20

Sample Size and Power for Longitudinal Studies

20.1 INTRODUCTION

The emphasis in earlier chapters has been on methods for analyzing longitudinal data. In this chapter we consider the design of a longitudinal study. Specifically, we focus on the determination of sample size and power for longitudinal studies. In general, questions about sample size and power arise in the earliest stages of the design of a study. Although the question can be posed in a variety of different ways, investigators typically need to know the answer to the following question: How *large* should my study be? In a cross-sectional study design, with only a single univariate response, the answer to this question is relatively straightforward: the *size* of a study is directly related to the number of subjects, that is, the sample size. However, for a longitudinal study the question of *size* is more complex. For example, in planning a longitudinal study to compare an active treatment to control, investigators need to determine not only how many subjects to enroll in the study but also the duration of the study and the frequency and spacing of repeated measurements on the subjects.

For the special case of a cross-sectional study design, with only a single response, statisticians have developed simple formulas for sample size and power calculations. Explicit formulas can be found in many introductory textbooks in statistics. In addition some statistical software packages include procedures for sample size and power calculations, and publicly available sample size and power calculators can be found on the Web. However, for the multivariate response obtained from a longitudinal study, accurate sample size (and power) determination is more complicated and, in general, requires inversion of matrices and iterative solutions when no closed-form expressions can be obtained. The purpose of this chapter is not to derive complex sample size formulas for longitudinal studies; references to accurate, but also more complex, methods for calculating sample size and power can be found at the end of the chapter. Instead, we present simple, albeit approximate, methods for sample size and power determination for longitudinal studies that allow direct application of standard sample size and power formulas.

In this chapter we begin with a review of sample size (and power) formulas for a univariate continuous response in a cross-sectional study design. We emphasize the main considerations in determining how large the sample size needs to be to achieve a specified power to detect some effect of scientific interest; this section can be skimmed through for those already familiar with power and sample size calculations for a univariate response. We then present simple closed-form expressions for sample size (and power) calculations for longitudinal studies with a continuous response based on the standard sample size (and power) formula for a univariate response. Similar closed-form expressions for longitudinal binary responses are also presented. Finally, two examples are presented to illustrate the application of these formulas to the design of a longitudinal study with a continuous and binary response, respectively.

20.2 SAMPLE SIZE FOR A UNIVARIATE CONTINUOUS RESPONSE

When planning a cross-sectional study, investigators must establish how many subjects they will need to achieve some specified power to detect an effect of subject-matter importance. For example, suppose that investigators are interested in comparing two treatments, an active treatment and a control. The investigators plan to randomize a total of N subjects, with $N_1 = \pi N$ in group 1 (e.g., active treatment), and $N_2 = (1 - \pi)N$ in group two (e.g., control). When $\pi = 0.5$, an equal number of subjects ($N_1 = N_2$) are randomized to receive each of the two treatments. At the completion of the study, the two treatment groups are to be compared in terms of the mean response. Let $\mu^{(1)}$ denote the mean response in the population of individuals assigned to the active treatment; similarly let $\mu^{(2)}$ denote the mean response in the population of individuals assigned to the control. The treatment effect can be expressed in a variety of different ways, but here we consider the simple difference in means, $\delta = \mu^{(1)} - \mu^{(2)}$. The null hypothesis of no treatment difference is represented by H_0: $\delta = 0$. In this example the investigators may be interested in establishing whether the active treatment is superior to control, with the alternative hypothesis that $\delta > 0$.

Before we discuss sample size and power, we must consider the two types of errors that can arise when conducting a statistical test of H_0: $\delta = 0$. The first kind of error is called a type I error and is made if we reject the null hypothesis when in fact it is true. The probability of a type I error, also known as the significance level of the test, is usually denoted by α. Thus, for our example where H_0: $\delta = 0$,

$$\alpha = \Pr(\text{Reject H}_0 \mid \text{H}_0 \text{ is true}).$$

Conventionally, α is chosen to be no greater than 0.05; that is, we are prepared to mistakenly reject the null hypothesis no more than 5% of the time. The second kind of error that can arise when conducting a statistical test is called a type II error. A type II error is made if we fail to reject the null hypothesis when in fact it is false. We denote the probability of a type II error by γ, with

$$\gamma = \Pr(\text{Fail to reject } \text{H}_0 \mid \text{H}_0 \text{ is false}).$$

(The usual convention is to denote the probability of a type II error by β; however, we have chosen to denote it by γ to avoid any potential confusion with our widespread use of β for the regression parameters in earlier chapters of the book.) Since γ is determined by considering the case where the null hypothesis is not true (i.e., $\delta \neq 0$), it necessarily depends on the particular choice of value for $\delta \neq 0$ under the alternative hypothesis. Intuitively, the closer the true value of δ is to zero (the assumed value for δ under the null hypothesis), the more difficult it is to reject H_0: $\delta = 0$. Finally, the power of a statistical test is defined as $1 - \gamma$, that is,

$$\text{power} = 1 - \gamma = \Pr(\text{Reject } \text{H}_0 \mid \text{H}_0 \text{ is false}).$$

In simple terms, the power of a test is the probability that the study will determine that there is a treatment effect of some subject-matter importance when it truly exists. Since γ necessarily depends on the particular choice of value for $\delta \neq 0$ under the alternative hypothesis, so too does the power of a test. Thus, with all other things being equal, the further the true value of δ is from zero, the greater is the power of a test of H_0: $\delta = 0$.

By considering the two types of errors that can arise when conducting a statistical test, we can determine the sample size required to have some specified power to detect an effect, $\delta \neq 0$. For the special case of the two group comparison considered in our example, a test of H_0: $\delta = 0$ can be based on the following z-test,

$$Z = \frac{\widehat{\delta}}{\sqrt{\text{Var}(\widehat{\delta})}} = \frac{\widehat{\mu}^{(1)} - \widehat{\mu}^{(2)}}{\sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}}} = \frac{\widehat{\mu}^{(1)} - \widehat{\mu}^{(2)}}{\sqrt{\frac{\sigma^2}{N\pi(1-\pi)}}},$$

where $\widehat{\delta} = \widehat{\mu}^{(1)} - \widehat{\mu}^{(2)}$ is the difference in sample means in the two groups, and σ^2 is the variance of the response (assumed to be common in the two groups). A formula for the approximate total sample size, $N = N_1 + N_2$, for a 2-tailed test is given by

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, \sigma^2}{\pi(1 - \pi)\delta^2}, \tag{20.1}$$

where $Z_{(1-\alpha/2)}$ and $Z_{(1-\gamma)}$ denote the $(1 - \alpha/2) \times 100\%$ and $(1 - \gamma) \times 100\%$ percentiles of a standard normal distribution (e.g., the 97.5th percentile of a standard normal distribution is 1.96; or put in somewhat simpler terms, 97.5% of the area under the standard normal curve lies to the left of 1.96). When applying (20.1) σ^2 is replaced by an estimate of the variability. Given the total projected sample size, N, the number of subjects in group 1 is $N_1 = \pi N$ and the number of subjects in group 2 is $N_2 = (1 - \pi)N$; these estimates of N_1 and N_2 are rounded up to the next nearest integer.

The main reason for displaying the formula given by (20.1) is to highlight its main constituents. A closer examination of this formula reveals that the determination of sample size requires that all of the following be specified:

1. significance level, α;
2. power, $1 - \gamma$;
3. effect size, δ; and
4. common variance, σ^2.

Ordinarily, the first two factors do not pose a great challenge for investigators. Conventionally, the significance level of a statistical test is fixed at the mythical 0.05 level (with $Z_{(1-\alpha/2)} = 1.96$ for a 2-tailed test). Similarly the lower bound on what might be considered acceptable power is usually set at approximately 80% (with $Z_{(1-\gamma)} = 0.842$ for power = 0.8, or $Z_{(1-\gamma)} = 1.282$ for power = 0.9). This leaves only two key ingredients for which the investigators must provide information: the minimum effect size of scientific interest and an estimate of the variability in the data. Note that the former appears in the denominator of (20.1), while the latter appears in the numerator. As a result, for any fixed value of the variability, the required sample size decreases with increasing effect size, δ. Intuitively, fewer subjects (or less information) are needed when it is of interest to determine whether a true treatment effect is quite far from the null value. Similarly, for any fixed effect size, the required sample size decreases with decreasing variability. For example, the required sample size can be made smaller by using a more reliable measurement instrument.

20.3 SAMPLE SIZE FOR A LONGITUDINAL CONTINUOUS RESPONSE

We first consider the common scenario where investigators are interested in comparing two treatments, an active treatment and control, in terms of *changes* in the mean response over time. Toward the end of the section, we also consider the less common scenario where investigators are interested in the comparison of the *time-averaged* response (i.e., the *average* response over the duration of the study) rather than *changes* in the mean response. Throughout, we assume that investigators plan to randomize a total of N subjects, with $N_1 = \pi N$ in group 1 (e.g., active treatment), and $N_2 = (1 - \pi)N$ in group 2 (e.g., control). They plan to take n repeated measurements of the response (not necessarily equally spaced measurements).

20.3.1 Sample Size for Comparison of Change in Response

In this section we consider the case where, at the completion of the study, the two treatment groups are to be compared in terms of changes in the mean response over the duration of the study. For simplicity we assume that changes in the mean response can be expressed in terms of a linear trend and the treatment effect can be expressed in terms of the difference in slopes or rates of change, say δ. Under the null hypothesis of no treatment difference, that is, no treatment × linear trend interaction, H_0: $\delta = 0$. We show that sample size calculations for such a longitudinal study design can be simplified so that the standard sample size formula given by (20.1) can be used. This is achieved by considering the two-stage model for longitudinal data described in Chapter 8 (see Section 8.4). Let us assume the following two-stage formulation. At the first stage, we assume that a simple parametric curve (e.g., linear trend in time) fits the observed responses for each subject. In the second stage, these individual-specific parameters are then related to covariates that describe the different groups from which the individuals have been drawn (e.g., active treatment versus control).

