beta_0 + \beta_1 x_{1\,ij} + \cdots + \beta_p\, x_{p\,ij} + \beta_{p+1}\, t_j + u_i + \varepsilon_{ij}$$

where $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\, \sigma_u^2)$ are the *random intercepts*, and $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ are the random errors. It is assumed that u_i and ε_{ij} are independent. In this model, the variance of the observations is constant. Indeed, $\mathbb{V}ar(y_{ij}) = \mathbb{V}ar(u_i + \varepsilon_{ij}) = \sigma_u^2 + \sigma^2$. Likewise, the covariance between the repeated observations at times t_j and $t_{j'}$ in the ith subject is constant:

$$\mathbb{C}ov(y_{ij},\, y_{ij'}) = \mathbb{C}ov(u_i + \varepsilon_{ij},\, u_i + \varepsilon_{ij'}) = \mathbb{V}ar(u_i) = \sigma_u^2$$

The observations for different subjects i and i' at times t_j and $t_{j'}$ are uncorrelated because

$$\mathbb{C}ov(y_{ij},\, y_{i'j'}) = \mathbb{C}ov(u_i + \varepsilon_{ij},\, u_{i'} + \varepsilon_{i'j'}) = 0$$

In matrix notation, the random intercept model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}$$

where

$$\underset{nk \times 1}{\mathbf{y}} = \begin{bmatrix} y_{11} \\ \cdots \\ y_{1k} \\ \cdots \\ y_{n1} \\ \cdots \\ y_{nk} \end{bmatrix}, \quad \underset{nk \times (p+2)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{111} & \cdots & x_{p\,11} & t_1 \\ \cdots & \cdots & & & \\ 1 & x_{11k} & \cdots & x_{p\,1k} & t_k \\ \cdots & \cdots & & & \\ 1 & x_{1n1} & \cdots & x_{pn1} & t_1 \\ \cdots & \cdots & & & \\ 1 & x_{1nk} & \cdots & x_{pnk} & t_k \end{bmatrix},$$

$$\underset{(p+2) \times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \cdots \\ \beta_{p+1} \end{bmatrix}, \quad \underset{nk \times 1}{\mathbf{u}} = \begin{bmatrix} u_1 \\ \cdots \\ u_1 \\ \cdots \\ u_n \\ \cdots \\ u_n \end{bmatrix}, \quad \underset{nk \times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_{11} \\ \cdots \\ \varepsilon_{1k} \\ \cdots \\ \varepsilon_{n1} \\ \cdots \\ \varepsilon_{nk} \end{bmatrix}$$

The covariance matrix $\mathbf{V}$ for the response vector $\mathbf{y}$ is a block-diagonal matrix with non-zero $k \times k$ blocks

$$\underset{k \times k}{\mathbf{V}_0} = \begin{bmatrix} \sigma^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma^2 + \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma^2 + \sigma_u^2 \end{bmatrix} = \sigma^2 \mathbf{I}_k + \sigma_u^2 \mathbf{J}_k \tag{4.1}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix, and $\mathbf{J}_k$ denotes the $k \times k$ matrix with all unit entries.

Interpretation of Coefficients

The interpretation of the regression coefficients $\beta_1, \ldots, \beta_{p+1}$ is the same as in the ordinary linear regression model. The mean response $\mathbb{E}(y_{ij}) = \beta_0 + \beta_1\, x_{1\,ij} + \cdots + \beta_p\, x_{p\,ij} + \beta_{p+1}\, t_j$ and, therefore, each coefficient β represents the change in the mean response for a unit increase in the respective variable, provided the other variables are fixed.

For a categorial variable with l levels, the coefficients (say, $\beta_1, \ldots, \beta_{l-1}$) correspond to the dummy variables, which are indicators of levels $1, \ldots, l-1$. Then $\beta_a - \beta_b$, where a, $b = 1, \ldots, l-1$, is interpreted as the difference in mean response for subjects at level a and at level b of the covariate, with the other variables being equal. The coefficient β_a is the difference in mean response for subjects with the covariate at level a and at level l.

4.3.2 Estimation of Parameters

Two methods of model parameter estimation—the maximum-likelihood method and the restricted maximum-likelihood method—are considered in detail in this subsection.

Maximum-Likelihood Method

A usual approach to estimation of the model parameters $\beta_0, \dots, \beta_{p+1}, \sigma^2$, and σ_u^2 is the maximum-likelihood method (ML). The distribution of the $nk \times nk$ response vector $\mathbf{y}$ is a multivariate normal with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}$. The $nk \times nk$ matrix $\mathbf{V}$ is block-diagonal with n blocks $\mathbf{V}_0$ of dimensions $k \times k$. Therefore, the log-likelihood function is proportional to

$$\ln L(\boldsymbol{\beta}, \mathbf{V}_0) \propto -\frac{n}{2} \ln |\mathbf{V}_0| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.2}$$

where $|\mathbf{V}_0|$ denotes the determinant of $\mathbf{V}_0$, and the inverse matrix $\mathbf{V}^{-1}$ is block-diagonal with blocks $\mathbf{V}_0^{-1}$. It can be shown (see Exercise 4.2) that in the random intercept model,

$$|\mathbf{V}_0| = \sigma^{2k} + k\sigma^{2(k-1)}\sigma_u^2 \tag{4.3}$$

$$\mathbf{V}_0^{-1} = \frac{1}{\sigma^4 + k\,\sigma^2\sigma_u^2} \left[(\sigma^2 + k\,\sigma_u^2)\, \mathbf{I}_k - \sigma_u^2\, \mathbf{J}_k \right] \tag{4.4}$$

If $\mathbf{V}_0$ is known, then the maximum-likelihood estimator of $\boldsymbol{\beta}$ is (show it!)

$$\hat{\boldsymbol{\beta}}(\mathbf{V}_0) = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \tag{4.5}$$

Substituting this expression into Equation 4.2 yields

$$\ln L(\hat{\boldsymbol{\beta}}(\mathbf{V}_0), \mathbf{V}_0) \propto -\frac{n}{2} \ln |\mathbf{V}_0| - \frac{1}{2} RSS(\mathbf{V}_0) \tag{4.6}$$

where $RSS(\mathbf{V}_0) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{V}_0))' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{V}_0))$ denotes the *residual sum of squares*. This reduced log-likelihood function depends only on σ^2 and σ_u^2. Maximizing with respect to these variables produces the maximum-likelihood estimators $\hat{\sigma}^2$ and $\hat{\sigma}_u^2$ (and thus $\hat{\mathbf{V}}_0$), which are plugged into Equation 4.5 to obtain $\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}}_0)$.

Restricted Maximum-Likelihood Method

The maximum-likelihood method produces a biased estimator of the variance. An alternative approach is the *restricted maximum-likelihood method* (REML). SAS uses this method as a default.

The REML estimators of σ^2 and σ_u^2 maximize the log-likelihood function of a certain linear transformation of $\mathbf{y}$. The transformation is chosen in such a way that the resulting log-likelihood function, unlike that for the maximum-likelihood method given in Equation 4.2, does not depend on $\boldsymbol{\beta}$. The REML estimators of the variances $\hat{\sigma}^2$ and $\hat{\sigma}_u^2$ maximize this log-likelihood function, and are unbiased estimators of the parameters.

Proposition 4.1 Consider the $nk \times nk$ matrix $\mathbf{I}_{nk} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ that converts $\mathbf{y}$ into the ordinary least-squares residuals. Here $\mathbf{I}_{nk}$ denotes the $nk \times nk$ identity matrix. From the general theory of linear algebra, there exists an $(nk-p-2) \times nk$ matrix $\mathbf{A}$ with the properties $\mathbf{A}'\mathbf{A} = \mathbf{I}_{nk} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ and $\mathbf{A}\mathbf{A}' = \mathbf{I}_{nk-p-2}$, where $\mathbf{I}_{nk-p-2}$ is the $(nk-p-2) \times (nk-p-2)$ identity matrix.

Introduce the REML transformation $\mathbf{Z} = \mathbf{A}\mathbf{y}$. It is a random vector of length $nk-p-2$. Then the corresponding *restricted log-likelihood function* has the form

$$\ln L_r(\mathbf{V}_0) \propto -\frac{n}{2}\ln\left|\mathbf{V}_0\right| - \frac{1}{2}\ln\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right| - \frac{1}{2}RSS(\mathbf{V}_0) \qquad (4.7)$$

where the residual sum of squares $RSS(\mathbf{V}_0) = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{V}_0)\right)'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{V}_0)\right)$ is as in Equation 4.6.

Proof: Consider the weighted least-squares estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{B}\mathbf{y}$$

where $\mathbf{B} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}$ is a $(p+2) \times nk$ matrix. It is a well-known result that $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$. Thus the density of $\hat{\boldsymbol{\beta}}$ is

$$f_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}) = (2\pi)^{-\frac{p+2}{2}}\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right|^{1/2}\exp\left\{-\tfrac{1}{2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\} \qquad (4.8)$$

where $\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right|$ denotes the determinant of the matrix $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$.

