5

Studies With Multiple Measures Within a Unit

5.1 INTRODUCTION

Many health research studies involve multiple measurements within a basic study unit. One example is the repeated measures study, where repeated measurements of a given outcome variable are taken on the same set of individuals at a series of points in time. Another example is the multiple endpoint study, where several related outcome variables are measured for each individual, and it is desired to perform an integrated analysis incorporating all of the outcome measures. This could involve, for example, measurements of severity on several possible disease symptoms.

Another example is the study involving subunits nested within each basic unit, such as similar parts of the body within patients (e.g. eyes, teeth, joints, and so on), children within schools, or residents within communities. Included in this framework is the cluster randomization (or group randomization trial), where randomization is performed at the level of some aggregate, such as schools or communities (Murray, 1998; Donner and Klar, 2000; Zucker, 2004). Cluster randomization is employed by necessity in trials where the intervention under test is delivered on a group basis, for example, an educational program or a community outreach program. Cluster randomization is also sometimes employed for administrative convenience in trials where the intervention is delivered at the level of the individual; see, for example, Simon (1981).

The main issue of concern for sample size calculation in studies with multiple measurements within a unit is to take the correlation between the measurements into account. In studies with a cluster structure, this correlation

is generally known as the intraclass correlation or intra-cluster correlation (ICC). The correlation influences the variance of estimated between-group differences. When the correlation is positive, as is usually the case, the variance is larger than it would be if the measurements were independent. Hence failure to account for the correlation can lead to a underestimated sample size, and thus to an underpowered study.

This chapter is devoted to sample size methods with correlated data. We begin with simple methods for simple settings, and move on progressively to more complex situations. We discuss, in succession, the following topics: summary measures analysis, simple mixed ANOVA analysis, simple binary endpoint analysis, repeated measures ANOVA, random line models, multiple endpoint studies, general linear mixed models, and generalized estimating equations (GEE) analysis.

5.2 SUMMARY MEASURES ANALYSIS

The simplest approach to analyzing studies with multiple observations within each unit is to reduce each unit's data to a single summary measure, such as a mean or an area under the curve. Summary measures analysis is discussed by Matthews, Altman, Campbell, and Royston (1990) in a repeated measures context. The summary measures approach reduces the original multivariate problem to a univariate one. The analysis then may proceed using standard methods for univariate data, such as two-sample t-tests, analysis of variance, and so on.

As regards sample size calculation, two situations may be identified. One situation is where data are available from past studies whose structure matches that of the planned study. For example, in the planning of a repeated measurements study, prior studies may be found with an identical or highly similar study visit schedule. In this case, data on the summary measure in these prior studies bear directly on the behavior of the summary measure in the planned study. Thus the sample size calculation can be carried out using univariate data methods, as in Chapter x. The other situation is where the relevant prior studies involve a different structure than intended for the planned study. In this case, the sample size calculation can be carried out using the methods described in Sec. 5.8 below.

In many situations, the data structure may be different for different units, either for intrinsic reasons (e.g. different heart bypass patients having a different number of bypasses) or due to missing data. When this is the case, a stratified summary measures analysis can be conducted, with stratification according to the different possible structures (Dawson and Lagakos, 1993). Sample size calculation for this case can also be carried out using the methods of Sec. 5.8.

5.3 SIMPLE MIXED ANOVA ANALYSIS

5.3.1 The basic procedure

We consider here the situation where it is desired to compare two or more groups with respect to a continuous outcome measure, with multiple observations on the outcome within each unit (e.g. multiple time points or subunits within units). The multiple observations are assumed exchangeable. We let Y_{ijk} denote the response for observation ijk, with i indexing the groups, j indexing the units within a group, and k indexing the observations within a unit.

The typical analysis model for such data is the mixed ANOVA model

$$Y_{ijk} = \mu + a_i + b_{j(i)} + \epsilon_{k(ij)}, \tag{5.1}$$

where μ represents the overall response level, the a_i's are fixed parameters representing the group effects, $b_{j(i)}$ is a random unit effect term assumed distributed $N(0, \sigma_b^2)$, and $\epsilon_{k(ij)}$ is an observation-specific error term assumed distributed $N(0, \sigma_\epsilon^2)$, with all random variables assumed independent. The variance of an individual response Y_{ijk} is $\tau^2 = \sigma_b^2 + \sigma_\epsilon^2$. The intraclass correlation coefficient (ICC) is given by $\rho = \sigma_b^2/\tau^2$.

In sample size calculation, it is often assumed that the number of observations in a unit is equal to a common value m for all units. In this case, the foregoing ANOVA procedure is equivalent to a summary measures analysis, involving a t-test or a classical one-way ANOVA analysis on the unit means $\bar{Y}_{ij\cdot}$. We have

$$\sigma^2 \equiv \mathrm{Var}(\bar{Y}_{ij\cdot}) = \frac{\tau^2}{m}(1 + (m-1)\rho), \tag{5.2}$$

which is the standard variance for a sample mean multiplied by an inflation factor $IF = (1 + (m-1)\rho)$ arising from the intraclass correlation. Given this expression for σ^2, the sample size can be computed using the methods for univariate data described in Sec. x.x and x.x. For instance, if the study involves two groups (test and control) and the asymptotic normal formula is used, then the total number N of units required is given by

$$N = [\pi(1-\pi)]^{-1} m^{-1} [1 + (m-1)\rho](z_{\alpha/2} + z_\beta)^2 \tau^2/\delta^2, \tag{5.3}$$

where π denotes the proportion of units allocated to the test group, δ denotes the mean difference we desire to detect, and z_p denotes the p-th quantile of the standard normal distribution. If the number of clusters is small, then the method based on the noncentral t-distribution should be used (Section x.x).

With this approach, it is possible to estimate τ^2 and ρ from past studies that are not necessarily identical in structure to the planned study. For instance, data can be used from studies that involved a different number of repeated measurements or a different number of subunits per unit than that anticipated for the planned study.