Stage 1: In the first stage subjects are assumed to have their own unique individual-specific response trajectories. That is, in stage 1 we posit that the repeated measures on each individual follow a regression model with a linear trend in time, but with separate regression coefficients for each individual

$$Y_{ij} = \beta_{1i} + \beta_{2i}\, t_j + \epsilon_{ij},$$

where the errors, ϵ_{ij}, are assumed to be independent and identically distributed, having a normal distribution with mean equal to zero and variance σ_ϵ^2, that is, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

Stage 2: In the second stage we assume that the individual-specific effects, $\beta_i = (\beta_{1i}, \beta_{2i})'$, are random. The mean and covariance of β_{1i} and β_{2i} are the population parameters that are modeled in the second stage. Specifically, variation in β_i is modeled as a function of between-individual covariates, which we assume here includes only the treatment group. Thus we can allow the mean of β_i (i.e., the mean intercept and slope) to depend on the treatment group,

$$\begin{aligned} E(\beta_{1i}|\text{Group}_i = g) &= \beta_1^{(g)}, \text{ for } g = 1, 2, \\ E(\beta_{2i}|\text{Group}_i = g) &= \beta_2^{(g)}, \text{ for } g = 1, 2, \end{aligned}$$

where $\text{Group}_i = 1$ if the i^{th} individual was assigned to the active treatment, and $\text{Group}_i = 2$ otherwise. In this model, $\beta_2^{(1)}$ is the mean slope, or constant rate of change in the mean response over time, in the active treatment group, while $\beta_2^{(2)}$ is the mean slope in the control group. That is, $\beta_2^{(1)} - \beta_2^{(2)}$ has interpretation in terms of a treatment group difference in the rate of change in the mean response and corresponds to the definition of δ given earlier. The residual between-individual variation in the β_i that cannot be explained by treatment group is expressed as

$$\text{Cov}(\beta_i) = G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix},$$

where $g_{11} = \text{Var}(\beta_{1i})$, $g_{22} = \text{Var}(\beta_{2i})$, and $g_{12} = g_{21} = \text{Cov}(\beta_{1i}, \beta_{2i})$.

This two-stage formulation yields a tractable form for the component of variability required for sample size and power calculations. If each subject is measured at a common set of occasions, $t_1, ..., t_n$, and there are N_1 and N_2 subjects in the two treatment groups (for a total sample size of N), we can derive simple expressions for sample size and power similar to the univariate setting. Letting $\widehat{\beta}_{2i}$ denote the ordinary least squares (OLS) estimate of the slope for the i^{th} subject, the variability of $\widehat{\beta}_{2i}$ is given by

$$\mathrm{Var}(\widehat{\beta}_{2i}) = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \; g_{22},$$

where

$$\bar{t} = \frac{1}{n} \sum_{j=1}^{n} t_j.$$

Thus the variability of $\widehat{\beta}_{2i}$ is composed of two components: the within-subject variance, $\sigma_\epsilon^2 \{\sum_{j=1}^{n} (t_j - \bar{t})^2\}^{-1}$, and the between-subject variance, $g_{22} = \mathrm{Var}(\beta_{2i})$. To test whether the mean slopes are equal in the two treatment groups, we can construct the following z-test based on averages of the $\widehat{\beta}_{2i}$,

$$Z = \frac{\widehat{\delta}}{\sqrt{\mathrm{Var}(\widehat{\delta})}} = \frac{\overline{\beta}_2^{(1)} - \overline{\beta}_2^{(2)}}{\sqrt{\frac{\sigma_\beta^2}{N_1} + \frac{\sigma_\beta^2}{N_1}}} = \frac{\overline{\beta}_2^{(1)} - \overline{\beta}_2^{(2)}}{\sqrt{\frac{\sigma_\beta^2}{N\pi(1-\pi)}}},$$

where $\overline{\beta}_2^{(1)}$ and $\overline{\beta}_2^{(2)}$ are the sample averages of $\widehat{\beta}_{2i}$ in the treatment and control groups respectively, $\sigma_\beta^2 = \mathrm{Var}(\widehat{\beta}_{2i})$, and π is the proportion of subjects in Group 1.

Given estimates of g_{22}, the between-subject variability in slopes, and σ_ϵ^2, the within-subject variability, the sample size can be determined from the standard formula (20.1) introduced in Section 20.2,

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, \sigma_\beta^2}{\pi(1-\pi)\delta^2}, \tag{20.2}$$

where now

$$\sigma_\beta^2 = \mathrm{Var}(\widehat{\beta}_{2i}) = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + \; g_{22},$$

and δ is a treatment effect size of interest (i.e., δ is the treatment group difference in slopes or rates of change in the mean response). Notice that the sample size formula given by (20.2) is virtually identical to (20.1), except that σ_β^2 has two components: a within-subject variance component, $\sigma_\epsilon^2 \{\sum_{j=1}^{n} (t_j - \bar{t})^2\}^{-1}$, and a between-subject variance component, $g_{22} = \mathrm{Var}(\beta_{2i})$. When applying (20.2), σ_β^2 is replaced by estimates of these two sources of variability. Furthermore, if the measurement occasions are equally spaced (at least approximately) throughout the duration of the study, then the expression for $\sum_{j=1}^{n} (t_j - \bar{t})^2$ simplifies to $\{\tau^2 \, n\,(n+1)\}/\{12\,(n-1)\}$, where τ denotes the duration of the study. (The latter expression can be derived by using the fact that the variance of the first n integers is $(n+1)(n-1)/12$.)

The sample size formula given by (20.2) can also be manipulated to determine the power of a test of H_0 for a given sample size, since (20.2) implies that

$$Z_{(1-\gamma)} = \sqrt{\frac{N\pi(1-\pi)\delta^2}{\sigma^2_\beta}} - Z_{(1-\alpha/2)}. \qquad (20.3)$$

Therefore the power, $1-\gamma$, is given by $\Phi\{Z_{(1-\gamma)}\}$, where $\Phi(\cdot)$ denotes the cumulative standard normal distribution function. That is, the value of $Z_{(1-\gamma)}$ can be calculated from (20.3), and the power is determined by the area under the standard normal curve that lies to the left of $Z_{(1-\gamma)}$. For example, if $Z_{(1-\gamma)} = 1.08$ for some given sample size, then the projected power is 0.86, corresponding to the area under the standard normal curve that lies to the left of 1.08.

For more general models (e.g., quadratic trends or a spline function of time), sample size formulas for two group comparisons of other coefficients in the model for the mean response can be derived by using the general formulation for the stage 1 model (see Section 8.4),

$$Y_i = Z_i\beta_i + \epsilon_i,$$

where the matrix Z_i specifies how an individual's responses change over time and β_i is a $q \times 1$ vector of individual-specific regression coefficients. Then, for any particular trend of interest, the variance, σ^2_β, in the sample size formula is simply obtained from the appropriate diagonal element of

$$\begin{aligned} \text{Cov}(\widehat{\beta}_i) &= \sigma^2_\epsilon\,(Z_i'Z_i)^{-1} + \text{Cov}(\beta_i) \\ &= \sigma^2_\epsilon\,(Z_i'Z_i)^{-1} + G. \end{aligned}$$

In our previous example, with random intercepts and slopes,

$$\begin{aligned} \text{Cov}(\widehat{\beta}_i) &= \text{Cov}\begin{pmatrix} \widehat{\beta}_{1i} \\ \widehat{\beta}_{2i} \end{pmatrix} \\ &= \sigma^2_\epsilon \left[\begin{pmatrix} 1,\ldots,1 \\ t_1,\ldots,t_n \end{pmatrix} \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \right]^{-1} + \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}, \end{aligned}$$

and $\text{Var}(\widehat{\beta}_{2i})$ is the lower-diagonal element (2nd row, 2nd column) of this 2×2 matrix.

In the absence of any information about the variability of the random slopes, the simple formulas for sample size and power given by (20.2) and (20.3) cannot be used. A number of textbooks, and commercially available software for sample size and power calculations, rely on formulas that assume a longitudinal model with only randomly varying intercepts (or random subject-effects). This implies a compound symmetry covariance matrix, with constant variance over time, say $\text{Var}(Y_{ij}) = \sigma^2_b + \sigma^2_\epsilon$ (where σ^2_b and σ^2_ϵ denote the between-subject and within-subject variances, respectively), and constant correlation, say ρ, among pairs of repeated measurements, where

$$\rho = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_\epsilon}.$$

By making the strong assumption of no between-subject variability in the slopes ($g_{22} = 0$), the sample size formula given by (20.2) can be used where

$$\sigma_{\hat{\beta}}^2 = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + g_{22}$$

is replaced by

$$\sigma_{\hat{\beta}}^2 = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} = (1-\rho)(\sigma_b^2 + \sigma_\epsilon^2) \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} .$$

This yields the following formula for sample size,

$$\begin{aligned} N &= \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \sigma_\epsilon^2}{\pi(1-\pi)\,\delta^2 \sum_{j=1}^{n}(t_j - \bar{t})^2} \\ &= \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 (1-\rho)\,\text{Var}(Y_{ij})}{\pi(1-\pi)\,\delta^2 \sum_{j=1}^{n}(t_j - \bar{t})^2}; \end{aligned} \qquad (20.4)$$

a similar formula for power can be based on (20.3) with $\sigma_{\hat{\beta}}^2$ replaced by the expression above. This simple formula for sample size is instructive about how sample size (and power) is impacted by the correlation among repeated measures. It is apparent from (20.4) that N decreases with increasing (positive) correlation, ρ. In general, when estimating and comparing *change* in the response over time, the sample size required decreases with increasing magnitude of the positive correlations among the responses. Although the formula given by (20.4) appears in numerous textbooks that discuss sample size (and power) calculations for longitudinal study designs, we must caution the reader that use of (20.4) can dramatically underestimate the required sample size. That is, it can be shown that the assumption of equal variances and equal correlations (compound symmetry) on which (20.4) is based always produces an underestimate of the projected sample size if a more complex random effects covariance structure (e.g., random intercepts and slopes) or an arbitrary covariance matrix actually holds.