Further, using the properties of $\mathbf{A}$, derive that

$$\begin{aligned}\mathbf{A}\mathbf{X} &= \mathbf{I}_{nk-p-2}\mathbf{A}\mathbf{X} = \left(\mathbf{A}\mathbf{A}'\right)\mathbf{A}\mathbf{X} = \mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)\mathbf{X} \\ &= \mathbf{A}\left(\mathbf{I}_{nk} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{X} = \mathbf{A}(\mathbf{X} - \mathbf{X}) = \mathbf{0} \qquad (4.9)\end{aligned}$$

Now recall that the response vector $\mathbf{y}$ has a multivariate normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}$. Because $\mathbf{Z}$ is a linear transformation of $\mathbf{y}$, its distribution is also multivariate normal. The mean of $\mathbf{Z}$ is $\mathbb{E}(\mathbf{Z}) = \mathbb{E}(\mathbf{A}\mathbf{y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, by Equation 4.9. Also, $\mathbf{Z}$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated (and hence independent because they both are normally distributed). Indeed,

$$\begin{aligned}
\mathbb{C}ov(\mathbf{Z},\,\hat{\boldsymbol{\beta}}) &= \mathbb{E}\Big[\mathbf{Z}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\Big] = \mathbb{E}\Big[\mathbf{A}\,\mathbf{y}\,(\mathbf{y}'\,\mathbf{B}'-\boldsymbol{\beta}')\Big]\\
&= \mathbf{A}\,\mathbb{E}(\mathbf{y}\,\mathbf{y}')\,\mathbf{B}' - \mathbf{A}\,\mathbb{E}(\mathbf{y})\,\boldsymbol{\beta}'\\
&= \mathbf{A}\Big[\mathbb{V}ar(\mathbf{y}) + \mathbb{E}(\mathbf{y})\mathbb{E}(\mathbf{y})'\Big]\mathbf{B}' - \mathbf{A}\,\mathbf{X}\,\boldsymbol{\beta}\,\boldsymbol{\beta}'\\
&= \mathbf{A}\,\big[\mathbf{V}+\mathbf{X}\boldsymbol{\beta}\,\boldsymbol{\beta}'\,\mathbf{X}'\big]\,\mathbf{B}' - \mathbf{A}\,\mathbf{X}\,\boldsymbol{\beta}\,\boldsymbol{\beta}'\\
&= \mathbf{A}\,\mathbf{V}\,\mathbf{B}' \qquad \text{(by Equation 4.9)}\\
&= \mathbf{A}\,\mathbf{V}\,\mathbf{V}^{-1}\,\mathbf{X}\,(\mathbf{X}'\mathbf{V}^{-1}\,\mathbf{X})^{-1} = \mathbf{0} \qquad \text{(by Equation 4.9 again)}
\end{aligned}$$

Next, write the introduced transformations in a matrix form

$$\begin{bmatrix}\mathbf{Z}\\ \hat{\boldsymbol{\beta}}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\,\mathbf{y}$$

From linear algebra, the identity equating the respective differentials is

$$d\mathbf{Z}\,d\hat{\boldsymbol{\beta}} \;=\; \big|J(\mathbf{A},\,\mathbf{B})\big|\,d\mathbf{y} \tag{4.10}$$

where $\big|J(\mathbf{A},\,\mathbf{B})\big| \;=\; \left|\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\right|$ denotes the Jacobian determinant that is computed as follows:

$$\begin{aligned}
\left|\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\right| &= \left(\left|\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\right|\,\left|\begin{bmatrix}\mathbf{A}' & \mathbf{B}'\end{bmatrix}\right|\right)^{1/2}\\
&= \left|\begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix}\,\begin{bmatrix}\mathbf{A}' & \mathbf{B}'\end{bmatrix}\right|^{1/2} = \left|\begin{bmatrix}\mathbf{A}\,\mathbf{A}' & \mathbf{A}\,\mathbf{B}'\\ \mathbf{B}\,\mathbf{A}' & \mathbf{B}\,\mathbf{B}'\end{bmatrix}\right|^{1/2}
\end{aligned}$$

Now make a use of the following result on determinants. For any matrices **F**, **G**, and **H**,

$$\left|\begin{bmatrix}\mathbf{F} & \mathbf{G}\\ \mathbf{G}' & \mathbf{H}\end{bmatrix}\right| = \left|\,\mathbf{F}\,\mathbf{H}-\mathbf{F}\,\mathbf{G}'\,\mathbf{F}^{-1}\,\mathbf{G}\,\right|$$

Applying this result and the definitions of **A** and **B**, yields (see Exercise 4.3)

$$\begin{aligned}
\big|J(\mathbf{A},\,\mathbf{B})\big| &= \big|\,(\mathbf{A}\,\mathbf{A}')\,\mathbf{B}\,\mathbf{B}' - (\mathbf{A}\,\mathbf{A}')\,\mathbf{B}\,\mathbf{A}'\,(\mathbf{A}\,\mathbf{A}')^{-1}\,\mathbf{A}\,\mathbf{B}'\,\big|^{1/2}\\
&= \big|\mathbf{X}'\,\mathbf{X}\big|^{-1/2}
\end{aligned} \tag{4.11}$$

Finally, it remains to express the density of **Z** in terms of **y**. Write

$$\begin{aligned}
f(\mathbf{Z},\,\hat{\boldsymbol{\beta}})\;d\mathbf{Z}\,d\hat{\boldsymbol{\beta}} &= f_{\mathbf{Z}}(\mathbf{Z})\;f_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}})\;d\mathbf{Z}\,d\hat{\boldsymbol{\beta}} \qquad \text{(by independence)}\\
&= f_{\mathbf{y}}(\mathbf{y})\;\big|\mathbf{X}'\,\mathbf{X}\,\big|^{-1/2}\,d\mathbf{y} \qquad \text{(by Equations 4.10 and 4.11)}
\end{aligned}$$

Consequently, using Equation 4.8, obtain

$$
\begin{aligned}
f_{\mathbf{Z}}(\mathbf{Z}) &= \left|\mathbf{X}'\,\mathbf{X}\right|^{-1/2} \frac{f_{\mathbf{y}}(\mathbf{y})}{f_{\hat{\beta}}(\hat{\beta})} \\
&= (2\pi)^{-\frac{(n\,k-p-2)}{2}} \left|\mathbf{X}'\,\mathbf{X}\right|^{-1/2} \left|\mathbf{V}\right|^{-1/2} \left|\mathbf{X}'\,\mathbf{V}^{-1}\,\mathbf{X}\right|^{-1/2} \\
&\quad \times \exp\Big\{ -\tfrac{1}{2} \Big[(\mathbf{y}-\mathbf{X}\,\beta)'\,\mathbf{V}^{-1}\,(\mathbf{y}-\mathbf{X}\,\beta) \\
&\qquad\qquad - (\hat{\beta}-\beta)'\,(\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{X})\,(\hat{\beta}-\beta)\Big]\Big\} \\
&\propto \left|\mathbf{X}'\,\mathbf{X}\right|^{-1/2} \left|\mathbf{V}\right|^{-1/2} \left|\mathbf{X}'\,\mathbf{V}^{-1}\,\mathbf{X}\right|^{-1/2} \\
&\quad \times \exp\Big\{ -\tfrac{1}{2}\,(\mathbf{y}-\mathbf{X}\,\hat{\beta})'\,\mathbf{V}^{-1}\,(\mathbf{y}-\mathbf{X}\hat{\beta})\Big\}
\end{aligned}
$$

The expression for the restricted log-likelihood function in Equation 4.7 follows. □

4.3.3 Data Example

Example 4.3 Consider the data in Example 4.1. The random intercept model may be appropriate for these data for the following reasons. As seen in Figure 4.1, the scatter of glioma diameter does not increase drastically over time. It may, therefore, be assumed that the response varies constantly over time. Also, for each subject, the historical tumor diameters may be equally influential on a current diameter. It may, therefore, be assumed that the constant covariance assumption holds.

Given that the subjects in group 1 (chemotherapy patients) are to be compared to the subjects in group 2 (chemo-radiotherapy patients), the variable `group` should be used as a covariate in this model. Denote the individual values of this covariate by x_i. Note that $x_i = 1$ if the ith subject is in group 1 and that $x_i = 2$ if the ith subject is in group 2. Let y_{ij} be the tumor size of the ith subject at visit time t_j, where $t_1 = 0$, $t_2 = 3$, $t_3 = 6$, $t_4 = 12$, $t_5 = 18$, and $t_6 = 24$ (all in months). Then the model is

$$y_{ij} = \beta_0 + \beta_1\,x_i + \beta_2\,t_j + u_i + \varepsilon_{ij}, \quad i = 1, \dots, 24, \quad j = 1, \dots, 6$$

where the random intercepts $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2)$ are independent of the random errors $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

In SAS, procedure `mixed` is used to compute the estimates of the model parameters. The ML and REML methods are applied. The relevant fragments of the SAS output are shown here:

```
                    Effect       Estimate    Pr > |t|
$\hat{\beta}_0 \rightarrow$      Intercept     4.1240     <.0001
$\hat{\beta}_1 \rightarrow$      group        -0.5625     0.0038   ML method
                                              0.0055   REML method
$\hat{\beta}_2 \rightarrow$      time         -0.0855     <.0001
```

```
                         Covariance
                         Parameter     Estimate
          ML method      Intercept       0.1636   $\leftarrow \hat{\sigma}_u^2$
                         Residual        0.3237   $\leftarrow \hat{\sigma}^2$

        REML method      Intercept       0.1829   $\leftarrow \hat{\sigma}_u^2$
                         Residual        0.3264   $\leftarrow \hat{\sigma}^2$
```

The estimators of the model coefficients are statistically significant at the 1% significance level.

Notice that the estimator of β_1 is negative. This confirms the conclusion from Figures 4.1, 4.2, and 4.3 that the second treatment (chemo-radiotherapy) is superior to the first one (chemotherapy). In fact, at every fixed visit time, the average tumor size for subjects in group 2 is estimated to be 0.5625 cm smaller than that for subjects in group 1. Also, for either group, an average monthly decrease in tumor diameter is estimated as 0.0855 cm.