If it is desired to take into account the possibility that different units may have different numbers of observations, say m_{ij} observations in unit ij, the sample size calculation becomes significantly more complex. Firstly, some specification is needed of the anticipated distribution of the m_{ij} across the units. Secondly, the mixed ANOVA analysis no longer reduces to a simple summary measures analysis. Several analysis approaches are possible. The state of the art approach is iterative maximum likelihood or restricted maximum likelihood fitting. A detailed discussion of sample size calculation within this framework is provided in Sec. 5.8 below. There we give a discussion of sample size calculation when it is possible to identify different strata with different observation patterns. In the present case, the strata would be defined by the number of observations m_{ij} within the unit.

Example: We present the sample size calculation for the the Child and Adolescent Trial for Cardiovascular Health (CATCH) (Luepker et al., 1996; Zucker et al., 1995). This study investigated a school-based intervention aimed at promoting heart-healthy habits in elementary school children. Because the intervention was run at the school level, a cluster randomization design was required. The trial involved randomization of 96 schools, with 40 assigned to to the control group (C), 28 assigned to receive the school-based intervention (S), and 28 assigned to receive the school-based intervention plus a supplementary family-based intervention (S+F). The main trial comparison was between the combination of S and S+F groups and the control group. The primary trial endpoint was taken to be serum total cholesterol, because it was felt that dietary and exercise habit measures would be open to reporting bias, whereas cholesterol would be free of such bias but responsive enough to reflect true diet and exercise changes. Various diet, exercise, and other endpoints were included as secondary endpoints.

The sample size was determined by the following slightly extended version of the formula (5.3):

$$N = [\pi(1-\pi)]^{-1}[CS]^{-1}[1+(S-1)\rho_1+(C-1)S\rho_2](z_{\alpha/2}+z_\beta)^2\tau^2/\delta^2, \quad (5.4)$$

where C denotes the number of classrooms in a school, S denotes the number of students per classroom with available data, ρ_1 denotes the within-classroom correlation, and ρ_2 denotes the within-school correlation for students in different classrooms in the same school.

The CATCH calculations assumed 3-4 classrooms per school ($C = 3.5$ was used as an average figure) and $S = 17$ students with available data per class. From past studies, the standard deviation τ was estimated to be 28 mg/dl. The correlations ρ_1 and ρ_2 were estimated from a variance components analysis of a small data set from a prior cohort study at one of the CATCH study centers. The estimates were ρ_1=0.023 and ρ_2=0.003. The projected treatment difference on cholesterol (S and S+F vs. C) was determined to be 5.1 mg/dl, based on the program's dietary targets and published results on the relationship between dietary constituents and blood cholesterol levels. A

conservative adjustment factor was incorporated to account for possible loss to follow-up bias; see Zucker et al. (1995) for details. The effect of the adjustment was to reduce the difference δ to be detected from 5.1 mg/dl to an effective difference, after the adjustment, of 2.9 mg/dl. The intervention (S+F and S) to control (C) allocation ratio was 7:5 (to enhance the power of the S+F vs. S comparison), so that π=0.583. Substituting these parameters into the formula (5.4) yields a sample size requirement of 102 schools total for 90% power at the two-sided 0.05 level. Based on administrative considerations, the final sample size was taken to be 96 schools total.

The sample size calculation incorporated both school and classroom effects because at the design stage it was felt important to do so. As regards the analysis, with school-level assignment the randomization theory of statistical testing requires school to be the primary unit of analysis. However, it is not necessary to adjust for classroom for the analysis to be valid in terms of Type I error (under the "strong" null hypothesis that the intervention has absolutely no effect on any child). In fact, the final CATCH analysis did not adjust for classroom.

5.3.2 Some remarks

When the number of clusters is small, the preferred method of analysis will typically be a permutation test applied to the cluster means (Braun and Feng, 2001; Zucker, 2004). Nonetheless, power calculations using the noncentral t or F distribution will generally be reasonable for the purpose of sample size specification.

As usual, to carry out the sample size calculations, preliminary estimates of the relevant parameters are needed. The estimate of τ^2 is typically based on past data on the endpoint of interest. If data are available from a prior study involving the endpoint of interest and a cluster structure similar to that of the planned study, even if the number of observations per unit is not the same, such data can be used to estimate the ICC. This can be done by performing a variance components analysis to estimate σ_b^2 and σ_ϵ^2, and then converting these to estimates of τ^2 and ρ, as in the foregoing example. If no past data arising from the relevant cluster structure are available, then an educated guess of the ICC will have to be made by using an ICC value available in the literature for an endpoint whose ICC is judged likely to be similar to that for the endpoint of interest. Donner and Klar (1994) have presented ICC's for a range of public health settings.

In repeated measures studies, there is a trade-off between the number N of units entered and the number m of observations per unit. Given estimates of τ^2 and ρ, different combinations of N and m can be considered, with a power estimate presented for each combination. The combination most appealing to the investigators can then be chosen.

In cluster randomization studies, the number m of individuals per cluster is often fixed by circumstances (e.g. a typical standard classroom size). But in

some cases, both N and m are in the control of the investigators, in which case the above-described trade-off between N and m can be explored. Relatedly, if the ICC is moderate to large, substantial effort and cost can be saved with modest loss of efficiency by subsampling individuals within the clusters. Let m denote the total cluster size. Then, for given ρ, the relative efficiency (RE) of measuring only m' individuals in each cluster as compared with measuring all m individuals is given by

$$RE = \left(\rho + \frac{1-\rho}{m}\right) \Big/ \left(\rho + \frac{1-\rho}{m'}\right).$$

The number of clusters N needed if m' individuals in each in each cluster are measured is $1/RE$ times the number of clusters needed if all m individuals in each cluster are measured. Often ρ is small, and so it will be worthwhile to measure all individuals in each cluster. However, when ρ is moderate to large, the cost of adding further clusters may be offset by the reduced measurement cost engendered by sampling.