Finally, we conclude this section by noting that the sample size formulas given by (20.2) and (20.4) are special cases of a more general formula based upon contrasts or linear summary statistics for the response vector, say

$$C_i = \sum_{j=1}^{n} a_j Y_{ij},$$

for a set of known weights a_j (for $j = 1, ..., n$), and where

$$\delta = E(C_i|\text{Group}_i = 1) - E(C_i|\text{Group}_i = 2).$$

For example, the ordinary least squares (OLS) estimator of the slope for the i^{th} subject, considered earlier in this section, can be expressed as

$$C_i = \sum_{j=1}^{n} a_j Y_{ij}, \text{ where } a_j = \frac{(t_j - \bar{t})}{\sum_{j=1}^{n}(t_j - \bar{t})^2};$$

other choices of weights can be used to summarize the quadratic trend over time (e.g., polynomial contrast coefficients), area under the curve (AUC), or the mean of all post-baseline measurements minus the baseline. For example, with five equally spaced measurements, a summary statistic for the quadratic time trend is given by

$$C_i = 2 \times Y_{i1} - 1 \times Y_{i2} - 2 \times Y_{i3} - 1 \times Y_{i4} + 2 \times Y_{i5},$$

with $a_1 = 2, a_2 = -1, a_3 = -2, a_4 = -1, a_5 = 2$. Similarly a summary statistic for the mean of all post-baseline measurements minus the baseline is given by

$$C_i = -1 \times Y_{i1} + \frac{1}{4} \times (Y_{i2} + Y_{i3} + Y_{i4} + Y_{i5}), (\text{with } a_1 = -1, a_2 = a_3 = a_4 = a_5 = \frac{1}{4}).$$

For any choice of weights used to form the summary statistic, C_i, sample size can be determined by the following general formula:

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \sigma_C^2}{\pi(1-\pi)\delta^2}, \tag{20.5}$$

where now

$$\sigma_C^2 = \text{Var}(C_i) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_j \times a_k \times \sigma_{jk}, \tag{20.6}$$

and $\sigma_{jk} = \text{Cov}(Y_{ij}, Y_{ik})$. This formula can be used for two group comparisons of any summary statistic (e.g., slope, polynomial contrast, AUC minus baseline); in addition it allows for more general patterns for the covariance among the responses. In Section 20.3.3 we use this formula for the comparison of a "time-averaged" response.

20.3.2 Effects of Study Duration and Frequency of Observation

In this section we closely examine the simple formula for sample size given by (20.2) to reveal how sample size (and power) is impacted by:

1. the length of the study,
2. the number of repeated measures, and
3. the spacing of the repeated measures.

Consider the design of a longitudinal study with n repeated measurements. Let $t_1 = 0$ denote the baseline measurement occasion and $t_n = \tau$ denote the time at the final measurement occasion (i.e., τ denotes the duration of the study). Study investigators can reduce the required sample size (or, correspondingly, increase the power for a fixed total sample size, N) by reducing the magnitude of σ_β^2. Recall that σ_β^2 depends on both the between-subject and within-subject variability. In general, investigators have relatively little control over the natural heterogeneity of the study population, denoted in this instance by $g_{22} = \text{Var}(\beta_{2i})$ and, more generally, by $G = \text{Var}(\beta_i)$. The between-subject variability can only be reduced by focusing on a more homogeneous population. However, doing so could alter the intended target of inference and reduce the generalizability of the results. In principle, the within-subject variability, σ_ϵ^2, can be reduced by using a more reliable measurement instrument; however, this will not always be possible or practical. Therefore, to reduce the magnitude of $\sigma_{\hat\beta}^2$, we must focus on ways to increase the magnitude of

$$\sum_{j=1}^{n}(t_j - \bar{t})^2.$$

Because $\sum_{j=1}^{n}(t_j - \bar{t})^2$ is a divisor of σ_ϵ^2, increasing its magnitude reduces the contribution of the within-subject variance to $\sigma_{\hat\beta}^2$. Note that $\sum_{j=1}^{n}(t_j - \bar{t})^2$ is the sum of the squared deviations of the measurement times about their mean. It is a function of the duration of the study, τ, the number of repeated measurements, n, and the relative spacing of the repeated measurements. For a study of fixed length τ and fixed number of repeated measures n, $\sum_{j=1}^{n}(t_j - \bar{t})^2$ is maximized when $n/2$ measurements are taken at baseline and $n/2$ measurements are taken at the end of the study (when n is an even number). In general, such a study design would not be desirable because it relies too heavily on the assumption that changes in the response are linear over time and precludes examination of non-linear (e.g., quadratic) trends. Also the notion of taking $n/2$ replicate measurements at the same occasion is not feasible or practical in many settings. So, for the remainder of this discussion, we assume that the measurement occasions will be equally spaced (at least approximately) throughout the duration of the study. That is, in a study of length τ, the n repeated measurements are to be taken at times $t_1 = 0, t_2 = \tau/(n-1)$, $t_3 = 2\tau/(n-1), ..., t_n = \tau$. Recall that with equally spaced measurement it can be shown that

$$\sum_{j=1}^{n}(t_j - \bar{t})^2 = \frac{\tau^2\, n\,(n+1)}{12\,(n-1)}.$$

Thus, for a fixed number of repeated measurements, doubling the length of the study decreases the impact of the within-subject variability by a factor of 4. Impressive as this may seem, there are a number of practical limitations that qualify this result. First, the length of a longitudinal study is usually determined by economic, logistical, and subject-matter factors that constrain the maximum length of follow-up. Second, changes in the mean response, as a function of exposure to some treatment or intervention, may be of limited duration and constrain the maximum value of τ. As a result many study investigators are restricted to a relatively narrow range of possible values for τ. Third, increasing the duration of a study may potentially increase the rate of attrition.

The simple formula for $\sum_{j=1}^{n}(t_j - \bar{t})^2$ given above also indicates that for fixed τ the impact of the within-subject variability decreases non-linearly with increasing n. For example, increasing the number of repeated measurements from $n = 2$ (a simple pre-post longitudinal design) to $n = 4$, $n = 6$, $n = 8$, and $n = 10$, results in a 10%, 29%, 42%, and 50% reduction in the impact of the within-subject variability. However, for a study of fixed length τ, there may be some practical constraints on the number of repeated measurements that can be taken. Also we must caution that these results for the impact of increasing either the length of the study or the number of repeated measures rely heavily on the assumption that changes in the mean response over the duration of the study are linear in time. For example, when the assumed trend is curvilinear (e.g., quadratic trend in time), it can be shown that the number of repeated measurements, n, has an even more pronounced effect on reducing the impact of the within-subject variability.

The simple formulas for sample size and power given by (20.2) and (20.3) can be useful for making informed decisions about how best to design a longitudinal study. For any fixed values of π, δ, and τ, the first term on the right-hand side of the formula for power (20.3),

$$\sqrt{\frac{N\pi(1-\pi)\delta^2}{\sigma_\beta^2}}$$

increases as

$$\frac{\sigma_\beta^2}{N} = \left\{ \frac{12\,(n-1)\,\sigma_\epsilon^2}{\tau^2\, n\,(n+1)} + g_{22} \right\} /N$$

decreases. Therefore, among other factors, the power of a longitudinal study is determined by a combination of the number of subjects, N, and the number of repeated measurements, n. Increasing either the sample size or the number of repeated measurements per subject increases power. However, relative to an increase in the number of repeated measurements, increasing the sample size will generally have a far greater effect on power. The reason for this can be seen by noting that as n increases, for a fixed N, it reduces the impact of the within-subject variability (σ_ϵ^2) but leaves the between-subject variability (g_{22}) unchanged. Therefore, even with an infinitely large number of repeated measurements, the power of the study will have an upper bound that is less than 1; how much less than 1 will be determined by the magnitude of the between-subject variability (relative to N and δ). In contrast, as sample size increases, for a fixed n, power increases without any bound (leading eventually to power approaching 1 for a sufficiently large N). Put another way, an increase in the number of repeated measurements only decreases the impact of the within-subject variability, while an increase in sample size decreases the impact of both the within-subject and between-subject variability. This suggests that adding more repeated measures to a study of fixed length is most helpful only in cases where the within-subject variation is large relative to the between-subject variation. In general, adding more subjects is the most effective way to increase the power of a longitudinal study.