The required SAS code follows:

```
data glioma;
input individual group visit0 visit3
visit6 visit12 visit18 visit24;
cards;
  1   1   3.1   3.0   2.7   2.3   2.1   1.8
  2   1   3.3   2.9   2.4   1.8   1.7   0.2
                      ...
 24   2   2.5   2.4   2.0   1.0   0.3   0.0
;
data new;
set glioma;
array x{6} visit0 visit3 visit6 visit12 visit18 visit24;
array t{6} t1-t6 (0 3 6 12 18 24);
do visits = 1 to 6;
tumorsize = x{visits};
time = t{visits};
```

```
output;
end;
keep individual group tumorsize time;
run;

/*Maximum-Likelihood Method*/
proc mixed data = new method = ml;
   model tumorsize = group time / solution;
/*'solution' option requests estimates of beta parameters*/
   random intercept / subject = individual;
run;

/*Restricted Maximum-Likelihood Method*/
proc mixed data = new method = reml;
   model tumorsize = group time / solution;
   random intercept / subject = individual;
run;
```

□

4.4 Random Slope and Intercept Model

The random intercept model introduced in Section 4.3 backs up the assumptions that variances and covariances remain constant over time. In practice, however, these assumptions rarely hold. A more realistic model that allows for heterogeneity is called the *random slope and intercept model.* In this model, y_{ij}, the observation of the response variable on the ith subject at time t_j, $i = 1, \dots, n$, $j = 1, \dots, k$, is of the following form:

$$y_{ij} = \beta_0 + \beta_1\, x_{1\,ij} + \cdots + \beta_p\, x_{p\,ij} + \beta_{p+1}\, t_j + u_{i1} + u_{i2}\, t_j + \varepsilon_{ij}$$

where $u_{i1} \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \sigma^2_{u_1})$ are the *random intercepts*, and $u_{i2} \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \sigma^2_{u_2})$ are the *random slopes.* Also, $\mathbb{C}ov\big(u_{i1},\ u_{i2}\big) = \sigma_{u_1\,u_2}$, $i = 1, \dots, n$, and $\mathbb{C}ov\big(u_{i1},\ u_{i'\,2}\big) = 0$ for $i \neq i'$. The error terms $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ are independent of u_1 and u_2. In this model

$$\mathbb{V}ar(y_{ij}) = \mathbb{V}ar\big(\, u_{i1} + u_{i2}\, t_j + \varepsilon_{ij}\big) = \sigma^2_{u_1} + 2\,\sigma_{u_1 u_2}\, t_j + \sigma^2_{u_2}\, t_j^2 + \sigma^2 \tag{4.12}$$

and

$$\begin{aligned}
\mathbb{C}ov\big(y_{ij},\ y_{ij'}\big) &= \mathbb{C}ov\big(\, u_{i1} + u_{i2}\, t_j + \varepsilon_{ij},\ u_{i1} + u_{i2}\, t_{j'} + \varepsilon_{ij'}\big) \\
&= \sigma^2_{u_1} + \sigma_{u_1\,u_2}\,\big(t_j + t_{j'}\big) + \sigma^2_{u_2}\, t_j\, t_{j'} \quad \text{for } j \neq j' \qquad (4.13) \\
\mathbb{C}ov(y_{ij}, y_{i'j'}) &= \mathbb{C}ov(u_{i1} + u_{i2}t_j + \varepsilon_{ij}, u_{i'1} + u_{i'2}t_{j'} + \varepsilon_{i'j'}) = 0 \quad \text{for } i \neq i'
\end{aligned}$$

Notice that now the variances and the covariances depend on time. The covariance matrix $\mathbf{V}$ is still block-diagonal, but the block matrix $\mathbf{V}_0$ has a nonsymmetric structure as opposed to that in Equation 4.1. The log-likelihood function in Equation 4.2 and the restricted log-likelihood function in Equation 4.7 are still valid, and the two methods of parameter estimation are applicable.

In this model, the regression coefficients are interpreted the same way as in the random intercept model (see Section 4.3).

Example 4.4 Often clinical trials include substudies focusing on the *health-related quality of life* for subjects. Even though a treatment prolongs the lifetime of a subject, it might reduce the quality of that life—for example, by causing depression, limitations in mobility, and difficulties with everyday activities. During follow-up visits to the clinic, subjects fill out therapy satisfaction questionnaires. Scores on these questionnaires are recorded, and the response variable is the percentage satisfaction with the therapy. Suppose that for the subjects from Example 4.1 the data are as shown in Table 4.2 on page 90.

The individual response profiles are plotted in Figure 4.4. The mean response profiles are displayed in Figure 4.5, and the boxplots are shown in Figure 4.6.

Recall that the conclusion regarding the two treatments—chemotherapy and chemo-radiotherapy—was that the chemo-radiotherapy, on average, reduces the tumor size faster than the chemotherapy treatment; thus chemo-radiotherapy was deemed the superior treatment. However, as seen in Figures 4.4, 4.5, and 4.6, the therapy satisfaction scores for subjects who receive the chemo-radiotherapy (group 2) are, on average, lower than those for subjects undergoing chemotherapy (group 1). The point illustrated here is that even though a treatment may be more powerful in fighting a disease, it may lead to a lower quality of life of subjects.

For the data set in Table 4.2, the variance of the response is not constant over time; that is, the observations become more scattered toward the end of

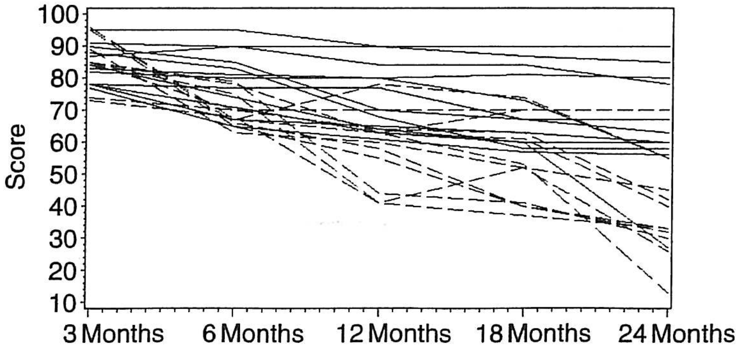

Figure 4.4 Individual response profiles in Example 4.4 (solid lines = group 1, dashed lines = group 2)

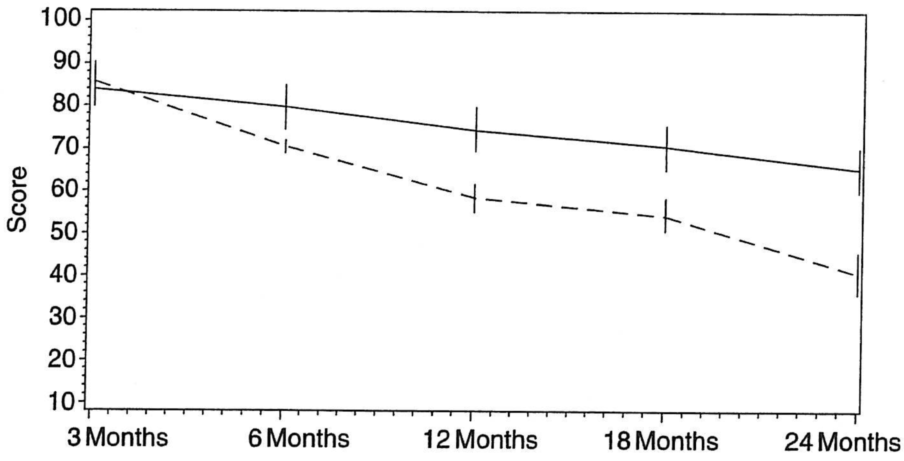

Figure 4.5 Mean response profiles in Example 4.4 (solid line = group 1, dashed line = group 2)

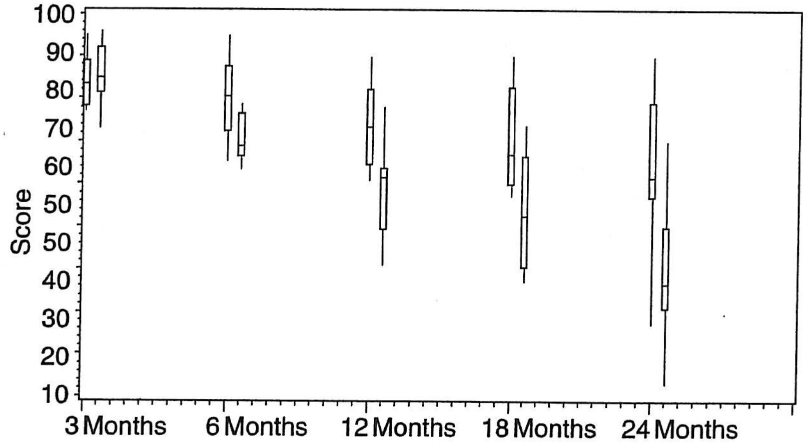

Figure 4.6 Boxplots for two treatment groups in Example 4.4 (left = group 1, right = group 2)

the study. Thus the random slope and intercept model will be fitted to these data. The model is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 t_j + u_{i1} + u_{i2} t_j + \varepsilon_{ij}$$

where y_{ij} is the therapy satisfaction score for the ith subject at time t_j, $i = 1, \ldots, 24$, $j = 1, \ldots, 5$; $x_{1ij} = x_{1i}$ is the group number of the ith subject (the same for all times t_j); x_{2ij} is the tumor size of the ith subject at time t_j; and $t_1 = 3$, $t_2 = 6$, $t_3 = 12$, $t_4 = 18$, and $t_5 = 24$ (all in months). In addition to the regression coefficients, the other parameters of the model are $\mathbb{V}ar(u_{i1}) = \sigma^2_{u_1}$, $\mathbb{V}ar(u_{i2}) = \sigma^2_{u_2}$, $\mathbb{V}ar(\varepsilon_{ij}) = \sigma^2$, and $\mathbb{C}ov(u_{i1}, u_{i2}) = \sigma_{u_1 u_2}$.

To estimate the unknown parameters in SAS, use the following code. It assumes that the data set `glioma` has been defined earlier. For the sake of brevity, only the restricted maximum-likelihood estimation method is applied here.