5.4 SIMPLE BINARY ENDPOINT ANALYSIS

For a binary (0–1) response representing the occurrence of some event, the ANOVA model (5.1) is not applicable, but an approach similar in form can followed. Here $p_i = E[Y_{ijk}]$ is the event probability for an individual in study arm i and the intraclass correlation coefficient $\rho = \text{Cov}(Y_{ijk}, Y_{ijk'})/\text{Var}(Y_{ijk})$, $k \neq k'$, becomes the Cohen (1960) kappa coefficient, which can be expressed as

$$\rho = [\Pr(Y_{ijk} = 1, Y_{ijk'} = 1) - p_i^2]/[p_i(1-p_i)]. \tag{5.5}$$

Assuming the number of subunits in each unit is equal to a common value m, and that the number of units in each group is large enough to justify a normal approximation, one can as before analyze the data by applying a t-test or ANOVA analysis to the cluster means $\bar{Y}_{ij\cdot}$. We have

$$\text{Var}(\bar{Y}_{ij\cdot}) = m^{-1} p_i(1-p_i)(1+(m-1)\rho). \tag{5.6}$$

Thus, for a two-group study for example, the total number of units required is

$$N = \frac{1}{m}\left(\frac{p_0(1-p_0)/\pi_0 + p_1(1-p_1)/\pi_1}{(p_1-p_0)^2}\right)(z_{\alpha/2} + z_\beta)^2[1+(m-1)\rho], \tag{5.7}$$

where π_i is the proportion of subjects allocated to group i. This development is easily extended to the case of stratified analysis; see, for example, Donner and Donald (1987). The case where the number of subunits can vary from unit to unit is discussed in Sec. 5.9; see especially the example presented there.

Example: We present the sample size calculation for the Jerusalem Handwashing Study (JHS) (Rosen et al., 2005). The goal was to test the efficacy of a

preschool-based handwashing education program in promoting handwashing and preventing illness-related absenteeism among the children. The design called for equal allocation ($\pi_0 = \pi_1 = 0.5$) of preschools to intervention or control. The primary response variable was illness absenteeism. Based on US public source data for a similar population and pilot data collected by JHS investigators, the daily illness absenteeism rate in the control group was projected to be p_0=0.06. The trial was sized to detect a 25% drop in absenteeism, corresponding to a treatment group daily illness absenteeism rate of p_1=0.045, in a trial of length D=60 days, with 80% power at the two-sided 0.05 level. The projected number of children per preschool was $m = 22$. Because the trial involved cluster sampling with multiple observations for each child, two types of correlation were considered: the intraclass correlation (ICC), reflecting the correlation between two children in the same preschool, and the between-day correlation (BDC) for a particular child. These were estimated using data from a previous handwashing researcher, with the ICC estimated at 0.0274 using variance components analysis and the BDC estimated at 0.0548 by comparing the observed variance of absentee rates within each school with the binomial variance that would hold under zero between-day correlation. In view of the repeated observations on each child within each preschool, the formula (5.7) had to be generalized to

$$N = \frac{1}{mD}\left(\frac{p_0(1-p_0)/\pi_0 + p_1(1-p_1)/\pi_1}{(p_1-p_0)^2}\right)(z_{\alpha/2}+z_\beta)^2 \times [1+(m-1)\rho][1+(D-1)\eta],$$

where η denotes the BDC. The resulting calculated sample size was $N = 36$ preschools. In the final trial design, the sample size was set at 40 preschools, to provide a modest safeguard against misspecification of the relevant parameters.

5.5 REPEATED MEASURES ANOVA

The development in the preceding two sections assumed that the observations within a unit are exchangeable. Often there are characteristics that distinguish the observations one from another. The simplest case is that in which the observations fall into distinct classes according to the level of a categorical variable. For instance, in a repeated measures study with a fixed visit schedule for all subjects, the mean response level may change from timepoint to timepoint. As another example, in an ophthalmologic study, the mean response level may differ between the right eye and the left eye. For concreteness, we will frame the presentation below in terms of repeated measures over time, but the same theory applies to the ophthalmologic example and other like cases. We focus on a continuous outcome measure.

As before, we let Y_{ijk} denote the response for observation ijk, with i indexing the groups, j indexing the units within a group, and k indexing the

observation (thus k now corresponds to time). We assume that all individuals are measured at the same timepoints (except for minor jitter in the scheduling of the visits, which we will ignore). Missing data, however, could be allowed. The classical repeated measures ANOVA model is as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_k + \gamma_{ik} + b_{j(i)} + \epsilon_{ijk}. \tag{5.8}$$

Here there are four fixed effect terms: μ represents the general base response level, the α_i represent group effects, the β_k represent time effects, and the γ_{ik} represent group by time interaction effects. In standard ANOVA fashion, the sum of the α_i, the sum of the β_k, and the sum of the γ_{ik} with respect to either index, are all constrained to be equal to zero. The term $b_{j(i)}$ is a random unit effect, assumed $N(0, \sigma_b^2)$, while the ϵ_{ijk} are $N(0, \sigma_\epsilon^2)$ error terms. The random terms are all assumed independent.

The development for this model is very similar to that for the simple ANOVA analysis model of Sec. 5.3. The variance of Y_{ijk} is given by $\tau^2 = \sigma_b^2 + \sigma_\epsilon^2$, and the intraclass correlation is given by $\rho = \sigma_b^2/\tau^2$. We assume here that the number of observations is equal to m for all subjects, with no missing data. The statistical test for assessing the group main effect then reduces again to a standard two-sample t-test or one-way ANOVA on the unit means $\bar{Y}_{i\cdot\cdot}$. The variance of these unit means is as given in (5.2), and the sample size for testing the group main effects is exactly as described in Sec. 5.3.1. The group by time interaction is tested using the F statistic

$$F_{int} = \frac{\text{SSI}/[(I-1)(m-1)]}{\text{SSE}/(N-IJ)}, \tag{5.9}$$

where the interaction and error sums of squares SSI and SSE are given by

$$\text{SSI} = \sum_{i=1}^{I}\sum_{k=1}^{m} n_i(\bar{Y}_{i\cdot k} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot k} + \bar{Y}_{\cdot\cdot\cdot})^2, \tag{5.10}$$

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\sum_{k=1}^{m}(Y_{ijk} - \bar{Y}_{i\cdot k} - \bar{Y}_{ij\cdot} + \bar{Y}_{i\cdot\cdot})^2. \tag{5.11}$$

Here I is the number of groups, N is the total number of units in the study, n_i is the number of units in group i, and the $\bar{Y}$'s denote averages over the indices that are replaced by dots. The random unit effects $b_{j(i)}$ cancel out of the interaction statistic. Under the null hypothesis of no interaction, the statistic F_{int} has an F distribution with $(I-1)(m-1)$ and $N-IJ$ degrees of freedom. Under a general alternative, the statistic F_{int} has a noncentral F distribution with the same degrees of freedom and noncentrality parameter $\eta = \sum_{i=1}^{I}\sum_{k=1}^{m} n_i\gamma_{ik}^2 = N\sum_{i=1}^{I}\sum_{k=1}^{m}\pi_i\gamma_{ik}^2$, where π_i is the fraction of units assigned to group i. The case of greatest interest is the case of equal allocation, with $\pi_i = 1/I$. Sample size for testing interaction, if this hypothesis is of interest, can be determined using the noncentral F method described in Sec. x.x.