20.3.3 Sample Size for Comparison of Time-Averaged Response

So far our discussion of sample size and power has focused on study designs where we are primarily interested in comparing groups in terms of *changes* in the mean response over time. Although less common, sometimes it is of interest to design a longitudinal study where the main comparison of interest is the *average* response over the duration of the study. For example, a between-group comparison of the *time-averaged* response provides a relatively powerful analysis in a study where the treatment effect has a quick onset and numerous repeated measurements are obtained after a maximum effect has been attained. This comparison can be evaluated by calculating a single composite score (or mean) of the correlated repeated measurements for each individual, say $C_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$. This is an example of a linear summary statistic discussed at the end of Section 20.3.1, where $C_i = \sum_{j=1}^{n} a_j Y_{ij}$ and $a_j = \frac{1}{n}$ (for $j = 1, ..., n$). Values of C_i can then be analyzed using a standard t-test for independent groups. With a single summary score, C_i, sample size (and power) calculations are relatively straightforward and can rely on the general formula for the difference between means for two independent groups (see equation (20.5) in Section 20.3.1). Specifically, a formula for the approximate total sampe size, $N = N_1 + N_2$, is given by

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, \sigma_C^2}{\pi(1-\pi)\delta^2},$$

where $\sigma_C^2 = \text{Var}(C_i)$ denotes the variance of C_i (assumed to be common for the two groups), and $\delta = \{E(C_i|\text{Group}_i = 1) - E(C_i|\text{Group}_i = 2)\}$ is the difference in the mean of C_i for the two groups. Recall that the variance of a sum (or average) of n repeated measurements is a function of the variances of the measurements at the n occasions and their intercorrelations. Specifically,

$$\text{Var}(C_i) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \rho_{jk} \times \sigma_j \times \sigma_k,$$

where $\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik})$ and $\sigma_j^2 = \text{Var}(Y_{ij})$. Note that this expression for $\text{Var}(C_i)$ is a special case of equation (20.6) in Section 20.3.1, where $a_j = \frac{1}{n}$ (for $j = 1, ..., n$). A simpler expression for $\text{Var}(C_i)$ is obtained when it can be assumed that the variances are constant over time, with $\sigma_j^2 = \sigma_k^2 = \sigma^2$,

$$\text{Var}(C_i) = \frac{\sigma^2}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \rho_{jk};$$

if in addition it is assumed that the correlations are approximately constant, with $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$, then

$$\text{Var}(C_i) = \frac{1}{n}\{1 + (n-1)\rho\}\sigma^2.$$

The design and analysis of a study with a time-averaged response, C_i, requires greater care when Y_{i1} is a "baseline" response; see Chapter 5, Sections 5.6 and 5.7, for a comparison of alternative strategies for handling the baseline response. For example, in the setting of a randomized trial, where Y_{i1} is a baseline (pre-randomization) response, the comparison of groups in terms of C_i is a less meaningful indicator of treatment effect because the expected difference between treatment groups at baseline is known to be zero. Inclusion of the baseline response in C_i would bias the comparison of the treatment groups toward the null. There are two alternative approaches to the analysis in this setting. First, the baseline response can be excluded from the construction of the time-averaged score, with

$$C_i = \frac{1}{n-1}\sum_{j=2}^{n} Y_{ij}$$

based only on the $(n-1)$ post-baseline responses. Values of C_i can then be analyzed using a standard t-test for independent groups. Second, the analysis of C_i based on the $(n-1)$ post-baseline responses can be adjusted for the baseline response; that is, we can test for group differences in C_i using analysis of covariance (ANCOVA), with baseline response as a covariate. In general, the latter is the preferred method of analysis because it can substantially reduce the variability of C_i, leading to a more powerful test of group differences and a study requiring fewer subjects (and/or fewer repeated measurements). With adjustment for the baseline response, Y_{i1}, the sample size formula given above can be used by replacing σ_C^2 with

$$\sigma_C^2 = \text{Var}(C_i)\,[1-\{\text{Corr}(Y_{i1}, C_i)\}^2],$$

where $\text{Var}(C_i)$ is defined above (albeit now for C_i based only on the $(n-1)$ post-baseline responses), and

$$\begin{aligned}
\text{Corr}(Y_{i1}, C_i) &= \frac{\text{Cov}(Y_{i1}, C_i)}{\sqrt{\text{Var}(Y_{i1})\,\text{Var}(C_i)}} \\
&= \frac{\frac{1}{n-1}\sum_{j=2}^{n}\rho_{1j}\times\sigma_1\times\sigma_j}{\sqrt{\sigma_1^2\,\frac{1}{(n-1)^2}\sum_{j=2}^{n}\sum_{k=2}^{n}\rho_{jk}\times\sigma_j\times\sigma_k}} \\
&= \frac{\sum_{j=2}^{n}\rho_{1j}\times\sigma_j}{\sqrt{\sum_{j=2}^{n}\sum_{k=2}^{n}\rho_{jk}\times\sigma_j\times\sigma_k}}.
\end{aligned}$$

From the expression above for σ_C^2 it is apparent that adjustment for baseline response reduces variability when the baseline response is correlated with C_i. This reduction in variability decreases the required sample size (or, correspondingly, increases the power for a fixed total sample size, N). If it can be assumed that $\sigma_j^2 = \sigma_k^2 = \sigma^2$, and that the correlations are approximately constant, with $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$, then σ_C^2 in the sample size formula can be replaced by

$$\sigma_C^2 = \text{Var}(C_i)\,[1-\{\text{Corr}(Y_{i1}, C_i)\}^2] = \frac{(1-\rho)\{1+(n-1)\rho\}\,\sigma^2}{(n-1)}.$$

Although the formula for σ_C^2 given above is simple, we must remind the reader that the assumption of equal variances and equal correlations produces inaccurate estimates of the projected sample size if a more complex covariance matrix actually holds.

When there is no a priori reason to assume the groups have the same mean response at baseline (e.g., in an observational study), the preferred strategy for adjusting for baseline response is to compare the groups in terms of the mean of all post-baseline measurements minus the baseline (see Section 20.3.1). That is, the comparison of groups can be made in terms of C_i, where

$$C_i = \left(\frac{1}{n-1} \sum_{j=2}^{n} Y_{ij} \right) - Y_{i1}.$$

In concluding this section on sample size and power for a longitudinal continuous response, we note that the focus has been exclusively on the simple two group study design. Longitudinal studies comparing three or more treatment groups are not uncommon. Although sample size and power calculations for general linear models (e.g., ANOVA with three or more groups or linear regression with a quantitative covariate) involve a level of complexity greater than that required for the simple two group setting, the key ingredients required for their computation remain the same. Using the variance formulas for linear summary statistics, $\sigma^2_{\hat{\beta}}$ and σ^2_C, outlined in earlier sections, the extensions to three or more treatment groups, or to the regression setting with a quantitative covariate, are relatively straightforward. Finally, throughout all our discussion of sample size and power, we have assumed no missing data or attrition. The impact of missing data is difficult to quantify precisely because it depends on the patterns of missingness, and in this case simple formulas no longer apply. An admittedly ad hoc, but conservative, approach for adjusting for attrition is to inflate the required sample size in each group to account for the assumed rate of attrition (or proportion of subjects who drop out before the completion of the study). That is, if the rate of attrition is assumed to be 10% in each group, then the projected total sample size should be $N/0.9$.

20.3.4 Example: Longitudinal Study with a Continuous Response

To illustrate the application of the sample size formula (20.2), let us return to the example introduced earlier. Suppose that investigators are interested in comparing two treatments, an active treatment and control, in terms of changes in the mean response over time. The investigators plan to randomize an equal number of subjects ($N_1 = N_2$, with $\pi = 0.5$) to receive either of the two treatments. They plan to take five repeated measurements of the response, one at baseline, and the remainder at 6-month intervals until the completion of the study ($\tau = 2$ years). The response variable is continuous and assumed to have an approximate normal distribution. At the completion of the study, the two treatment groups are to be compared in terms of changes in the mean response over the duration of the study. For simplicity we assume that changes in the mean response can be expressed in terms of a linear trend over time (in years) and the treatment effect can be expressed in terms of the difference in slopes, say δ.

Suppose that the investigators want to detect a minimum treatment effect of $\delta = 1.2$, that is, a difference in the annual rates of change in the treatment and control groups of no less than 1.2. Based on historical data from similar populations, the investigators posit that the between-subject variability in the rate of change, Var(β_{2i}) $\approx$ 2 and the within-subject variability, $\sigma_\epsilon^2 \approx 7$. Finally, the investigators desire to have power of 90% when conducting a 2-sided test at the 5% significance level (i.e., $\gamma = 0.1$ and $\alpha = 0.05$). Given these specifications,

$$\sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} = \frac{12\,(n-1)\,\sigma_\epsilon^2}{\tau^2\, n\,(n+1)} = \frac{12 \times 4 \times 7}{4 \times 5 \times 6} = 2.8$$

and

$$\sigma_\beta^2 = \sigma_\epsilon^2 \left\{ \sum_{j=1}^{n} (t_j - \bar{t})^2 \right\}^{-1} + g_{22} = 2.8 + 2.0 = 4.8.$$

The projected total sample size required is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2\, 4\sigma_\beta^2}{\delta^2} = \frac{(1.96 + 1.282)^2 \times 4 \times 4.8}{1.44} = 140.1.$$

Thus, to ensure that they have power of at least 90%, the investigators will need to enroll a total of 142 subjects, randomizing an equal number (71) to each of the two treatment groups. Note that a study of the same duration ($\tau = 2$ years) with $n = 3$ repeated measurements, 12 months apart, would require a total of 162 subjects to achieve comparable power. Alternatively, if it were feasible to conduct the study over 3 years instead of 2 years (and have the same retention rate), with $n = 5$ repeated measurements taken 9 months apart, it would require a total of 96 subjects to achieve power of at least 90%.