Table 4.2 Data for Example 4.4

		Therapy Satisfaction Score				
Subject	Group	3 Months	6 Months	12 Months	18 Months	24 Months
1	1	90	85	70	67	63
2	1	87	90	95	90	90
3	1	78	67	65	63	60
4	1	77	65	61	57	56
5	1	78	77	77	67	67
6	1	82	80	80	73	55
7	1	88	83	68	58	58
8	1	95	95	90	87	85
9	1	84	74	64	60	27
10	1	78	71	63	60	60
11	1	91	90	84	84	78
12	1	83	81	80	81	80
13	2	78	67	63	70	70
14	2	85	78	63	61	40
15	2	85	70	70	70	70
16	2	85	78	63	53	13
17	2	85	79	44	41	30
18	2	96	66	58	40	32
19	2	89	63	60	52	45
20	2	95	65	55	40	33
21	2	73	68	41	52	26
22	2	85	75	41	37	33
23	2	74	70	64	63	42
24	2	96	67	78	74	55

```
data scores;
input individual group mos3 mos6 mos12 mos18 mos24;
cards;
 1   1   90   85   70   67   63
 2   1   87   90   95   90   90
              ...
24   2   96   67   78   74   55
 ;
```

```
data combined;
merge glioma scores;
by individual;
run;

data new;
set combined;
array x{5} visit3 visit6 visit12 visit18 visit24;
array z{5} mos3 mos6 mos12 mos18 mos24;
array t{5} (3 6 12 18 24);
do visits = 1 to 5;
tumorsize = x{visits};
time = t{visits};
score = z{visits};
output;
end;
keep individual group tumorsize time score;
run;

proc mixed data = new method = reml;
model score = group tumorsize time / solution;
random intercept time / subject = individual type = un;
/*option un = ‘unstructured’ requests covariance estimation*/
run;
```

The estimated regression coefficients are as follows:

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	96.7818	<.0001
$\hat{\beta}_1 \rightarrow$	group	-5.6229	0.0370
$\hat{\beta}_2 \rightarrow$	tumorsize	-0.7679	0.6028
$\hat{\beta}_3 \rightarrow$	time	-1.4546	<.0001

The P-value for the estimator $\hat{\beta}_2$ is larger than 0.05, indicating that the variable `tumorsize` has an insignificant effect on the level of satisfaction with the therapy. Removing this covariate from the model and reestimating the parameters using the REML method yields the following results:

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	93.9689	<.0001
$\hat{\beta}_1 \rightarrow$	group	-5.3926	0.0398
$\hat{\beta}_3 \rightarrow$	time	-1.3919	<.0001

```
Covariance
Parameter    Estimate
UN(1,1)       31.3316   ← $\hat{\sigma}^2_{u_1}$
UN(2,1)       -2.8514   ← $\hat{\sigma}_{u_1 u_2}$
UN(2,2)        0.7508   ← $\hat{\sigma}^2_{u_2}$
Residual      46.2082   ← $\hat{\sigma}^2$
```

In the reduced model, the estimates of the β values are all significant. The estimated coefficient in front of the covariate group is negative. This proves that chemo-radiotherapy subjects have a lower quality of life compared to that of the members of the radiotherapy group. In fact, the satisfaction score in the chemo-radiotherapy group at any given time is roughly 5.39 points lower than that in the other group.

For both groups, the satisfaction score decreases, on average, by about 1.39 points every month. □

4.5 Model with Spatial Power Covariance Structure for Error

Other useful models for longitudinal data include a mixed-effects model with *spatial power covariance structure for error.* The observation y_{ij} of the response variable on the ith subject, $i = 1, \dots, n$, at time t_j, $j = 1, \dots, k$, is modeled as follows:

$$y_{ij} = \beta_0 + \beta_1 \, x_{1ij} + \cdots + \beta_p \, x_{pij} + \beta_{p+1} \, t_j + u_{i1} + u_{i2} \, t_j + w_i(t_j)$$

where $x_{1ij}, \dots, x_{pij}$ are the *fixed-effects* covariates observed on the ith subject at time t_j, and u_{i1} and u_{i2} are the *random-effects* terms. It is assumed that $u_{i1} \overset{i.i.d.}{\sim} \mathcal{N}(0,\, \sigma^2_{u_1})$, $u_{i2} \overset{i.i.d.}{\sim} \mathcal{N}(0,\, \sigma^2_{u_2})$, $\mathbb{C}ov\big(u_{i1},\, u_{i2}\big) = \sigma_{u_1 u_2}$, and $\mathbb{C}ov\big(u_{i1},\, u_{i'2}\big) = 0$ if $i \neq i'$. The component w_i is a discrete-time process with a constant mean and constant variance. This process satisfies the recursive formula

$$w_i(t_j) = \rho \, w_i(t_j - 1) + z_i(t_j) \tag{4.14}$$

Here ρ, $\big|\rho\big| < 1$, is a fixed number, and $z_i(t_j) \overset{i.i.d.}{\sim} \mathcal{N}\big(0,\, (1 - \rho^2)\, \sigma^2\big)$ are independent of $w_i(t_1)$, the initial observation on the ith subject. It follows that for any $j' > j$, $w_i(t_j)$ and $z_i(t_{j'})$ are independent. Indeed, by Equation 4.14,

$$\begin{aligned} w_i(t_j) &= \rho \, \Big(\rho \, w_i(t_j - 2) + z_i(t_j - 1) \Big) + z_i(t_j) \\ &= \rho^2 \, \Big(\rho \, w_i(t_j - 3) + z_i(t_j - 2) \Big) + \rho \, z_i(t_j - 1) + z_i(t_j) \\ &= \cdots = \rho^{t_j - t_1} \, w_i(t_1) + \rho^{t_j - t_i - 1} \, z_i(t_1 + 1) + \cdots + \rho \, z_i(t_j - 1) + z_i(t_j) \end{aligned} \tag{4.15}$$

and every term in this sum is independent of $z_i(t_{j'})$.

It can be shown (see Exercise 4.6) that the process w_i has mean zero and variance σ^2; that is, $\mathbb{E}\big(w_i(t_j)\big) = 0$ and $\mathbb{E}\big(w_i(t_j)\big)^2 = \sigma^2$. To compute the covariance matrix of this process, note that $\mathbb{C}ov\big(w_i(t_j),\, w_i(t_{j'})\big) = \mathbb{E}\big(w_i(t_j)\, w_i(t_{j'})\big)$ for any $j' > j$. As in Equation 4.15, $w_i(t_{j'})$ can be written as

$$w_i(t_{j'}) = \rho^{\,t_{j'}-t_j}\; w_i(t_j) + \rho^{\,t_{j'}-t_j-1}\; z_i(t_j+1) + \cdots + \rho\; z_i(t_{j'}-1) + z_i(t_{j'}) \quad (4.16)$$

By independence of $w_i(t_j)$ and $z_i(t_{j''})$ for any $j'' > j$, $\mathbb{E}\big(w_i(t_j)\, z_i(t_{j''})\big) = \mathbb{E}\big(w_i(t_j)\big)\, \mathbb{E}\big(z_i(t_{j''})\big) = 0$. Therefore, in view of Equation 4.16, the covariance between $w_i(t_j)$ and $w_i(t_{j'})$ is

$$\mathbb{E}\Big(w_i(t_j)\, w_i(t_{j'})\Big) = \rho^{\,t_{j'}-t_j}\; \mathbb{E}\big(w_i(t_j)\big)^2 = \rho^{\,t_{j'}-t_j}\, \sigma^2$$

Consequently, the covariance matrix of the error terms is an $nk \times nk$ block-diagonal matrix with $k \times k$ blocks

$$\sigma^2 \begin{bmatrix} 1 & \rho^{\,t_2-t_1} & \rho^{\,t_3-t_1} & \ldots & \rho^{\,t_k-t_1} \\ \rho^{\,t_2-t_1} & 1 & \rho^{\,t_3-t_2} & \ldots & \rho^{\,t_k-t_2} \\ & & \ldots & & \\ \rho^{\,t_k-t_1} & \rho^{\,t_k-t_2} & \rho^{\,t_k-t_3} & \ldots & 1 \end{bmatrix} \quad (4.17)$$

A matrix of this form is called a *spatial power matrix*, and the corresponding process is said to have a *spatial power covariance structure*.

The covariance matrix $\mathbf{V}$ of the response variable is an $nk \times nk$ block-diagonal matrix, in which the $k \times k$ blocks $\mathbf{V}_0$ have diagonal entries (compare them to Equation 4.12) of the following form:

$$\mathbb{V}ar(y_{ij}) = \mathbb{V}ar\big(\,u_{i1} + u_{i2}\, t_j \,+\, w_i(t_j)\,\big) = \sigma^2_{u_1} + 2\,\sigma_{u_1\,u_2}\, t_j + \sigma^2_{u_2}\, t_j^2 + \sigma^2$$

where $j = 1, \ldots, k$, and the off-diagonal entries of $\mathbf{V}_0$ are (cf. Equation 4.13)

$$\begin{aligned} \mathbb{C}ov\big(y_{ij},\, y_{ij'}\big) &= \mathbb{C}ov\big(u_{i1} + u_{i2}\, t_j + w_i(t_j),\; u_{i1} + u_{i2}\, t_{j'} + w_i(t_{j'})\,\big) \\ &= \sigma^2_{u_1} + \sigma_{u_1\,u_2}\, \big(t_j + t_{j'}\big) + \sigma^2_{u_2}\, t_j\, t_{j'} + \rho^{\,t_{j'}-t_j}\, \sigma^2 \end{aligned}$$

As in the random slope and intercept model, the variance and covariance of the response variable depend on time. However, unlike in the models discussed earlier, where no correlation between the error terms was assumed, in the model with spatial power covariance structure for error, a weak dependence between error terms for the same individual is present. In absolute value, the covariance between the error terms decays exponentially fast as the time span increases.

For this model, the validity of the ML and REML estimation methods introduced in Subsection 4.3.2 holds. The regression coefficients admit the interpretation identical to that in the random intercept model introduced in Section 4.3.