When the number of observations can differ from unit to unit due to missing data, the situation becomes considerably more complicated. The relevant methodology for this situation is described below in Sec. 5.8.

5.6 RANDOM LINE MODELS

A common alternate model for studies comparing several groups with repeated measurements over time is the random line model

$$Y_{ijk} = \beta_{1ij} + \beta_{2ij} t_{ijk} + \epsilon_{ijk}. \tag{5.12}$$

Here the index i represents groups, the index j represents individuals, and the index k represents observations within individuals. The t_{ij} represent observation times. The quantities β_{1ij} and β_{2ij} represent individual-specific intercept and slope parameters whose mean differs between the groups. Specifically, it is assumed that $\beta_{1ij} = \mu_1 + \alpha_{1i} + b_{1ij}$ and $\beta_{2ij} = \mu_2 + \alpha_{2i} + b_{2ij}$, where α_{1i} and α_{2i} are fixed parameters representing group effects on the intercept and slope, respectively, while b_{1ij} and b_{2ij} are mean-zero random individual-specific effects and ϵ_{ijk} are mean-zero random error terms. This model is often appropriate, for example, in studies involving chronic degenerative diseases.

It is typically assumed that the random vectors $\mathbf{b}_{ij} = (b_{1ij}, b_{2ij})^T$ are i.i.d. bivariate normal with mean zero and covariance matrix $\mathbf{\Omega}$, and the ϵ_{ijk} are distributed $N(0, \sigma^2_\epsilon)$ independently over i, j, k, with the ϵ_{ijk} independent of the $\mathbf{b}_{ij}$. Within-subject dependence among the observations is reflected in the random effects $\mathbf{b}_{ij}$.

In this section we will focus on the simple balanced design where the measurements are taken at the same set of K times $\{t_k\}$ for all individuals (except for ignorable minor jitter), with no missing data. This is the case dealt with in classic paper of Schlesselman (1973), and is often assumed for simplicity in sample size calculations, even though some deviation from this ideal situation may be expected in the actual study. For the balanced case, the analysis can be carried out as follows. First, individual-specific ordinary least squares (OLS) estimates B_{1ij} and B_{2ij} of the individual-specific intercept and slope are computed. These quantities are then subjected to a standard univariate analysis, such as a two-sample t-test or one-way ANOVA.

The sample size calculation is thus driven by $\mathrm{Var}(B_{1i})$ and $\mathrm{Var}(B_{2ij})$, which are given by

$$\mathrm{Var}(B_{1ij}) = \sigma^2_{\beta_1} + \sigma^2_\epsilon \left[\frac{1}{K} + \frac{\bar{t}^2}{\sum_{k=1}^K (t_k - \bar{t})^2} \right], \tag{5.13}$$

$$\mathrm{Var}(B_{2ij}) = \sigma^2_{\beta_2} + \frac{\sigma^2_\epsilon}{\sum_{k=1}^K (t_k - \bar{t})^2}, \tag{5.14}$$

where $\sigma^2_{\beta_1} = \text{Var}(\beta_{1ij})$, $\sigma^2_{\beta_2} = \text{Var}(\beta_{2ij})$, and $\bar{t} = K^{-1}\sum_k t_k$. An example illustrating this method will be presented in Sec. 5.8, along with an extension designed to account for missing data.

The quantities $\sigma^2_{\beta_1}$, $\sigma^2_{\beta_2}$, and σ^2_ϵ can be estimated in principle by random line analysis of data from a prior study. It is then possible to use formulas (5.13) and (5.14) to explore tradeoffs between changing the sample size and changing the number or timing of the repeat measurements. The sample size can be computed for several possible measurement schedules, and then the investigators can decide what measurement schedule is most appropriate. Typically there will be practical constraints on the measurement schedule; for example, it is not realistic to expect subjects to comply with a visit schedule that calls for too many visits or for visits that are too closely spaced.

Often a sample size calculation will have to be based on journal articles which do not provide sufficient information to estimate $\sigma^2_{\beta_1}$, $\sigma^2_{\beta_2}$, and σ^2_ϵ directly. In this case, some approximation and guesswork is needed. Typically a journal article will report the standard deviation of B_{1ij} and/or B_{2ij} (often only the slope is of interest) and will provide some information about the visit schedule. From the information about the visit schedule it is possible to determine K, $\bar{t}$, and $\sum_k (t_k - \bar{t})^2$ for the study reported in the article. Often the calculation will be approximate because the subjects in the study were followed for different lengths of time, but some reasonable approximation usually can be formed. Next, some sort of guess is needed for σ^2_ϵ. This might be obtained as follows. Some reasonable guess can be made, with the help of the investigators, of the R-squared value for the regression of the outcome measure on time within an individual subject over the average follow-up period in the published study. Call this $\tilde{R}^2$. A estimate could be made of the general level of the slope of the regression, based on the observed mean slope in the study. Call this $\tilde{\beta}_2$. An estimate of σ^2_ϵ then can be obtained by solving

$$\tilde{R}^2 = \frac{\tilde{\beta}_2^2 \sum_{k=1}^K (t_k - \bar{t})^2 + \sigma^2_\epsilon}{\tilde{\beta}_2^2 \sum_{k=1}^K (t_k - \bar{t})^2 + (K-1)\sigma^2_\epsilon} \; . \tag{5.15}$$

Here, the numerator and denominator are, respectively, the expected model sum of squares and the expected total sum of squares for a typical individual subject. With the estimates of the standard deviations of B_{1ij} and B_{2ij}, of K, $\bar{t}$, and $\sum_k (t_k - \bar{t})^2$, and of σ^2_ϵ for the published study, estimates of $\sigma^2_{\beta_1}$ and $\sigma^2_{\beta_2}$ can be obtained by backcalculation from (5.13) and (5.14). These estimates then can be used for sample size calculations for the planned study.