For this example it is of interest to study the relationship of power, sample size, and the number of repeated measurements (assuming a study of the same duration $\tau = 2$ years). Table 20.1 displays the power as a function of the total sample size, N, and the number of equally spaced repeated measurements, n. Table 20.1 is revealing about the trade-offs of increasing the sample size versus increasing the number of repeated measurements. For example, doubling the sample size leads to a discernibly greater increase in power than doubling of the number of repeated measurements. This can be explained by the fact that increases in the number of repeated measurements only reduce the impact of the within-subject variance component in the formula for power. Recall that σ_β^2 depends on both the between-subject and within-subject variability. In contrast, increasing the sample size reduces the impact of both sources of variability.

Next, for illustrative purposes only, we consider the projected sample size based on (20.4) under the assumption that there is no between-subject variation in the slopes ($g_{22} = 0$). In that case $\sigma_\beta^2 = 2.8$ and the projected total sample size required is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2\, 4\sigma_\beta^2}{\delta^2} = \frac{(1.96 + 1.282)^2 \times 4 \times 2.8}{1.44} = 81.7.$$

This suggests that the investigators will need to enroll a total of 82 subjects (in contrast to the earlier calculation of a total of 142 subjects when $g_{22} = 2.0$). However, this is a substantial underestimate of the sample size required to ensure power of at least 90%

Table 20.1 Power as a function of sample size and the number of equally spaced repeated measurements in a longitudinal study of fixed duration.

	Number of Repeated Measures (n)				
Sample Size (N)	2	4	6	8	10
40	0.37	0.39	0.43	0.47	0.50
80	0.63	0.66	0.72	0.76	0.79
120	0.80	0.83	0.87	0.90	0.93
160	0.90	0.92	0.95	0.97	0.98
200	0.95	0.96	0.98	0.99	0.99

Note: Power when conducting a 2-sided test at the 5% significance level ($\alpha = 0.05$) when $\tau = 2, \delta = 1.2$, $\text{Var}(\beta_{2i}) = 2$, and $\sigma_\epsilon^2 = 7$.

if the true variability in slopes is $g_{22} = 2.0$. Put another way, if $g_{22} \approx 2.0$, then the projected sample size of 82 subjects will provide power of approximately 70% instead of 90%, resulting in a study that is under-powered to detect the effect of interest. On a somewhat technical note, it can be argued that this illustration exaggerates the potential underestimation of sample size when based on (20.4) because assuming $g_{22} = 0$, when in fact $g_{22} > 0$, is also likely to lead to a somewhat larger projected estimate of the within-subject variability, that is, a projected estimate of $\sigma_\epsilon^2 > 7$. While conceding this point, it does not alter the fact that use of (20.4) is problematic when $g_{22} > 0$. The main purpose of this illustration is to underscore our earlier warning about the use of (20.4) for sample size (and power) calculations; in general, it produces an underestimate of the projected sample size (or, correspondingly, an overestimate of the power in the sense that the projected sample size yields less power than the nominal level) when there is any between-subject variation in the slopes.

It should be apparent from (20.2) and (20.3) that sample size and power calculations are sensitive to assumptions about the covariance among the repeated measures. Because σ_β^2 depends on assumptions about the magnitudes of the between-subject and within-subject variability, it is advisable to perform a sensitivity analysis to examine how sample size varies according to changes in the values of the between-subject and within-subject variances.

Finally, toward the end of the section we discussed sample size calculations for longitudinal studies where the main comparison of interest is a summary outcome. Suppose that investigators are interested in comparing two treatments, an active treatment and control, in terms of the *average* post-baseline response over time. The investigators plan to randomize an equal number of subjects ($N_1 = N_2$, with $\pi = 0.5$) to receive either of the two treatments. They plan to take five repeated measurements of the response, one at baseline, and the remainder at 6-month intervals until the completion of the study. The response variable, Y_{ij}, is continuous and assumed to have an approximate normal distribution. At the completion of the study, the two treatment groups are to be compared in terms of the mean response over the duration of post-baseline follow-up, $C_i = \frac{1}{4}\sum_{j=2}^{5} Y_{ij}$.

Suppose that the investigators want to detect a minimum treatment effect of $\delta = 1.0$, that is, an average difference between the treatment and control groups over time of no less than 1 unit. Based on historical data from similar populations, the investigators posit that the variance of the baseline response is approximately 8.0. They are less certain about the magnitude of the correlation between pairs of repeated measures but conjecture that the correlation is likely to be in the range 0.4 to 0.8. They also conjecture than the variability of the response can be assumed to be approximately constant over the duration of the study. The investigators desire to have power of 90% when conducting a 2-sided test at the 5% significance level.

Given these specifications, it is instructive to compare the sample sizes required for (case 1) a test for group differences in C_i using a standard t-test, and (case 2) a test for group differences in C_i using ANCOVA, with baseline (Y_{i1}) as a covariate. For this illustration we assume $\text{Var}(Y_{ij}) = \sigma^2 = 8$, for $j = 1, ..., 5$, and $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$, for all $j \neq k$, with ρ in the range 0.4 to 0.8. Using the general sample size formula for the comparison of two groups in terms of a linear summary statistic (see equation (20.5) in Section 20.3.1; also see Section 20.3.3), the required sample size is

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2\,\sigma_C^2}{\pi(1-\pi)\delta^2} = \frac{(1.96 + 1.282)^2\,4\,\sigma_C^2}{1.0} = 42.04\,\sigma_C^2,$$

where

$$\sigma_C^2 = \text{Var}(C_i) = \frac{1}{n-1}\{1 + (n-2)\rho\}\sigma^2 = \frac{8}{4}(1 + 3\,\rho),$$

in case 1, and

$$\begin{aligned} \sigma_C^2 &= \text{Var}(C_i)\,[1 - \{\text{Corr}(Y_{i1}, C_i)\}^2] \\ &= \frac{(1-\rho)\{1 + (n-1)\rho\}\,\sigma^2}{(n-1)} = \frac{8}{4}\,(1-\rho)(1 + 4\,\rho), \end{aligned}$$

in case 2.

The projected total sample sizes, as a function of varying values of ρ, are presented in Table 20.2. For analysis without any adjustment for baseline, the required sample size *increases* with increasing correlation. For example, with $\rho = 0.4$, the required sample size is 186 subjects, whereas with $\rho = 0.8$, the required sample size is 286 subjects, approximately 50% larger. In general, when estimating and comparing the *time-averaged* response, the sample size required increases with increasing magnitude of the positive correlations among the responses. This can be explained by the fact that the variability of the summary score is increasing in ρ. The benefits of adjustment for baseline response are apparent in Table 20.2. When $\rho = 0.5$ a total of 128 subjects are required to achieve power of at least 90%. This is almost half the

Table 20.2 Sample size (N), as a function of the pairwise correlation among repeated measurements, for between-group comparison of time-averaged outcome based on t-test, and ANCOVA, adjusting for the baseline response.

	Correlation (ρ)				
Method	0.4	0.5	0.6	0.7	0.8
t-test	186	212	236	262	286
ANCOVA	132	128	116	96	72

Note: Sample size (N) to ensure power of 0.90 when conducting a 2-sided test at the 5% significance level ($\alpha = 0.05$).

number required by the analysis that fails to adjust for baseline response. The relative magnitude of the gain in power of ANCOVA over the t-test increases rapidly with ρ. When $\rho = 0.7$ a total of 96 subjects are required; this is approximately a third of the number of subjects required by the t-test. Also, unlike the t-test, for the ANCOVA analysis the required sample size decreases with increasing correlation. This can be explained by the fact that adjusting for baseline response removes a fraction of the between-subject variability in the summary score; moreover, the fraction of variability removed from the between-subject comparison is an increasing function of ρ. In general, when the design of a longitudinal study ensures no expected difference between the groups at baseline (see Sections 5.6 and 5.7), making an adjustment for baseline response yields a substantial power advantage.

20.4 SAMPLE SIZE FOR A LONGITUDINAL BINARY RESPONSE

Sample size determination for longitudinal studies with a binary response variable is somewhat more complicated. Complications arise from two main sources: (1) the non-linear link function (e.g., logit) usually adopted for the relationship between the mean response and covariates, and (2) the dependence of the variance on the mean. In general, simple closed-form expressions for sample size (and power), comparable to those for a continuous response, cannot be derived. Instead, precise determination of sample size and power involves more complicated procedures that, in general, require inversion of matrices. However, in this section we outline an approximate sample size (and power) formula that does yield a closed-form expression. The formula is based on the GEE approach under a "working independence" assumption (albeit with inference based on the empirical or "sandwich" variance estimator). In general, this approach can potentially yield conservative estimates of sample size, in the sense that the projected sample size may yield greater power than the nominal level; for balanced longitudinal designs, however, the formula is quite accurate.

20.4.1 Sample Size for Comparison of Change in Response

Similar to the derivation for a continuous response, we suppose that investigators are interested in comparing two treatments, say an active treatment and control, in terms of changes in the mean of the binary response (or probability of success) over time. The investigators plan to randomize a total of N subjects, with $N_1 = \pi N$ in group 1, and $N_2 = (1 - \pi)N$ in group 2. They plan to take n repeated measurements of the response at a common set of occasions, $t_1, ..., t_n$ (not necessarily equally spaced measurements). At the completion of the study, the two treatment groups are to be compared in terms of changes in the success probabilities over the duration of the study. For simplicity we assume that changes in the success probabilities can be expressed in terms of a linear trend in the log odds and that the treatment effect can be expressed in terms of the difference in slopes, say δ. Specifically, we assume that

$$\text{logit}\{\Pr(Y_{ij} = 1|\text{Group}_i = g)\} = \beta_1^{(g)} + \beta_2^{(g)} t_j, \quad g = 1, 2;$$

where $\text{Group}_i = 1$ if the i^{th} individual was assigned to the active treatment, and $\text{Group}_i = 2$ otherwise. In this model, $\beta_2^{(1)}$ is the slope, or constant rate of change in the log odds of success over time, in the active treatment group, while $\beta_2^{(2)}$ is the slope in the control group. That is, $\beta_2^{(1)} - \beta_2^{(2)}$ has interpretation in terms of a treatment group difference in the slopes and corresponds to the definition of δ given earlier. Under the null hypothesis of no treatment difference, that is, no treatment × linear trend interaction, H_0: $\delta = \beta_2^{(1)} - \beta_2^{(2)} = 0$.