Example 4.5 Consider the data in Example 4.4. The objective is to fit the mixed-effects model with the spatial power covariance structure of the error terms. The model is of the form

$$y_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 t_j + u_{i1} + u_{i2} t_j + w_i(t_j)$$

where y_{ij} denotes the therapy satisfaction score for the ith subject at time t_j; x_{1i} is the group (1 or 2) of the ith subject; x_{2ij} is the tumor size of the ith subject on the jth occasion; $t_1 = 3$, $t_2 = 6$, $t_3 = 12$, $t_4 = 18$, $t_5 = 24$ (all in months); $i = 1, \dots, 24$, and $j = 1, \dots, 5$. The random variables u_{i1} and u_{i2} are the random intercept and the slope, respectively. The error terms $w_i(t_j)$ have the zero mean and a 120 × 120 block-diagonal covariance matrix with 24 5×5 blocks:

$$\sigma^2 \begin{bmatrix} 1 & \rho^3 & \rho^9 & \rho^{15} & \rho^{21} \\ \rho^3 & 1 & \rho^6 & \rho^{12} & \rho^{18} \\ \rho^9 & \rho^6 & 1 & \rho^6 & \rho^{12} \\ \rho^{15} & \rho^{12} & \rho^6 & 1 & \rho^6 \\ \rho^{21} & \rho^{18} & \rho^{12} & \rho^6 & 1 \end{bmatrix}$$

The parameters of this model are β_0, β_1, β_2, β_3, $\sigma^2_{u_1}$, $\sigma^2_{u_2}$, $\sigma_{u_1 u_2}$, σ^2, and ρ. The REML parameter estimation method is applied. The SAS code for data `new` defined in Example 4.4 is as follows:

```
proc mixed data = new method = reml;
model score = group tumorsize time / solution;
random intercept time / subject = individual type = un;
repeated / subject = individual type = sp(pow)(time);
/*option sp(pow)(time) requests estimation of rho */
run;
```

The estimated regression coefficients are

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \longrightarrow$	Intercept	97.4115	<.0001
$\hat{\beta}_1 \longrightarrow$	group	-5.2655	0.0410
$\hat{\beta}_2 \longrightarrow$	tumorsize	-1.0294	0.4965
$\hat{\beta}_3 \longrightarrow$	time	-1.4827	<.0001

As in the random slope and intercept model in Example 4.4, here the variable `tumorsize` is not a significant covariate at the 5% level of significance either. Removing it from the model results in the reduced model with the following parameter estimates:

```
                     Effect       Estimate    Pr > |t|
$\hat{\beta}_0 \to$  Intercept     93.7869     <.0001
$\hat{\beta}_1 \to$  group         -5.0584     0.0455
$\hat{\beta}_2 \to$  time          -1.3989     <.0001

Covariance
Parameter    Estimate
UN(1,1)       42.3477    $\leftarrow \hat{\sigma}^2_{u_1}$
UN(2,1)       -3.6335    $\leftarrow \hat{\sigma}_{u_1 u_2}$
UN(2,2)        0.7928    $\leftarrow \hat{\sigma}^2_{u_2}$
SP(POW)       -0.7687    $\leftarrow \hat{\rho}$
Residual      46.4864    $\leftarrow \hat{\sigma}^2$
```

Note that the parameter estimates in this model are close to those produced in the random slope and intercept model. The fitted coefficients are interpreted in the same fashion for both models. □

4.6 Random Intercept Logistic Regression Model

Definition

If the response y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, k$ of the ith subject at time t_j is binary (that is, assumes values that may be coded 0 and 1), a random intercept logistic regression model may be used. Denote by $\pi_{ij}(u) = \mathbb{P}(y_{ij} = 1 \,|\, u)$ the conditional probability of $y_{ij} = 1$ given some random variable u. The ratio

$$\frac{\pi_{ij}(u)}{1 - \pi_{ij}(u)} = \frac{\mathbb{P}(y_{ij} = 1 \,|\, u)}{\mathbb{P}(y_{ij} = 0 \,|\, u)}$$

is called the *conditional odds in favor of* $y_{ij} = 1$, *given* u. A *logit* transformation of $\pi_{ij}(u)$ is the natural logarithm of the odds in favor of $y_{ij} = 1$ conditioned on u—that is,

$$\text{logit}\big(\pi_{ij}(u)\big) = \ln\left(\frac{\pi_{ij}(u)}{1 - \pi_{ij}(u)}\right)$$

The *random intercept logistic regression model* has the form

$$\text{logit}\big(\pi_{ij}(u_i)\big) = \beta_0 + \beta_1\, x_{1ij} + \cdots + \beta_p\, x_{pij} + \beta_{p+1}\, t_j + u_i$$

where $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\, \sigma_u^2)$ are the *random intercepts*, $i = 1, \dots, n$, $j = 1, \dots, k$. Equivalently, the model can be written as

$$\pi_{ij}\,(u_i) = \frac{\exp\left\{\, \beta_0 + \beta_1\, x_{1ij} + \cdots + \beta_p\, x_{pij} + \beta_{p+1}\, t_j + u_i \,\right\}}{1 + \exp\left\{\, \beta_0 \,+\, \beta_1\, x_{1ij} + \cdots + \beta_p\, x_{pij} + \beta_{p+1}\, t_j + u_i \,\right\}} \tag{4.18}$$

Estimation of Parameters

The parameters of the model are $\beta_0, \dots, \beta_{p+1}$ and σ_u^2. The maximum-likelihood method is commonly used to estimate these parameters. For given u_i, $i = 1, \dots, n$, the distribution of y_{ij} is Bernoulli with parameter $\pi_{ij}(u_i)$ defined by Equation 4.18. Therefore, the *conditional likelihood function* is

$$\begin{aligned}
L(\beta_0, \dots, \beta_{p+1} \,|\, u_1, \dots, u_n) &= \prod_{i=1}^{n} \prod_{j=1}^{k} \left(\pi_{ij}\,(u_i)\right)^{y_{ij}} \left(1 - \pi_{ij}\,(u_i)\right)^{1-y_{ij}} \\
&= \prod_{i=1}^{n} \prod_{j=1}^{k} \left[\frac{\exp\left\{\beta_0 + \cdots + \beta_{p+1}\, t_j + u_i\right\}}{1 + \exp\left\{\beta_0 + \cdots + \beta_{p+1}\, t_j + u_i\right\}} \right]^{y_{ij}} \\
&\quad \times \left[\frac{1}{1 + \exp\left\{\beta_0 + \cdots + \beta_{p+1}\, t_j + u_i\right\}} \right]^{1-y_{ij}} \\
&= \prod_{i=1}^{n} \prod_{j=1}^{k} \frac{\exp\left\{ \left(\beta_0 + \cdots + \beta_{p+1}\, t_j + u_i\right) y_{ij} \right\}}{1 + \exp\left\{ \beta_0 + \cdots + \beta_{p+1}\, t_j + u_i \right\}}
\end{aligned}$$

To obtain the conventional likelihood function, integrate this expression over all possible values of $u_1, \dots, u_n$, which are independent $\mathcal{N}(0, \sigma_u^2)$ random variables:

$$\begin{aligned}
L(\beta_0, \dots, \beta_{p+1},\, \sigma_u^2) &= \left(2\pi\, \sigma_u^2\right)^{-n/2} \exp\left\{ \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\beta_0 + \cdots + \,\beta_{p+1}\, t_j\right) y_{ij} \right\} \\
&\quad \times \prod_{i=1}^{n} \int_{-\infty}^{\infty} \frac{\exp\left\{\sum_{j=1}^{k} u_i y_{ij} - u_i^2/(2\sigma_u^2)\right\}}{\prod_{j=1}^{k} \left(1 + \exp\left\{\beta_0 + \cdots + \beta_{p+1} t_j + u_i\right\}\right)}\, du_i .
\end{aligned} \tag{4.19}$$

The estimators $\hat{\beta}_0$, $\hat{\beta}_{p+1}$, and $\hat{\sigma}_u^2$ are the numerical solution to the maximization problem for this function.

Interpretation of Coefficients

The interpretation of the regression coefficients $\beta_1, \dots, \beta_{p+1}$ is analogous to that in the ordinary logistic regression model. For convenience, denote by

$$\pi(x_m) = \frac{\exp\left\{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \cdots + \beta_p x_p + \beta_{p+1} t + u\right\}}{1 + \exp\left\{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \cdots + \beta_p x_p + \beta_{p+1} t + u\right\}}$$

where $m = 1, \dots, p+1$. Then the exponentiated regression coefficient β_m represents the odds ratio conditioned on u, when x_m is increased by one unit holding all the other covariates fixed. That is,

$$\exp\left\{\beta_m\right\} = \frac{\pi(x_m + 1)/\big(1 - \pi(x_m + 1)\big)}{\pi(x_m)/\big(1 - \pi(x_m)\big)} \tag{4.20}$$

Hence, $100\big(\exp\{\beta_m\} - 1\big)\,\%$ is the corresponding percentage change in odds.

In the case of a categorical covariate with l levels, the coefficients, say, $\beta_1, \dots, \beta_{l-1}$, correspond to the dummy variables for levels 1 through $l-1$ and, therefore, the quantity $100 \exp\left\{\beta_a - \beta_b\right\}\%$, a, $b = 1, \dots, l-1$, represents the percentage of conditional odds for subjects with the covariate at level a as compared to those for subjects at level b, provided the equality of the other covariates. The percentage of conditional odds for subjects at level a versus those for subjects at level l is $100 \exp\{\beta_a\}\%$.

Example 4.6 Osteoporosis is a metabolic bone disease in which the bones in the body become extremely porous and can be easily fractured. Most commonly, this disease occurs in women older than age 50. A medication that is supposed to help rebuild bones is tested in a nonrandomized clinical trial on 20 women who suffer from osteoporosis. Age at primary visit (in years), calcium supplement intake (yes = 1, no = 0), and family history of osteoporosis (yes = 1, no = 0) are recorded for each subject. During a follow-up visit, each subject has a bone density test. The response variable is whether osteoporosis is present (yes = 1, no = 0). The data are shown in Table 4.3.