5.7 MULTIPLE ENDPOINT ANALYSIS

The preceding sections have dealt with various situations involving multiple measures on the same endpoint. In this section, we consider the multiple endpoint study. We focus on the case of continuous endpoints. The theory

extends to more general endpoints provided that the number of units per group is large enough for the central limit theorem to take effect. We denote the responses by Y_{ijk}, where i indexes groups, j indexes units, and k indexes endpoints. We assume here that a fixed number K of endpoints is measured on all units, with no missing data.

We denote the vector of all the observations on unit ij by $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijK})^T$. Typically data of this type are analyzed under the assumption that $\mathbf{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, and the key quantities are the within-group mean vectors $\bar{\mathbf{Y}}_i$ and the sample covariance matrix $\mathbf{S}$, given by

$$\mathbf{S} = \frac{1}{N-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)^T,$$

where I is the number of groups, N is the total number of units in the study, and n_i is the number of units in group i. The null hypothesis is $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_I$.

One approach to handling such data is to reduce each unit's data to a single summary measure, such as an overall symptom score, as described in Sec. 5.2. The summary score can be expressed as $W = \mathbf{w}^T\mathbf{Y}$, where $\mathbf{w}$ is a pre-specified vector. Without loss of generality, we will assume that $\mathbf{w}$ is a unit vector with respect to the norm induced by the weighted inner product $\langle \mathbf{w}_1, \mathbf{w}_2 \rangle = \mathbf{w}_1^T \boldsymbol{\Sigma}^{-1} \mathbf{w}_2$. In what follows, we denote this weighted norm by $\| \cdot \|$, and use the term "unit vector" relative to this norm. The specification of $\mathbf{w}$ can be based either on subject-matter considerations (e.g. the linear combination that seems most meaningful from a clinical standpoint) or on statistical considerations. To simplify the discussion, we assume that there are only two groups, and define $\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\hat{\boldsymbol{\Delta}} = \bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2$. Also define $d = \mathbf{w}^T\boldsymbol{\Delta}$ and $\hat{d} = \mathbf{w}^T\hat{\boldsymbol{\Delta}} = \bar{W}_1 - \bar{W}_2$, where $\bar{W}_i$ is the sample mean of W in group i. The null hypothesis becomes $H_0 : \boldsymbol{\Delta} = \mathbf{0}$. The distribution of $\hat{\boldsymbol{\Delta}}$ is $N(\boldsymbol{\Delta}, \psi N^{-1}\boldsymbol{\Sigma})$, where $\psi = [\pi(1-\pi)]^{-1}$ with π denoting the proportion of subjects allocated to group 1. The distribution of $\hat{d}$ is $N(\mathbf{w}^T\boldsymbol{\Delta}, N^{-1}\psi\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})$.

Now suppose that the anticipated alternative is of the form $H_1 : \boldsymbol{\Delta} = \theta\mathbf{u}$, where $\mathbf{u}$ is a specified unit vector expressing the anticipated direction of the between-group difference and θ is a parameter expressing the magnitude of the difference. Then the optimal choice of $\mathbf{w}$, in terms of maximizing power, is $\mathbf{w}^* = \boldsymbol{\Sigma}^{-1}\mathbf{u}/\|\boldsymbol{\Sigma}^{-1}\mathbf{u}\|$. It sometimes happens that $\mathbf{w}^*$ contains negative values. This is a problem because a linear combination of endpoints which includes some negative coefficients is difficult to interpret. One simple solution is simply to delete the endpoints with the negative coefficients or weight all the endpoints equally (Pocock et al., 1987). Alternatively, we can use the unit vector $\mathbf{w}^{**}$ that maximizes the power subject to having nonnegative coefficients. Under the alternative $H_1 : \boldsymbol{\Delta} = \theta\mathbf{u}$, the asymptotic relative efficiency of the statistic based on $\mathbf{w}$ compared with that based on $\mathbf{w}^*$ is given by $\langle \mathbf{w}, \mathbf{w}^* \rangle$, and the unit vector $\mathbf{w}^{**}$ that maximizes this subject to having non-

negative coefficients can be found by quadratic programming (Zucker, 2005). Tang et al. (1993) give some other options for the choice of $\mathbf{w}$.

Whichever $\mathbf{w}$ is chosen, the two-sample t-statistic based on the summary statistic $W = \mathbf{w}^T\mathbf{Y}$ will have a noncentral t distribution with noncentrality parameter $\omega = \sqrt{N\pi(1-\pi)}\theta\mathbf{w}^T\mathbf{u}/\sqrt{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}}$. The sample size thus can be determined using the method of Sec. x.x for the noncentral t distribution, or using the normal sample size formula (x.x).

In practice, the anticipated direction of effect $\mathbf{u}$ can be difficult to specify in advance. It is therefore desirable to set the sample size so as to yield reasonable power over a reasonable range of $\mathbf{u}$. We can specify a range for $\mathbf{u}$ as $\boldsymbol{\mathcal{U}} = \{\mathbf{v} : v_k = (1 + \xi h_k)u_k \ \forall k, \text{ with } |h_k| \leq 1 \ \forall k\}$. This corresponds to allowing up to a $100\xi\%$ relative error in the specification of the anticipated effect on each endpoint. A reasonable value of ξ, e.g. $\xi = 0.20$, can be selected by the study designers. It suffices to examine the 2^K corner points at which $|h_k| = 1 \ \forall k$. Given the chosen $\mathbf{w}$, we can set the sample size so as to yield, for example, 90% power at the anticipated $\mathbf{u}$ and 80-85% power for each of the corner points of $\boldsymbol{\mathcal{U}}$.