Next we present expressions for sample size and power, based on GEE methods, that are similar to the corresponding formulas in the univariate setting. Letting $\widehat{\delta} = \widehat{\beta}_2^{(1)} - \widehat{\beta}_2^{(2)}$ denote the GEE estimate of the difference is slopes, a test of H_0: $\delta = 0$ can be based on the following z-test,

$$Z = \frac{\widehat{\delta}}{\sqrt{\text{Var}(\widehat{\delta})}} = \frac{\widehat{\beta}_2^{(0)} - \widehat{\beta}_2^{(1)}}{\sqrt{\text{Var}(\widehat{\beta}_2^{(1)}) + \text{Var}(\widehat{\beta}_2^{(2)})}},$$

where $\text{Var}(\widehat{\beta}_2^{(g)})$ is the variance of the slope estimator for the g^{th} group. A formula for the approximate total sample size, N, is then given by

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \{\nu_1/\pi + \nu_2/(1-\pi)\}}{\delta^2}, \tag{20.7}$$

where $\nu_1 = N_1 \times \text{Var}(\widehat{\beta}_2^{(1)})$ and $\nu_2 = N_2 \times \text{Var}(\widehat{\beta}_2^{(2)})$. To apply this formula, we require expressions for ν_g, for $g = 1, 2$.

Recall that $\widehat{\beta}_2^{(g)}$ is the GEE estimate of the slope in the g^{th} group. To determine ν_g, we must make assumptions about the nature and magnitude of the correlations among the repeated binary responses. Let R denote an $n \times n$ matrix of pairwise correlations among the repeated binary responses. The components of R are $\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik})$; R is assumed to be the same in both groups. Then it can be shown that ν_g is given by the lower-diagonal element (2nd row, 2nd column) of the following 2 × 2 matrix:

$$\left[\begin{pmatrix} \sigma_1^{(g)}, & \ldots, & \sigma_n^{(g)} \\ t_1 \times \sigma_1^{(g)}, & \ldots, & t_n \times \sigma_n^{(g)} \end{pmatrix} R^{-1} \begin{pmatrix} \sigma_1^{(g)} & t_1 \times \sigma_1^{(g)} \\ \vdots & \vdots \\ \sigma_n^{(g)} & t_n \times \sigma_n^{(g)} \end{pmatrix}\right]^{-1}, \qquad (20.8)$$

where $\sigma_j^{(g)} = \sqrt{\mu_j^{(g)}(1 - \mu_j^{(g)})}$ denotes the standard deviation of the binary response in the g^{th} group at the j^{th} occasion, $\mu_j^{(g)}$ denotes the probability of success in the g^{th} group at the j^{th} occasion, and R^{-1} is the inverse of the $n \times n$ correlation matrix R. The 2×2 matrix given by (20.8) is derived from the "model-based" variance estimator for the GEE with R as the "working correlation" assumption; see Section 13.2 in Chapter 13. Note, in general, there is no simple closed-form expression for ν_g based on (20.8) because it involves inversion of R, an $n \times n$ matrix. Calculating the inverse of a 2×2 matrix is relatively straightforward and there is a well-known analytic formula; however, for a matrix with three or more rows and columns, we require the use of computers to find its inverse. Therefore ν_g can only be obtained from (20.8) with the aid of computer software with matrix algebra functions. However, in Section 20.6, using some statistical skullduggery, we describe how ν_g can also be obtained via application of GEE to a "pseudo–dataset" that has been created to have the assumed structure for the mean response over time.

Although there is no simple expression for ν_g based on (20.8), a closed-form expression can be obtained when $\widehat{\beta}_2^{(g)}$ is the GEE estimate of the slope under a "working independence" assumption for the correlation among the repeated binary responses. When a "working independence" assumption is made for the purpose of estimation, note that ν_g still depends on the true correlation matrix, R. With a "working independence" GEE, the following closed-form expression for ν_g is obtained,

$$\nu_g = \frac{\sum_{j=1}^n \sum_{k=1}^n \rho_{jk} \times \sigma_j^{(g)} \times \sigma_k^{(g)} \times (t_j - \bar{t}^{(g)}) \times (t_k - \bar{t}^{(g)})}{\{\sum_{j=1}^n (\sigma_j^{(g)})^2 \times (t_j - \bar{t}^{(g)})^2\}^2}, \qquad (20.9)$$

where

$$\bar{t}^{(g)} = \frac{\sum_{j=1}^n (\sigma_j^{(g)})^2 \times t_j}{\sum_{j=1}^n (\sigma_j^{(g)})^2},$$

is a weighted average of the times of measurement for the g^{th} group (with weights that depend on the variance at each occasion). Although at first glance this formula for ν_g may appear somewhat daunting, it only involves addition, multiplication, and division of known quantities (t_j, $\mu_j^{(g)}$, and ρ_{jk}); therefore all of the required computations can be done using a pocket calculator or within a spreadsheet. When applying this formula for ν_g, recall that

$$\mu_j^{(g)} = \Pr(Y_{ij} = 1 | \text{Group}_i = g) = \frac{\exp(\beta_1^{(g)} + \beta_2^{(g)} t_j)}{1 + \exp(\beta_1^{(g)} + \beta_2^{(g)} t_j)},$$

and $\sigma_j^{(g)} = \sqrt{\mu_j^{(g)}(1-\mu_j^{(g)})}$, for $g = 1, 2,\ j = 1, ..., n$.

We note that for many balanced longitudinal designs, the closed-form expression for ν_g based on (20.9) is almost identical to the corresponding expression for ν_g based on (20.8). For more complicated and unbalanced longitudinal designs, ν_g based on (20.9) will be larger than ν_g based on (20.8). As a result, reliance on the closed-form expression for ν_g in these more complicated settings will yield conservative estimates of sample size, in the sense that the projected sample size may have greater power than the nominal level. However, for all practical purposes, the differences between sample size and power calculations based on (20.8) and (20.9) will be relatively small for many standard longitudinal designs. We also note that sample size calculations based on both (20.8) and (20.9) have been implemented in a publicly-available SAS macro called `GEESIZE`. (`GEESIZE` is available at `http://www.imbs-luebeck.de/imbs/de/software`, together with detailed documentation and a series of illustrative examples.) The macro, initially developed by Rochon (1998), is very versatile and can be applied to many different longitudinal designs for a range of different types of outcomes (e.g., continuous, binary, and count data). The macro can also incorporate monotone missing data patterns and allows for a flexible family of correlation structures that includes compound symmetry and first-order autoregressive correlation.

20.4.2 Sample Size for Comparison of Time-Averaged Response

So far our discussion of sample size and power has focused on study designs where we are primarily interested in comparing groups in terms of *changes* in the log odds over time. Although less common, sometimes it is of interest to design a longitudinal study where the main comparison of interest is the *average* response probability over the duration of the study. This between-group contrast can be represented in the following marginal logistic regression model,

$$\text{logit}\{\Pr(Y_{ij} = 1|\text{Group}_i = g)\} = \beta_1^{(g)},\ g = 1, 2;$$

where $\text{Group}_i = 1$ if the i^{th} individual was assigned to the active treatment, and $\text{Group}_i = 2$ otherwise. In this model, $\beta_1^{(g)}$ is the log odds of success (assumed constant over the duration of the study) in the g^{th} group. The probability of success in the g^{th} group is

$$\mu^{(g)} = \frac{\exp(\beta_1^{(g)})}{1 + \exp(\beta_1^{(g)})},\ g = 1, 2,$$

and does not depend on measurement occasions. Here $\beta_1^{(1)} - \beta_1^{(2)}$ has interpretation in terms of the log odds ratio and corresponds to the definition of δ given earlier. Under the null hypothesis of no treatment difference, H_0: $\delta = \beta_1^{(1)} - \beta_1^{(2)} = 0$. Sample size (and power) calculation is far more straightforward in this setting because the variance of the response, $\mu^{(g)}(1 - \mu^{(g)})$, no longer depends on measurement occasions. Specifically, we can rely on the following sample size formula:

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \, \sigma_\beta^2}{\delta^2},$$

where

$$\delta = \text{logit}(\mu^{(1)}) - \text{logit}(\mu^{(2)}) = \log\left(\frac{\mu^{(1)}}{1-\mu^{(1)}} \Big/ \frac{\mu^{(2)}}{1-\mu^{(2)}}\right),$$

is the log odds ratio comparing the two groups,

$$\sigma_\beta^2 = \left(\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\rho_{jk}\right)\left\{\frac{1}{\mu^{(1)}(1-\mu^{(1)})\pi} + \frac{1}{\mu^{(2)}(1-\mu^{(2)})(1-\pi)}\right\},$$

and $\rho_{jk} = \text{Corr}(Y_{ij}, Y_{ik})$. When it can be assumed that the correlations are approximately constant, with $\text{Corr}(Y_{ij}, Y_{ik}) = \rho$, for all $j \neq k$, then a simpler expression for σ_β^2 is obtained:

$$\sigma_\beta^2 = \frac{1}{n}\{1 + (n-1)\rho\}\left\{\frac{1}{\mu^{(1)}(1-\mu^{(1)})\pi} + \frac{1}{\mu^{(2)}(1-\mu^{(2)})(1-\pi)}\right\}.$$

However, we remind the reader that assuming equal correlations produces inaccurate estimates of the projected sample size if a more complex correlation structure actually holds.