The random intercept logistic regression model for these data is

$$\pi_{ij}\,(u_i) = \frac{\exp\left\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 t_j + u_i\right\}}{1 + \exp\left\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 t_j + u_i\right\}}$$

where $\pi_{ij}\,(u_i)$ is the conditional probability of the ith subject having osteoporosis on the jth visit, given u_i, $i = 1, \dots, 20$, $j = 1, \dots, 4$; x_{1i}, x_{2i}, and x_{3i} are, respectively, the age, calcium intake, and history of osteoporosis for the ith subject; $t_1 = 3$, $t_2 = 9$, $t_3 = 12$, and $t_4 = 18$ (all in months); and u_i is the random intercept.

The parameters of this model are $\beta_0, \dots, \beta_4$, and σ_u^2. SAS is used to estimate the parameters by the maximum-likelihood method. Procedure `glimmix`

Table 4.3 Data for Example 4.6

				Presence of Osteoporosis			
Subject	Age	Calcium	History	3 Months	9 Months	12 Months	18 Months
1	76	0	1	1	1	1	1
2	57	1	1	1	1	0	0
3	58	0	1	1	1	1	0
4	62	1	0	1	0	0	0
5	60	1	0	0	0	0	0
6	58	0	1	1	0	1	1
7	52	1	0	0	1	0	0
8	74	0	1	1	0	1	0
9	51	0	0	0	1	0	0
10	56	1	0	0	1	0	0
11	75	0	1	1	1	1	1
12	63	1	0	1	0	0	0
13	67	1	1	0	1	0	0
14	68	0	0	1	1	0	0
15	56	1	0	1	0	1	0
16	62	1	0	1	0	1	1
17	60	1	1	0	0	0	0
18	61	1	1	1	1	0	0
19	54	1	0	1	0	0	0
20	53	0	0	1	1	0	0

fits generalized linear mixed-effects models. To specify the logistic model for a binary response, add the options `dist = binary` and `link = logit`. Also, the probability $\mathbb{P}\big(y_{ij} = 0\big)$ is modeled unless the option `(event = "1")` is typed in. The SAS code is as follows:

```
data osteoporosis;
input individual age calcium history mos3 mos9 mos12 mos18 @@;
datalines;
  1  76  0  1   1   1  1  1
  2  57  1  1   1   1  0  0
                ...
 20  53  0  0   1   1  0  0
;
```

```
data new;
set osteoporosis;
array x{4} mos3 mos9 mos12 mos18;
array t{4} (3 9 12 18);
   do visits = 1 to 4;
   disease = x{visits};
   time = t{visits};
   output;
end;
keep individual age calcium history disease time;
run;

proc glimmix data = new;
model disease(event = "1") = age calcium history time / solution
dist = binary link = logit;
random intercept / subject = individual type = un;
run;
```

As given by SAS, the regression estimates in the full model are

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	-0.9445	0.7544
$\hat{\beta}_1 \rightarrow$	age	0.0545	0.2692
$\hat{\beta}_2 \rightarrow$	calcium	-1.4343	0.0309
$\hat{\beta}_3 \rightarrow$	history	0.5401	0.4273
$\hat{\beta}_4 \rightarrow$	time	-0.1926	0.0012

In this model, only calcium and time are significant variables (at the 0.05 level of significance). The parameter estimates in the reduced model are

	Effect	Estimate	Pr > |t|
$\hat{\beta}_0 \rightarrow$	Intercept	2.7129	0.0031
$\hat{\beta}_2 \rightarrow$	calcium	-1.7626	0.0060
$\hat{\beta}_4 \rightarrow$	time	-0.1825	0.0014

Covariance Parameter	Estimate	
UN(1,1)	0.3225	$\leftarrow \hat{\sigma}_u^2$

From here, the odds in favor of osteoporosis, conditioned on u, for subjects who take calcium supplement is only $100\exp\{-1.7626\} = 17\,\%$ of those for subjects who do not take calcium. In addition, the odds in favor of osteoporosis change every month by $100\big(\exp\{-0.1825\} - 1\big)\,\% = -16.68\,\%$; that is, they decrease by about 17%. □

4.7 Handling Missing Observations

In practice, some observations are frequently missing from a longitudinal data set. If for some reason a subject fails to appear for a particular follow-up visit, the missing observation is called an *intermittent missing value.* If a subject drops out of the study, then starting at some point, all observations are missing for this subject (called a *drop-out* case). A data set with missing observations is called an *unbalanced* (or *incomplete*) data set.

4.7.1 Missingness Mechanism

Several approaches may be used to model longitudinal data with missing observations, depending on the missing value mechanism. Denote by $\mathbf{y}_{i(obs)}$ the observed response measurements for the ith subject, and let $\mathbf{y}_{i(mis)}$ be the *missing values*—the responses that could have been observed were they not missing. Let $\mathbf{r}_i$ be a set of binary random variables indicating whether y_{ij} is an observed or a missing value, $j = 1, \ldots, k$. Then the missing values fall into one of three categories:

- *Missing completely at random* (MCAR), if $\mathbf{r}_i$ is independent of both $\mathbf{y}_{i(obs)}$ and $\mathbf{y}_{i(mis)}$
- *Missing at random* (MAR), if $\mathbf{r}_i$ depends on $\mathbf{y}_{i(obs)}$ but not $\mathbf{y}_{i(mis)}$
- *Missing not at random* (MNAR), if $\mathbf{r}_i$ depends on $\mathbf{y}_{i(mis)}$, and may depend on $\mathbf{y}_{i(obs)}$

Missing Completely at Random (MCAR)

For MCAR values, the missingness does not depend on the response. In other words, missing a follow-up visit or dropping out of a study could have happened to any subject in the study. For example, if a subject missed a scheduled visit to the clinic due to a family emergency, and at that time could not been reached by investigators to reschedule the appointment, the planned measurement was not taken for this subject. Because the subject remained in the study, this is an intermittent missing value that occurs completely at random. An accidental death or geographical relocation may, for example, cause a subject to drop out of a study completely at random.

Missing at Random (MAR)

An MAR intermittent value or a drop-out occurs when the missingness depends on the observed measurement history. For example, suppose an overweight subject enters a clinical trial of a body-fat-reducing medication, but after two follow-up visits that did not show much weight reduction, the subject misses the next visit due to either depression or embarrassment. This is an instance of an intermittent value that is missing at random. Suppose now a young woman suffering from anorexia participates in a study testing a new drug for this psychological disorder. The drug does not help her lack of desire to eat because of fear of becoming obese, and she constantly loses weight. After several visits to the clinic, investigators decide that it is unethical and dangerous to keep her in the study. Her withdrawal from the study is an MAR drop-out.

Missing Not at Random (MNAR)

Intermittent values or drop-outs missing not at random arise when the missingness is related to an unrecorded change in the response variable that causes it to be too high or too low. For example, an overweight subject in the study of a body-fat-reducing medication might show an improvement during the first couple of visits, but suddenly put on weight and become too depressed or embarrassed to show up for the next visit. Then the response measurement for this visit is missing not at random. In the study for an anorexia drug, suppose a subject gains some weight at the beginning, but between two visits she loses weight severely and drops out of the study. The resulting drop-out case is nonrandom.

Distinguishing between MCAR, MAR, and MNAR Observations

How can one tell whether an observation is missing completely at random, at random, or not at random?

For intermittent missing observations, this question is easy to answer, because the subject can be asked personally during the next visit (assuming, of course, that the subject tells the truth). For drop-outs, it is more difficult to know the exact reason, if the person (or the person's relatives) cannot be contacted by investigators. Here, the case of a subject who is withdrawn from the study by the investigators themselves (an MAR drop-out) is dismissed from consideration.

To understand on an intuitive level how to classify drop-outs, consider first a hypothetical situation in which a subject progresses well but then suddenly drops out of the study. It is unlikely that the drop-out decision occurred completely at random, even though it is possible. It is more likely that the person dropped out because he or she was still doing fine (an MAR drop-out) or suddenly felt much worse (an MNAR drop-out). It is not possible, however, to tell which one of these two cases applies, because the actual reason for the drop-out is unknown.

Now suppose that a subject is doing badly and drops out. Again, it is more likely that this case is not an MCAR drop-out, but it is difficult to tell whether it is an MAR drop-out (the person decided to drop out because his or her health was progressively deteriorating) or an MNAR drop-out (the subject suddenly felt much better).

If a subject neither progresses steadily nor regresses and drops out, it is more likely an MCAR drop-out.

Thus, based on the observed response history of a drop-out, it is possible to distinguish between MCAR and the other two cases (with high probability). It is never possible to separate MAR and MNAR drop-outs in this way, however, because in both cases the decision to drop out is based on measurements unavailable to the investigators.

4.7.2 Fitting Mixed-Effects Model to Unbalanced Data Sets

Ignorable and Non-ignorable Missing Observations

MCAR and MAR observations are called *ignorable*, while MNAR values are called *non-ignorable* or *informative.*

The reason for this terminology is as follows. When fitting a mixed-effects model to unbalanced data, the ML or REML methods are applied to estimate the parameters. MCAR or MAR values do not appear in the likelihood function and, therefore, can be ignored. However, the likelihood function depends on missing observations if they are MNAR, and therefore, these are non-ignorable. To see this, note that the contribution of the ith subject to the likelihood function is

$$\begin{aligned} f\big(\mathbf{r}_i, \mathbf{y}_{i(obs)}\big) &= \int f\big(\mathbf{r}_i, \mathbf{y}_{i(obs)}, \mathbf{y}_{i(mis)}\big)\, d\mathbf{y}_{i(mis)} \\ &= \int f\big(\mathbf{r}_i \,|\, \mathbf{y}_{i(obs)}, \mathbf{y}_{i(mis)}\big)\, f\big(\mathbf{y}_{i(obs)}, \mathbf{y}_{i(mis)}\big)\, d\mathbf{y}_{i(mis)} \qquad (4.21) \end{aligned}$$

- If the values are MCAR, then

$$f\big(\mathbf{r}_i \,|\, \mathbf{y}_{i(obs)}, \mathbf{y}_{i(mis)}\big) = f(\mathbf{r}_i)$$

According to Equation 4.21, the ith factor of the likelihood function is

$$f(\mathbf{r}_i) \int f\big(\mathbf{y}_{i(obs)}, \mathbf{y}_{i(mis)}\big)\, d\mathbf{y}_{i(mis)} = f(\mathbf{r}_i)\, f\big(\mathbf{y}_{i(obs)}\big) \propto f\big(\mathbf{y}_{i(obs)}\big)$$

as $f(\mathbf{r}_i)$ does not depend on the model parameters and, therefore, can be disregarded in the likelihood maximization problem.