As an alternate approach, we can perform an omnibus test for any difference between the groups on any of the endpoints. In the case of an arbitrary number of groups, a test of $H_0 : \boldsymbol{\Delta} = \mathbf{0}$ against the general alternative $H_1 : \boldsymbol{\Delta} \neq 0$ is carried out via multivariate analysis of variance (MANOVA). A number of test statistics for the MANOVA setting are available, including the Wilks lambda statistic, Pillai's trace, the Hotelling-Lawley trace, and Roy's maximum root criterion (Morrison, 1976, Sec. 5.8). We again focus here for simplicity on the case where there are are only two groups. In this case, $H_0 : \boldsymbol{\Delta} = \mathbf{0}$ can be tested against $H_1 : \boldsymbol{\Delta} \neq \mathbf{0}$ using the Hotelling T^2 statistic

$$T^2 = \frac{n_1 n_2}{N}\hat{\boldsymbol{\Delta}}^T\mathbf{S}^{-1}\hat{\boldsymbol{\Delta}}. \tag{5.16}$$

Under H_0, the scaled T^2 statistic $F = (N - K - 1)T^2/[(N - 2)K]$ has an F distribution with K and $N - K - 1$ degrees of freedom. Under a general alternative, F has a noncentral F distribution with the same degrees of freedom and noncentrality parameter $\eta = \boldsymbol{\Delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}$. Sample size calculations then can be carried out using the noncentral F method of Sec. x.x. A preliminary estimate of $\boldsymbol{\Sigma}$ based on prior studies is needed, along with, as usual, a specification of the difference to be detected.

In many settings, particularly in clinical trials, it is of interest to test H_0 against the "one-sided" alternative that $\boldsymbol{\Delta}$ lies in the positive orthant (or, more generally, some subset of the positive orthant). A number of statistical tests for this problem have been described; see, for example, Kudo (1963), Perlman (1969), Geller, Tang, and Gnecco (1989), Tang (1994), Follmann (1996), Bloch, Lai, and Tubert-Bitter (2001), and Perlman and Wu (2004). Analytical power calculation for these tests appears intractable, but in principle power calculations could be carried out by simulation.

5.8 GENERAL LINEAR MIXED MODELS FOR CONTINUOUS DATA

We now proceed to the general linear mixed model setting, a broad setting which covers most of the situations in the preceding sections as special cases. We deal with a general data structure involving a series of basic study units (indexed by j) each containing a number of subunits (indexed by k). Note that here there is no index for groups; different groups can be identified in the model setup as described in the examples below. As noted in the introduction to this chapter, this framework encompasses repeated measures studies, multiple endpoint studies, and studies with a cluster sampling structure.

In this section we deal with a continuous response variable Y. We denote by Y_{jk} the value of observation jk, with $j = 1, \ldots, N$ and $j = 1, \ldots, K_i$. Note that the number of subunits is allowed to vary from unit to unit. We denote by $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jK_j})^T$ the vector of all the observations on unit j. The model is

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j \tag{5.17}$$

where

1. The quantities $\mathbf{X}_j$ and $\mathbf{Z}_j$ are matrices involving explanatory variables.
2. The vector $\boldsymbol{\beta}$ is an vector of unknown fixed regression parameters.
3. The vector $\mathbf{b}_j$ is a vector of random unit-specific terms with distribution $N(\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\zeta}))$, independent over j, where $\boldsymbol{\zeta}$ is a vector of unknown parameters and $\boldsymbol{\Omega}(\boldsymbol{\zeta})$ is a known function of these parameters.
4. The vector $\boldsymbol{\epsilon}_j$ is a vector of random error terms with distribution $N(\mathbf{0},\sigma_\epsilon^2\mathbf{R}_j(\boldsymbol{\xi}))$, independent over j, where $\boldsymbol{\xi}$ is a vector of unknown parameters and $\mathbf{R}_j(\boldsymbol{\xi})$ is a known function of these parameters.
5. The random vectors $\mathbf{b}_j$ and $\boldsymbol{\epsilon}_j$ are independent within each unit j, and the units are all independent of each other.

For example, the simple mixed ANOVA model (5.1) of Sec. 5.3 can be written in the form (5.17) as follows. We take $\boldsymbol{\beta} = [\mu, a_1, \ldots, a_{I-1}]^T$, with I denoting the number of groups. The first column of $\mathbf{X}_j$ is a column of 1's. The $(i+1)$-th column of $\mathbf{X}_j$, for $i = 1, \ldots, I-1$, is equal to 1 if unit j is in group i, -1 if unit j is in group I, and 0 otherwise. The matrix $\mathbf{Z}_j$ is a single column of 1's. The vector $\mathbf{b}_j$ consists of a single random component having distribution $N(0, \sigma_b^2)$. The vector $\boldsymbol{\epsilon}_j$ is distributed as $N(\mathbf{0},\sigma_\epsilon^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix of the appropriate size, i.e. K_j.

As another example, the random line model (5.12) of Sec. 5.6, in the case of two groups, can be written in the form (5.17) as follows. We take $\boldsymbol{\beta} = [\mu_1, \mu_2, \alpha_{11}, \alpha_{12}]^T$. The matrix $\mathbf{X}_j$ is a $K_j \times 4$ matrix. The first column of $\mathbf{X}_j$ is a column of 1's. The second column contains the observation times t_{jk}. The third column is a column of 1's if unit j is in the test group and a column of -1's if unit j is in the control group. In the fourth column, the k-th entry equals t_{jk} if unit j is in the test group and $-t_{jk}$ if unit j is in the control group. The matrix $\mathbf{Z}_j$ is a $K_j \times 2$ matrix whose columns are equal to the first two columns of $\mathbf{X}_j$. The vector $\mathbf{b}_j$ is a vector of length two with

distribution $N(\mathbf{0}, \mathbf{\Omega})$ (with $\boldsymbol{\zeta}$ comprising the elements Ω_{11}, Ω_{12}, and Ω_{22}. The vector $\boldsymbol{\epsilon}_j$ is again distributed as $N(\mathbf{0},\sigma_\epsilon^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix of size K_j.