20.4.3 Example: Longitudinal Study with a Binary Response

To illustrate the application of the sample size formula with binary responses, let us return to the example introduced earlier. Suppose that investigators are interested in comparing two treatments, an active treatment and control, in terms of *changes* in the probability of a binary response over the duration of the study. The investigators plan to randomize an equal number of subjects ($N_1 = N_2$) to receive either of the two treatments. They plan to take five repeated measurements of the response, one at baseline, and the remainder at 6-month intervals until the completion of the study ($\tau = 2$ years), that is, $(t_1, t_2, t_3, t_4, t_5) = (0, 0.5, 1, 1.5, 2)$. We assume that changes in the log odds of response can be expressed (approximately) in terms of a linear trend; the treatment effect is the difference in slopes, say δ.

The investigators assume that the baseline probability of response is approximately 0.3 for both treatment groups. At the end of two years of follow-up, they assume that the probability of response will be relatively unchanged in the control group (0.3) but will be 0.15 in the active treatment group. Based on historical data from similar populations the investigators posit that the correlation among pairs of responses is approximately 0.5. Finally, the investigators desire to have power of 90% when conducting a 2-sided test at the 5% significance level (i.e., $\gamma = 0.1$ and $\alpha = 0.05$).

To calculate the sample size, we need to determine δ and the response probabilities at each occasion in the two groups. Because the two groups are assumed to have the same baseline probability of response, they have common intercepts (when $t_1 = 0$),

$$\beta_1^{(1)} = \beta_1^{(2)} = \text{logit}(\mu_1^{(1)}) = \log(0.30/0.70) = -0.8473.$$

The slope for time (on the log odds scale) in group 1 is

$$\begin{aligned} \beta_2^{(1)} &= \frac{\log\{\mu_5^{(1)}/(1-\mu_5^{(1)})\} - \log\{\mu_1^{(1)}/(1-\mu_1^{(1)})\}}{\tau} \\ &= \frac{\log(0.15/0.85) - \log(0.30/0.70)}{2} = -0.4437. \end{aligned}$$

The corresponding slope for time in group 2 is

$$\begin{aligned} \beta_2^{(2)} &= \frac{\log\{\mu_5^{(2)}/(1-\mu_5^{(2)})\} - \log\{\mu_1^{(2)}/(1-\mu_1^{(2)})\}}{\tau} \\ &= \frac{\log(0.30/0.70) - \log(0.30/0.70)}{2} = 0. \end{aligned}$$

Therefore $\delta = \beta_2^{(1)} - \beta_2^{(2)} = -0.4437$. The intercepts and slopes for the two groups can also be used to derive the response probabilities at each occasion,

$$\mu_j^{(g)} = \frac{\exp(\beta_1^{(g)} + \beta_2^{(g)} t_j)}{1 + \exp(\beta_1^{(g)} + \beta_2^{(g)} t_j)}, \quad g = 1, 2, \ j = 1, ..., 5.$$

For those in group 1,

$$(\mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)}, \mu_4^{(1)}, \mu_5^{(1)}) = (0.300, 0.256, 0.216, 0.181, 0.150),$$

while for those in group 2,

$$(\mu_1^{(2)}, \mu_2^{(2)}, \mu_3^{(2)}, \mu_4^{(2)}, \mu_5^{(2)}) = (0.30, 0.30, 0.30, 0.30, 0.30).$$

Given these specifications for $\mu_1^{(g)}, t_j$, and ρ_{jk}, we can calculate ν_1 and ν_2 from the formula,

$$\nu_g = \frac{\sum_{j=1}^n \sum_{k=1}^n \rho_{jk} \times \sigma_j^{(g)} \times \sigma_k^{(g)} \times (t_j - \bar{t}^{(g)}) \times (t_k - \bar{t}^{(g)})}{\{\sum_{j=1}^n (\sigma_j^{(g)})^2 \times (t_j - \bar{t}^{(g)})^2\}^2},$$

where

$$\bar{t}^{(g)} = \frac{\sum_{j=1}^n (\sigma_j^{(g)})^2 \times t_j}{\sum_{j=1}^n (\sigma_j^{(g)})^2}; \quad \sigma_j^{(g)} = \sqrt{\mu_j^{(g)}(1 - \mu_j^{(g)})}.$$

Here $\rho_{jk} = \rho = 0.5$ for $j \neq k$ and $\rho_{jk} = 1$ for $j = k$. This yields $\bar{t}^{(1)} = 0.8773, \bar{t}^{(2)} = 1.0, \nu_1 = 1.2676$, and $\nu_2 = 0.9524$. Therefore the projected total sample size is

$$\begin{aligned} N &= \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \{\nu_1/\pi + \nu_2/(1-\pi)\}}{\delta^2}, \\ &= \frac{(1.96 + 1.282)^2\ (1.2676/0.5 + 0.9524/0.5)}{(-0.4437)^2} = 237.05. \end{aligned}$$

Thus to ensure that they have power of at least 90%, the investigators will need to enroll a total of 238 subjects, randomizing an equal number (119) to each of the two treatment groups.

Finally, note that a correlation of $\rho = 0.5$ is relatively high for repeated binary responses. For example, given marginal probabilities ranging from 0.15 to 0.30, the maximum possible correlation between any pair of binary responses is constrained to be less than 0.64 (recall that with binary data, the correlations are restricted to ranges that are determined by the marginal probabilities of success; see Section 12.2). It is instructive to re-calculate sample size for smaller values of ρ. For example, with $\rho = 0.3$, the investigators will need to enroll a total of 328 subjects to ensure power of at least 90%. Thus, it is apparent that sample size requirements are sensitive to the magnitude of the assumed correlation among the repeated binary responses.

20.5 SUMMARY

In this chapter we have shown that sample size determination for longitudinal studies can often be simplified so that well-established formulas for the univariate case can be applied. For the investigators the main challenges are in the specification of the minimum effect size of subject-matter interest and in providing a realistic estimate of the anticipated variability in the measurements. The choice of an appropriate effect size must be made on purely subject-matter grounds. If the investigators expect a large effect, then it is likely to be detected with a relatively small sample size. In contrast, detection of small effects requires somewhat larger sample sizes. In planning a study, investigators need to keep their optimism in check, since gross overestimation of the effect size will result in too few subjects and insufficient power to detect somewhat smaller, but nonetheless scientifically important, effects.

Perhaps the greatest challenge facing investigators is to provide a realistic estimate of the variability in the data. This will either require the provision of estimates of both between-subject and within-subject variability or, alternatively, an estimate of the covariance among the repeated measurements. Since scientific studies are rarely conducted in a vacuum, investigators can usually obtain some estimates of the variability based on historical data from related studies with similar populations. Alternatively, in the complete absence of any relevant historical data, it may be prudent to conduct a small pilot study. If there is much uncertainty regarding the anticipated variability in the data, a simple sensitivity analysis, examining the projected sample sizes across a range of plausible values for the variability, should be conducted.

Finally, as mentioned in Chapters 17 and 18, missing data are the rule, not the exception, in longitudinal studies. Therefore it is important to make some adjustment for the potential loss of information due to missing data when planning a longitudinal study, for example, by using relatively conservative estimates of sample size. In general, a consideration of the anticipated fraction of missing data (or the proportion of subjects who drop out before the completion of the study), say f, suggests that the sample size should be inflated by a factor of $\frac{1}{1-f}$. That is, if 15% of the observations are expected to be missing, then investigators should plan on increasing the sample size of the study by a factor of 1.18 (or $\frac{1}{1-0.15}$). Although this adjustment is crude, ignoring both the location of the missing observations and the correlation among repeated measurements, it will probably be adequate for most practical purposes. Failure to make any adjustment for missing data will result in an underestimation of the number of subjects required to attain the desired level of power.

20.6 COMPUTING: SAMPLE SIZE CALCULATION FOR A LONGITUDINAL BINARY RESPONSE USING PSEUDO-DATA

Recall from Section 20.4 that to apply the simple formula for sample size (and power) for a longitudinal binary response,

$$N = \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \{\nu_1/\pi + \nu_2/(1-\pi)\}}{\delta^2},$$

we require an estimate of ν_g, for $g = 1, 2$. In general, there is no closed-form expression for ν_g based on (20.8); instead, ν_g can only be obtained from (20.8) with the aid of computer software with matrix algebra functions. In this section we demonstrate how ν_g can also be obtained via application of GEE (with "working correlation" matrix fixed at the true correlations) to a "pseudo–dataset" that has been created to have the assumed structure for the mean response over time.