- If the values are MAR, then

$$f\big(\mathbf{r}_i \,|\, \mathbf{y}_{i(obs)}, \mathbf{y}_{i(mis)}\big) = f\big(\mathbf{r}_i \,|\, \mathbf{y}_{i(obs)}\big)$$

and Equation 4.21 is equal to

$$\begin{aligned} f\big(\mathbf{r}_i \,|\, \mathbf{y}_{i(obs)}\big) \int f\big(\mathbf{y}_{i(obs)}, \mathbf{y}_{i(mis)}\big)\, d\mathbf{y}_{i(mis)} &= f\big(\mathbf{r}_i \,|\, \mathbf{y}_{i(obs)}\big)\, f\big(\mathbf{y}_{i(obs)}\big) \\ &\propto f\big(\mathbf{y}_{i(obs)}\big) \end{aligned}$$

as $f\big(\mathbf{r}_i \,|\, \mathbf{y}_{i(obs)}\big)$ is constant with respect to the model parameters.

- If values are MNAR, then Equation 4.21 cannot be simplified any further. Thus the likelihood function depends on the missing observations, which cannot be ignored for the likelihood maximization purpose.

Managing Unbalanced Data Sets

There are several relatively simple ways to handle the missing observations when fitting a mixed-effects model to an unbalanced data set:

- Work with the unbalanced data and use the ML or REML method for parameter estimation. This approach works best if all missing values are ignorable. However, for MNAR values, this strategy results in biased estimates.

- Delete all *incomplete cases*, which are the subjects with at least one missing observation, and analyze the resulting *balanced* data. This approach is called a *case-deletion method.* It works best when the number of incomplete cases is not large, and there is no structure to the missing data (they are MCAR observations). Otherwise, this method produces biased estimates.

- Impute the missing values and work with the resulting *complete* data set. *Imputation* is the substitution of some values for the missing ones in an incomplete data set. The resulting parameter estimates are biased.

Imputation Procedures

Two types of imputations for intermittent missing values are commonly used. If $y_{ij\,'}$ is missing, it can be replaced by a value $\hat{y}_{ij'}$ which is

- The mean response for the ith subject:

$$\hat{y}_{ij\,'} = \frac{\sum_{j=1}^{k} r_{ij}\, y_{ij}}{\sum_{j=1}^{k} r_{ij}}$$

where $r_{ij} = 1$ if y_{ij} is observed, and $r_{ij} = 0$ if y_{ij} is missing.

This type of imputation is called the *subject-mean imputation.*

- The mean response for the $j\,'$-th occasion:

$$\hat{y}_{ij\,'} = \frac{\sum_{i=1}^{n} r_{ij\,'}\, y_{ij\,'}}{\sum_{i=1}^{n} r_{ij\,'}}$$

where $r_{ij'} = 1$ if $y_{ij'}$ is observed, and $r_{ij\,'} = 0$ if $y_{ij'}$ is missing.

This imputation type is called the *occasion-mean imputation.*

The simplest way to impute missing values for a dropped-out subject is to replace all missing values by the last observation for this subject. That is, if a subject drops out after the j'-th visit, replace the missing responses $y_{i,j'+1}, y_{i,j'+2}, \dots, y_{ik}$ by $y_{ij'}$. This method is known as the *last observation carried forward imputation.*

Remark 4.7 There are several ways to handle missing observations of a binary variable:

- Ignore the missing values and fit the random intercept logistic regression model to the incomplete data set.
- Remove all incomplete cases and fit the logistic model to the complete data set.
- Impute the missing values and fit the logistic model to the imputed data.

Two simple-to-implement imputations for intermittent missing values are

1. Replace a missing value by the *within-subject mode*, which is the most frequent observation (0 or 1) for the subject.
2. Replace a missing value by the *within-occasion mode*, which is the most frequent observation (0 or 1) for the occasion.

For a drop-out, use the last observation carried forward technique. □

Remark 4.8 In some clinical trials, skipping a visit or dropping out results in missing values for both the response variable and the covariates. The same approaches described previously in this section can be applied to handling the missing values in the unbalanced data. In the imputation method, missing observations of the response variable as well as the covariates are imputed. □

4.7.3 Data Example

Example 4.9 Fourteen deaf recipients of cochlear implants (inner ear hearing devices) were followed for 3 years. The response variable was a speech intelligibility score assessed by a panel of investigators. Three subjects were lost to follow-up, and two subjects failed to show up for a visit. The unbalanced data are shown in Table 4.4.

Table 4.4 Data for Example 4.9

	Speech Intelligibility Score				
Subject	3 Months	6 Months	12 Months	24 Months	36 Months
1	46	57	67	68	75
2	24	35	—	—	—
3	30	45	58	72	82
4	22	—	48	66	76
5	18	15	14	—	—
6	35	67	77	80	86
7	27	47	50	55	67
8	12	31	41	47	52
9	55	76	—	—	—
10	18	39	50	67	78
11	10	25	28	33	33
12	35	44	67	73	84
13	22	45	—	68	78
14	35	55	70	75	83

The three methods discussed previously are implemented to fit a mixed-effects model. For simplicity, the procedure is illustrated by fitting a random intercept model

$$y_{ij} = \beta_0 + \beta_1 t_j + u_i + \varepsilon_{ij}, \quad i = 1, \dots, 14, \quad j = 1, \dots, 5$$

where y_{ij} is the recorded score for the ith subject on the jth visit, $t_1 = 3$, $t_2 = 6$, $t_3 = 12$, $t_4 = 24$, $t_5 = 36$ (all in months), and the random intercepts $u_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2)$ are independent of the random errors $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. The REML parameter estimation method is employed.

The first approach is to run procedure `mixed` to produce parameter estimates for the given unbalanced data set. To input the unbalanced data in SAS, write a dot in place of a missing observation. The SAS code is as follows:

```
data recorded;
input individual mos3 mos6 mos12 mos24 mos36 @@;
datalines;
```

```
  1  46  57  67  68  75     2  24  35   .   .   .
  3  30  45  58  72  82     4  22   .  48  66  76
  5  18  15  14   .   .     6  35  67  77  80  86
  7  27  47  50  55  67     8  12  31  41  47  52
  9  55  76   .   .   .    10  18  39  50  67  78
 11  10  25  28  33  33    12  35  44  67  73  84
 13  22  45   .  68  78    14  35  55  70  75  83
;
data unbalanced;
set recorded;
array x{5} mos3 mos6 mos12 mos24 mos36;
array t{5} t1-t5 (3 6 12 24 36);
do visits = 1 to 5;
score = x{visits};
time = t{visits};
output;
end;
keep individual score time;
run;

proc mixed data = unbalanced method = reml;
model score = time / solution;
random intercept / subject = individual;
run;
```

The second approach is to delete all incomplete cases, and apply procedure mixed to the complete data set (shown in Table 4.5).

In SAS, the incomplete cases are deleted from the unbalanced data set. The code is as follows:

```
data complete;
set unbalanced;
if individual = 2 then delete;
if individual = 4 then delete;
if individual = 5 then delete;
if individual = 9 then delete;
if individual = 13 then delete;
run;

proc mixed data = complete method = reml;
model score = time / solution;
```

Table 4.5 Complete Data Set for Example 4.9

	Speech Intelligibility Score				
Subject	3 Months	6 Months	12 Months	24 Months	36 Months
1	46	57	67	68	75
3	30	45	58	72	82
6	35	67	77	80	86
7	27	47	50	55	67
8	12	31	41	47	52
10	18	39	50	67	78
11	10	25	28	33	33
12	35	44	67	73	84
14	35	55	70	75	83

```
random intercept / subject = individual;
run;
```

Finally, the missing observations are imputed and procedure `mixed` is applied to the imputed data set. For instance, an intermittent missing observation is replaced by the subject-mean value, and a drop-out case is substituted by the last recorded observation for the subject. The imputed data are shown in Table 4.6.

For illustration purposes, the filled-in values are given in boldface in Table 4.6. For example, the missing observation for the fourth subject is imputed by the mean response for this subject, which is $(22 + 48 + 66 + 76)/4 = 53$.