The matrices $\mathbf{X}_j$ and $\mathbf{Z}_j$ are random insofar as the number of subunits within a unit, the specific subunits measured within a unit (e.g. measurement times or body parts), and the explanatory variables are random quantities that vary from unit to unit. We assume here that $\mathbf{X}_j$ and $\mathbf{Z}_j$ are independent of $\mathbf{b}_j$ and $\boldsymbol{\epsilon}_j$. In the missing data context, this corresponds to an assumption that the missing data are missing completely at random (MCAR) (see Rubin, 1976; Little and Rubin, 2002). The case of non-MCAR missingness is more complex; see Zucker and Denne (2002) for a discussion. At the analysis stage, it is usual to condition on $\mathbf{X}_j$ and $\mathbf{Z}_j$, thus treating them as fixed quantities. In sample size calculations, however, the randomness of $\mathbf{X}_i$ and $\mathbf{Z}_i$ comes into play.

We can rewrite the model (5.17) as $\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{e}_j$ where $\mathbf{e}_j = \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}$ is distributed $N(\mathbf{0},\mathbf{\Sigma}_j(\boldsymbol{\theta}))$, with $\mathbf{\Sigma}_j = \mathbf{Z}_j\mathbf{\Omega}(\boldsymbol{\zeta})\mathbf{Z}_j^T + \sigma_\epsilon^2\mathbf{R}_j(\boldsymbol{\xi})$ (here $\boldsymbol{\theta}$ includes $\boldsymbol{\zeta}, \boldsymbol{\xi}$, and σ_ϵ^2.

We take the parameter of interest to be $\boldsymbol{\lambda} = \mathbf{L}\boldsymbol{\beta}$, where $\mathbf{L}$ is a pre-specified matrix. For the mixed ANOVA model, the natural choice is $\mathbf{L} = [\mathbf{0}|\mathbf{I}]$, a zero vector of length $K-1$ followed by the identity matrix of size $K-1$, so that $\boldsymbol{\lambda} = [a_1, \ldots, a_{I-1}]^T$, the vector of treatment effect parameters. For the two-treatment random line model, we may be interested in the parameter α_{11} that describes the treatment effect on the intercept, the parameter α_{12} that describes the treatment effect on the slope, both parameters jointly, or some linear combination of these parameters (e.g. a linear combination that describes the difference between the treatment groups with respect to area under the curve).

The statistical theory of models of the form (5.17) is well-established. See, for example, Laird and Ware (1982) or Jennrich and Schlucter (1986). The model parameters are usually estimated by the maximum likelihood method or by a variant method known as restricted maximum likelihood. These two methods are asymptotically equivalent, and so we will not distinguish between them in our discussion. It is known that, conditional on $\mathbf{X}_j$ and $\mathbf{Z}_j$, the asymptotic covariance matrix of the estimate $\hat{\boldsymbol{\beta}}$ is given by $\mathbf{G}^{-1}$, where

$$\mathbf{G} = \sum_{j=1}^{N} \mathbf{X}_j^T \mathbf{\Sigma}_j(\boldsymbol{\theta})^{-1} \mathbf{X}_j. \tag{5.18}$$

Unconditionally, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $N^{-1}\mathbf{H}^{-1}$, where

$$\mathbf{H} = E[\mathbf{X}_j^T \mathbf{\Sigma}_j(\boldsymbol{\theta})^{-1} \mathbf{X}_j], \tag{5.19}$$

with the expectation taken over the distribution of $(\mathbf{X}_j, \mathbf{Z}_j)$. The asymptotic covariance matrix of $\hat{\boldsymbol{\lambda}}$ is then given by $N^{-1}\mathbf{\Gamma}$ with $\mathbf{\Gamma} = \mathbf{L}\mathbf{H}^{-1}\mathbf{L}^T$.

Suppose now that we wish to test the null hypothesis $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}_0$, where $\boldsymbol{\lambda}_0$ is a pre-specified null value (such as $\mathbf{0}$). Denote $\boldsymbol{\Delta} = \boldsymbol{\lambda} - \boldsymbol{\lambda}_0$ and $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0$. In addition, denote by $\hat{\mathbf{G}}$ the estimate of $\mathbf{G}$ obtained by substituting into the expression for $\mathbf{G}$ the estimated values of $\boldsymbol{\zeta}$, $\boldsymbol{\xi}$, and σ^2, and define $\hat{\boldsymbol{\Gamma}} = N\mathbf{L}^T\hat{\mathbf{G}}^{-1}\mathbf{L}$. Then H_0 then can be tested using the Wald test statistic $W = N\hat{\boldsymbol{\Delta}}^T\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\Delta}}$. Alternatively, a likelihood ratio statistic or score statistic could be used, but these statistics are asymptotically equivalent to the Wald statistic, and so we will focus on the Wald statistic. Under H_0, the Wald statistic W is distributed asymptotically as χ^2_l, where l is the number of rows in $\mathbf{L}$. Under a general alternative, the statistic W has an asymptotic noncentral χ^2_l distribution with noncentrality parameter $\eta = N\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}$. Hence, once $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ are specified, power calculations can be carried out using methods for the noncentral chi-square distribution described in Sec. x.x. In the case of testing a scalar parameter ($l = 1$), we can use instead the test statistic $Z = \sqrt{N}\hat{\Delta}/\sqrt{\Gamma}$, which is asymptotically distributed as $N(\mu, 1)$ with $\mu = \sqrt{N}\Delta/\sqrt{\Gamma}$. The methods for a normally-distributed test statistic described in Sec. x.x then can be applied.

It is possible to incorporate a small-sample refinement into the above development by using F and t approximations rather than χ^2 and normal approximations. A number of approaches to defining the denominator degrees of freedom have been proposed; see Kenward and Roger (1997) for discussion, and Manor and Zucker (2004) for a more recent review. For sample size calculation purposes, it is reasonable to follow the simple approach of taking the denominator degrees of freedom equal to $N - \dim(\boldsymbol{\beta})$. In the random line model context, Manor and Zucker showed that a similar simple approach exhibited reasonable Type I error performance. Given the denominator degrees of freedom and the other specifications as described in the preceding paragraph, the methods described previously for the noncentral t and F distributions can be applied (Secs. x.x and y.y).