Before discussing the construction of the "pseudo–dataset," and to provide some motivation for the method to be described, we note that the required term in the numerator of the sample size formula, $\{\nu_1/\pi + \nu_2/(1-\pi)\}$, can be re-expressed as

$$\begin{aligned}\{\nu_1/\pi + \nu_2/(1-\pi)\} &= \frac{N_1 \operatorname{Var}(\widehat{\beta}_2^{(1)})}{\pi} + \frac{N_2 \operatorname{Var}(\widehat{\beta}_2^{(2)})}{(1-\pi)} \\ &= N \operatorname{Var}(\widehat{\beta}_2^{(1)}) + N \operatorname{Var}(\widehat{\beta}_2^{(2)}) \\ &= N \operatorname{Var}(\widehat{\delta}).\end{aligned}$$

Because $\operatorname{Var}(\widehat{\delta})$ is the variance of the estimator of δ based on N subjects, $\{\nu_1/\pi + \nu_2/(1-\pi)\}$ can be thought of as the variance of the estimator of δ if based on a single representative subject instead of a sample of N subjects. To be representative of both groups, this single subject belongs to group 1 with weight π and belongs to group 2 with weight $(1-\pi)$. By creating a "pseudo–dataset" that contains repeated measures on a single representative subject, $\{\nu_1/\pi + \nu_2/(1-\pi)\}$ can then

Table 20.3 Illustrative commands in SAS for inputing variables from a "pseudo–dataset".

```
DATA=pseudo;
    INPUT id wt group time y;
    one=1;
    DATALINES;
    1  0.5  1  0    0.30
    1  0.5  1  0.5  0.2555697756
    1  0.5  1  1    0.2156921486
    1  0.5  1  1.5  0.1805278643
    1  0.5  1  2    0.15
    2  0.5  2  0    0.30
    2  0.5  2  0.5  0.30
    2  0.5  2  1    0.30
    2  0.5  2  1.5  0.30
    2  0.5  2  2    0.30
    ;
```

be obtained by squaring the reported standard error for the estimate of δ obtained from the analysis of the "pseudo–dataset". Specifically, for each group we create n "pseudo-observations" for the n repeated measures, where the response variable is set equal to the mean at the n occasions. The n "pseudo-observations" for group 1 receive weight of π, whereas the n "pseudo-observations" for group 2 receive weight of $(1-\pi)$. In the "pseudo–dataset" we also include the covariates required for the planned analysis, such as an indicator of group and a variable for the time of measurement. By including a weight variable, set equal to π for the n observations in group 1 and to $(1-\pi)$ for the n observations in group 2, we can allow for possible unequal allocation of subjects to the two groups.

The "pseudo–dataset" for the example of a longitudinal study with a binary response from Section 20.4 is displayed in Table 20.3. Recall that in this example there are two treatment groups with equal allocation of subjects to each group $(\pi = 0.5)$. Five repeated measurements of the binary response are planned, one at baseline, and the remainder at 6-month intervals until the completion of the study, i.e, $(t_1, t_2, t_3, t_4, t_5) = (0, 0.5, 1, 1.5, 2)$. The baseline probability of response is assumed to be 0.3 for both treatment groups, yielding common logistic regression intercepts $\beta_1^{(1)} = \beta_1^{(2)} = -0.8473$. At the end of two years of follow-up, it is assumed that the probability of response will be unchanged in the control group (0.3), but will be 0.15 in the active treatment group. This yields a slope for time in group 1 of $\beta_2^{(1)} = -0.4437$, whereas the corresponding slope in group 2 is $\beta_2^{(2)} = 0$; thus

Table 20.4 Illustrative commands for fitting a marginal logistic regression model, with fixed within-subject correlations, to the "pseudo-observations" using PROC GENMOD in SAS.

```
PROC GENMOD DATA=pseudo DESCENDING;
    CLASS id group;
    MODEL y/one=group time group*time / DIST=BIN LINK=LOGIT SCALE=1;
    WEIGHT wt;
    REPEATED SUBJECT=id / MODELSE
    TYPE=FIXED(1 .5 .5 .5 .5
               .5 1 .5 .5 .5
               .5 .5 1 .5 .5
               .5 .5 .5 1 .5
               .5 .5 .5 .5 1);
```

$\delta = \beta_2^{(1)} - \beta_2^{(2)} = -0.4437$. The intercepts and slopes for the two groups can be used to derive the response probabilities at the five occasions. For those in group 1,

$$(\mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)}, \mu_4^{(1)}, \mu_5^{(1)}) = (0.3000, 0.2556, 0.2157, 0.1805, 0.1500),$$

while for those in group 2,

$$(\mu_1^{(2)}, \mu_2^{(2)}, \mu_3^{(2)}, \mu_4^{(2)}, \mu_5^{(2)}) = (0.30, 0.30, 0.30, 0.30, 0.30).$$

The means at each occasion in the two groups are the "pseudo-responses" in Table 20.3. Associated with each "response" are the covariates and the weight variable (set equal to π for observations in group 1 and to $(1 - \pi)$ for observations in group 2).

By conductung a GEE analysis of the "pseudo–dataset," under a fixed structure for the correlation among the repeated measures, we obtain both an estimate of δ and its standard error. The square of the standard error provides an estimate of $\{\nu_1/\pi + \nu_2/(1 - \pi)\}$ that can then be plugged into the sample size (and power) formula. Illustrative SAS commands for using PROC GENMOD to fit a marginal logistic regression model to the "pseudo-observations" are presented in Table 20.4. Note that because the response variable is a proportion rather than a binary response, we must use the `events/trials` syntax (albeit setting the number of trials equal to 1); see Section 13.6 for more details on the command syntax for PROC GENMOD. PROC GENMOD in SAS allows the "working correlation" matrix, R, to be fixed; here, it is assumed that there is constant correlation among the five binary responses, with $\rho = 0.5$ (this assumption can easily be relaxed and a more general correlation structure can be specified). Finally, because the "working correlation" matrix is assumed to

Table 20.5 Estimated regression coefficients and model-based standard errors from the analysis of the "pseudo-observations" in Table 20.3.

Variable	Estimate	SE	Z
Intercept	−0.8473	2.7603	−0.31
I(Group = 1)	0.0000	3.9158	0.00
t_j	0.0000	1.3801	0.00
I(Group = 1) × t_j	−0.4437	2.1071	−0.21

be correctly specified, we request that standard errors for the estimated regression parameters are obtained using the "model-based" (MODELSE) variance estimator. We caution that standard errors based on the "sandwich" estimator, the default for many software packages, cannot be used for estimating $\{\nu_1/\pi + \nu_2/(1-\pi)\}$.

The results of fitting a marginal logistic regression model, using GEE with a fixed correlation matrix, to the "pseudo-observations" are presented in Table 20.5. The estimated intercept of -0.8473 corresponds to the common baseline log odds of success in the two groups (and hence the estimated group effect is zero). The estimated effect of time, the slope in group 2, is zero because, by construction, the log odds are constant over time in group 2. The estimated group × time interaction effect, -0.4437, corresponds to the difference in slopes in the two groups; this is an estimate of δ. The corresponding standard error of the estimate of δ, 2.1071, provides an estimate of $\sqrt{\nu_1/\pi + \nu_2/(1-\pi)}$. That is, $\{\nu_1/\pi + \nu_2/(1-\pi)\} = (2.1071)^2 = 4.4399$. We can then plug this value into the sample size formula to obtain

$$\begin{aligned} N &= \frac{\{Z_{(1-\alpha/2)} + Z_{(1-\gamma)}\}^2 \{\nu_1/\pi + \nu_2/(1-\pi)\}}{\delta^2}, \\ &= \frac{(1.96 + 1.282)^2\,(4.4399)}{(-0.4437)^2} = 237.04. \end{aligned}$$

Thus, to ensure that they have power of at least 90%, the investigators will need to enroll a total of 238 subjects, randomizing an equal number (119) to each of the two treatment groups. Recall that this is the same projected number of subjects obtained using the closed-form expression based on (20.9) instead of (20.8).

Finally, using the pseudo-data it is relatively straightforward to assess how sample size (and power) are sensitive to assumptions about the strength (or nature) of the correlation among the repeated measures. To do so only requires fixing the correlation matrix to a set of alternative values. For example, if the correlation is assumed to be constant but with $\rho = 0.3$, this yields a model-based standard error of the estimate of δ of 2.4783. When this value is squared and plugged into the sample size formula, we obtain

$$N = \frac{(1.96 + 1.282)^2 \, (6.1420)}{(-0.4437)^2} = 327.91.$$

Thus, to ensure that they have power of at least 90%, the investigators would need to enroll a total of 328 subjects if the correlation is 0.3 instead of 0.5.

20.7 FURTHER READING

Schlesselman (1973a, b), in two companion papers on the design of longitudinal studies, discusses sample size calculations and issues concerning study duration and the frequency of measurement; also see Raudenbush and Liu (2001). Muller et al. (1992) provide a comprehensive approach to the calculation of statistical power for longitudinal studies with a continuous outcome. Snijders and Bosker (1993) present sample size formulas for mixed effects models for longitudinal data. Overall and Doyle (1994) provide sample size formulas based on simple composites or contrasts among the repeated measurements. Hedeker et al. (1999) discuss sample size estimation for longitudinal study designs with a continuous outcome and allow for attrition and a variety of covariance structures for the repeated measurements. Basagana and Spiegelman (2010) discuss power and sample size calculations for longitudinal studies estimating the effect of a time-varying covariate.

The closed-form expression for sample size estimation for longitudinal study designs with a binary outcome presented in Section 20.4 is derived in Jung and Ahn (2005); they also derive a more general formula incorporating missing data patterns.

Bibliographic Notes

Lipsitz and Fitzmaurice (1994) describe a method, based on generalized least squares, for calculating sample size for longitudinal studies with binary responses. Pan (2001) derives explicit formulas for sample size and power calculations for two-group studies with correlated binary responses. A very general method for computing sample size and statistical power for longitudinal studies, based on the generalized estimating equations approach, has been developed by Liu and Liang (1997); this method does not, in general, yield closed-form expressions for sample size and power, but the method can be implemented numerically.