The imputed values are computed manually and recorded in SAS using the following code lines:

```
data imputed;
set unbalanced;
      if individual = 2 then do;
         if score = . then score = 35;
      end;
         if individual = 4 then do;
            if score = . then score = 53;
         end;
      if individual = 5 then do;
         if score = . then score = 14;
      end;
         if individual = 9 then do;
```

Table 4.6 Imputed Data Set for Example 4.9

	Speech Intelligibility Score				
Subject	3 Months	6 Months	12 Months	24 Months	36 Months
1	46	57	67	68	75
2	24	35	**35**	**35**	**35**
3	30	45	58	72	82
4	22	**53**	48	66	76
5	18	15	14	**14**	**14**
6	35	67	77	80	86
7	27	47	50	55	67
8	12	31	41	47	52
9	55	76	**76**	**76**	**76**
10	18	39	50	67	78
11	10	25	28	33	33
12	35	44	67	73	84
13	22	45	**53.25**	68	78
14	35	55	70	75	83

```
         if score = . then score = 76;
      end;
   if individual = 13 then do;
      if score = . then score = 53.25;
   end;
run;

proc mixed data = imputed method = reml;
model score = time / solution;
random intercept / subject = individual;
run;
```

As output by SAS, the estimated parameters are

```
          Effect                      Estimate                   Pr > |t|
                        Unbalanced    Complete     Imputed
 β̂0 →    Intercept        32.6197     34.7019     34.6573       <.0001
 β̂1 →    time              1.1867      1.1172      0.9517       <.0001
```

```
Covariance
Parameter                  Estimate
             Unbalanced   Complete   Imputed
Intercept        205.03     178.80    259.47   ← $\hat{\sigma}_u^2$
Residual        85.0518    85.5396    107.82   ← $\hat{\sigma}^2$
```

There is some noticeable variation in the estimates of variances, whereas the regression coefficients are relatively stable. □

Exercises for Chapter 4

Section 4.2

Exercise 4.1 A new drug for mental distress is tested in a randomized controlled trial. The response variable—the general health questionnaire score—is recorded for 18 subjects during 4 visits. The score range is between 0 and 36. Scores larger than 25 are evidence of severe mental distress. The measurements are taken at the initial visit and after 1, 2, and 4 weeks of treatment. Group 1 is the treatment group, and group 2 is the control group. The data are as follows:

		General Health Questionnaire Score			
Subject	Group	0 Week	1 Week	2 Weeks	4 Weeks
1	1	25	23	16	8
2	1	34	22	14	7
3	1	31	24	14	7
4	1	34	27	12	6
5	1	33	25	11	8
6	1	30	23	13	11
7	1	28	22	10	8
8	1	29	21	9	8
9	1	28	21	9	7
10	2	33	25	18	13
11	2	30	27	19	11
12	2	29	22	20	15
13	2	35	27	20	18
14	2	25	23	22	15
15	2	36	25	20	16
16	2	34	25	19	18
17	2	33	28	20	27
18	2	28	24	20	19

Use the graphical tools to determine whether the new drug is effective as a reducer of mental distress.

Section 4.3

Exercise 4.2 Using the definition of $\mathbf{V}_0$ given in Equation 4.1, verify the expressions in Equations 4.3 and 4.4.

Exercise 4.3 Fill in the details in the derivation of the Jacobian determinant in Equation 4.11.

Exercise 4.4 Consider the data in Exercise 4.1.

(a) Discuss the appropriateness of the random intercept model for the general health questionnaire scores.

(b) Write down the model. Estimate all parameters by the ML and REML methods. Use a 5% significance level.

(c) Interpret the fitted values of the regression coefficients. What is your conclusion regarding the effectiveness of the new drug?

Section 4.4

Exercise 4.5 Fourteen overweight subjects underwent *gastric bypass*, a surgical procedure that reduces the stomach size and allows food to bypass part of the small intestine. As a result, the subjects consumed less food and lost weight. Two types of surgery were tested in the study: an open procedure where a large abdominal incision was made, and a *laparoscopic* approach where a small incision was made and a viewing camera (*laparoscope*) was inserted. The seven subjects in group 1 had open surgery; the seven subjects in group 2 had laparoscopic surgery. The subjects were followed for one year. During four follow-up visits, at 1 month, 3 months, 8 months, and 12 months, the percentage loss of the original excess weight (weight above normal at the beginning of the study) was recorded. The data are as follows:

		Percentage Loss of Excess Weight			
Subject	Group	1 Month	3 Months	8 Months	12 Months
1	1	5	12	16	20
2	1	7	7	9	9
3	1	3	6	12	15
4	1	10	15	20	25

(*continued*)

		Percentage Loss of Excess Weight			
Subject	Group	1 Month	3 Months	8 Months	12 Months
5	1	8	10	13	16
6	1	5	10	19	22
7	1	6	6	12	15
8	2	8	12	20	27
9	2	10	15	22	32
10	2	12	17	20	30
11	2	8	16	23	28
12	2	7	11	22	32
13	2	10	16	24	28
14	2	13	15	25	20

(a) Discuss the appropriateness of the random slope and intercept model for these data. Plot necessary graphs to support your argument.

(b) Fit the model using the ML and REML parameter estimation methods. Assume a 0.05 level of significance.

(c) Discuss the signs and values of the estimated parameters. Draw conclusion regarding which is the more effective type of surgery.

Section 4.5

Exercise 4.6 Show that in the model with a spatial power covariance structure of the error terms, the mean of the process w_i is zero and the variance is σ^2.

Exercise 4.7 Consider the data given in Exercise 4.5.

(a) Fit at the 5% significance level the mixed-effects model with a spatial power covariance matrix for the errors using the ML and REML estimation methods.

(b) Compare the fitted model with the random slope and intercept model obtained in Exercise 4.5.

Section 4.6

Exercise 4.8 A new oral antifungal medication is tested in a randomized-controlled clinical trial. Twenty-two subjects with severe toenail fungus infection participate in the trial. The subjects in the control group (group 1) use a common over-the-counter liquid that fights toenail fungus. All subjects are instructed to use the medication once daily for the duration of 16 weeks. Four follow-up visits are scheduled, during which the presence or absence of fungus is recorded (yes = 1, no = 0). The data are as follows:

		Presence of Toenail Fungus			
Subject	Group	3 Weeks	6 Weeks	12 Weeks	16 Weeks
1	1	1	1	0	1
2	1	1	1	1	0
3	1	1	1	0	0
4	1	1	0	0	0
5	1	0	1	1	0
6	1	1	1	0	0
7	1	1	0	0	0
8	1	1	1	1	1
9	1	1	0	0	0
10	1	1	1	1	1
11	1	1	1	0	0
12	2	1	0	1	0
13	2	0	0	0	0
14	2	0	0	0	0
15	2	1	0	0	0
16	2	1	1	0	0
17	2	1	1	1	0
18	2	0	0	0	0
19	2	1	0	0	0
20	2	1	1	0	0
21	2	1	0	0	0
22	2	1	1	1	1

(a) Fit the random intercept logistic regression model to these data. Use significance level of 0.05. Do the data suggest that the oral treatment is effective? Explain.

(b) What is the weekly percentage change in odds of having toenail fungus?

Section 4.7

Exercise 4.9 Suppose in Example 4.4, subject 3 dropped out of the study after the 6-month follow-up visit; subject 13, after the 12-month visit; and subject 16, after the 3-month visit. Also, subject 8 did not show up for the 12-month visit, and subject 19 missed the 6-month and 12-month visits. For these subjects, the measurements of tumor size and therapy satisfaction scores on the given occasions were not recorded.

(a) Fit a random slope and intercept model to the data. Apply REML parameter estimation method. Use (i) an unbalanced data set, (ii) a complete data set, and (iii) an imputed data set. Impute the intermittent missing value for subject 8 by the subject-mean method, and those for subject 19 by the occasion-mean method.

(b) Compare the resulting models to the full model obtained in Example 4.4. Which covariates are significant at the 5% significance level and which are not?

(c) Fit the reduced model to the three data sets in part (a). Compare the estimated values of the parameters for the three cases.

Exercise 4.10 Use the data from Example 4.6. Assume that subjects 5, 9, and 15 dropped out of the study after the 9-month visit, that subject 12 missed the 12-month visit, and that subject 18 missed the 9-month visit.

(a) Fit the random intercept logistic regression. Use (i) an unbalanced data set, (ii) a complete data set, and (iii) an imputed data set. Apply the within-subject mode imputation for subject 12, and the within-occasion mode imputation for subject 18.

(b) Compare the resulting models to the original full-data model in Example 4.6. Which covariates are significant predictors of osteoporosis at the 5% significance level and which are not?

(c) Fit the reduced model to the three data sets in part (a). Compare the magnitudes of the parameter estimators for the three cases, and also for the reduced model in Example 4.6.

Recommended Books

Allison, P. D. (2005). *Fixed Effects Regression Methods for Longitudinal Data Using SAS®*, SAS Institute Inc.

Cantor, A. (2003). *SAS® Survival Analysis Techniques for Medical Research*, 2nd edition, SAS Publishing.

Chow, S.-C., and J.-P. Liu (2004). *Design and Analysis of Clinical Trials: Concepts and Methodologies*, Wiley-Interscience.

Chow, S.-C., J. Shao, and H. Wang (2007). *Sample Size Calculations in Clinical Research*, 2nd edition, Chapman & Hall/CRC.

Cody, R. (2001). *Longitudinal Data and SAS®: A Programmer's Guide*, SAS Institute Inc.

Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd edition, Chapman & Hall/CRC.

Der, G., and B. S. Everitt (2005). *Statistical Analysis of Medical Data Using SAS®*, Chapman & Hall/CRC.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data*, 2nd edition, Oxford University Press.

Everitt, B. S., and A. Pickles (2004). *Statistical Aspects of the Design and Analysis of Clinical Trials*, London: Imperial College Press.

Fairclough, D. L. (2008). *Design and Analysis of Quality of Life Studies in Clinical Trials*, 2nd edition, Chapman & Hall/CRC.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2004). *Applied Longitudinal Analysis*, John Wiley & Sons.

Hedeker, D., and R. D. Gibbons (2006). *Longitudinal Data Analysis*, Wiley-Interscience.

Jennison, C., and B. W. Turnbull (1999). *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC.

Kleinbaum, D. G., and M. Klein (2005). *Survival Analysis: A Self-Learning Text*, 2nd edition, Springer.

Machin, D., Y. B. Cheung, and M. Parmar (2006). *Survival Analysis: A Practical Approach*, John Wiley & Sons.

Ng, R. (2004). *Drugs—From Discovery to Approval*, Wiley-Liss.

Prokscha, S. (2006). *Practical Guide to Clinical Data Management*, 2nd edition, CRC Press.

Rosenberger, W. F., and J. M. Lachin (2002). *Randomization in Clinical Trials: Theory and Practice*, Wiley-Interscience.

Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley & Sons.

Weiss, R. E. (2005). *Modeling Longitudinal Data*, Springer.

Index