The key point needing attention is the specification of $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$. The specification of $\boldsymbol{\Delta}$ will as usual be driven by considerations of clinical or public health significance and plausibility, and guided by prior studies and the judgment of subject-matter experts. In regard to the specification of $\boldsymbol{\Gamma}$, estimates or guesses of the covariance parameters $\boldsymbol{\zeta}, \boldsymbol{\xi}$, and σ^2 typically can be obtained from prior studies, while specification of the behavior of $\mathbf{X}_j$ and $\mathbf{Z}_j$ can be arrived at through consideration of the anticipated design structure.

As an example, suppose that a two-group repeated measures study is planned, to be analyzed using the random line model (5.12). Then, as described above, the structure of $\mathbf{X}_j$ and $\mathbf{Z}_j$ is determined by the pattern of the observation times $\{t_{jk}\}$. Typically, the study protocol will specify a schedule of planned observation times $\{\tau_k\}$. The simplest situation that can be assumed is that all individuals will be measured exactly at all planned times, without any dropout or protocol violations. In this situation, $\mathbf{X}_j$ and $\mathbf{Z}_j$ can be specified immediately as described previously. With The sample size calcu-

lation procedure reduces in this case to the Schlesselman procedure described in the preceding section. Moving to a higher level of sophistication, we can assume, following Dawson (1998) and Rochon (1998), that a finite number S of possible observation time patterns are possible, with the probability of obtaining the s-th pattern in a subject in the r-th group denoted by $\pi_{s|r}$. This allows for the possibility that the missingness could differ between the groups, for example due to side effects of the treatment. For group r and observation time pattern s, denote by $\mathbf{X}_r^{(s)}$ and $\mathbf{\Sigma}_r^{(s)}(\boldsymbol{\theta})$ the realizations of $\mathbf{X}_j$ and $\mathbf{\Sigma}_j(\boldsymbol{\theta}) = \mathbf{Z}_j\mathbf{\Omega}(\boldsymbol{\zeta})\mathbf{Z}_j^T + \sigma_\epsilon^2 R_j(\boldsymbol{\xi})$, respectively, that apply to the relevant combination of group assignment and observation pattern. Further, denote the planned proportion of units to be assigned to arm r by q_r. Then the matrix $\mathbf{H}$ defined in (5.19) is given by the following expression (cf. Dawson, 1998, Sec. 3, and Zucker and Denne, 2002, Eqn. (4)):

$$\mathbf{H} = \sum_{r=1}^{2}\sum_{l=1}^{S} q_r \pi_{s|r} {\mathbf{X}_r^{(s)}}^T {\mathbf{\Sigma}_r^{(s)}(\boldsymbol{\theta})}^{-1} \mathbf{X}_r^{(s)}. \tag{5.20}$$

Let us now restrict attention to random dropout and random intermittent missingness, and assume that whenever an observation is obtained it is taken exactly at the planned time. In this case, there are 2^K possible observation patterns, where K is the planned maximum number of observations (beyond the baseline observation which we assume is available on all individuals), corresponding to presence or absence of each of the K possible measurements. These patterns are immediately identifiable from the planned observation schedule. It remains to specify the probabilities $\pi_{s|r}$ of the different patterns. In some cases, it may be possible to estimate these probabilities directly from previous studies of a similar design. In other cases, the $\pi_{s|r}$ can be specified through statistical modelling.

Zucker and Denne (2002) discussed a model in which time to dropout is modelled using a standard parametric survival distribution such as the Weibull, and independent intermitting missingness is modelled by assuming that each planned observation prior to the dropout time is successfully obtained with probability $1 - p_r$ in group r, independently of the other observations. An expression for $\pi_{s|r}$ under this model can be given as follows. Let $\tau_0 = 0$ and $\tau_{K+1} = \infty$. For each observation pattern s, there is some k_s such that there is an observation at τ_{k_s}, but no observations at any τ_k with $k > k_s$ (k_s may equal 0 or K). Further, for each pattern s, we can identify the number of missing visits prior to τ_{k_s}; denote this by m_l. Finally, let G_r denote the survival function for the time to dropout T^0 for group r, and define $\varphi_{rk} = \Pr_r(T^0 \in (\tau_{k-1}, \tau_k]) = G_r(\tau_{k-1}) - G_r(\tau_k)$. Then the probability $\pi_{s|r}$ of pattern s in group r is given by

$$\pi_{s|r} = \sum_{k=k_s+1}^{K+1} (1-p_r)^{k_s - m_s} p_r^{m_s + k - k_s - 1} \varphi_{kr}.$$

The parameters of the dropout time distribution and the intermittent missingness probabilities p_r are estimated through a combination of prior studies and investigator judgment. In principle, it is possible to extend the methodology to allow for random jitter in the actual times of observation, but for sample size planning purposes this is probably an undue level of refinement. Another obvious question is how to incorporate background covariates into the sample size calculation. This is a complex problem which requires some guess about the distribution of the background covariates. The general approach is along the lines described in Chapter x on regression methods. A simple practical strategy is to approximate the covariate distribution by a discrete distribution, which can be described in terms of strata. The above method for accounting for different observation patterns can then be immediately extended to cover the case of different covariate strata, along the lines of the setup for GEE models described in the next section.

[The example below can be replaced by John Lachin's Schlesselman example if John can dig it up.]

Example: We present here a modified version of an example given by Dawson (1998). We suppose that it is desired to test a diet/exercise program aimed at reducing weight increase among young men. A baseline measurement of height and weight is taken at age $\tau_0 = 16$, and then a intervention consisting of periodic diet/exercise counseling is run in half of the study participants, with the other half serving as a control group (thus $q_1 = q_2 = 0.5$). Additional height and weight measurements are taken at ages $\tau_1 = 18, \tau_2 = 20, \cdots, \tau_8 = 32$. The endpoint of interest is the logarithm of the Quetelet Body Mass Index (BMI), defined as the weight in kilograms divided by the square of the height in meters. The data will be analyzed using the random line model (5.12), with the primary comparison being a test for a between-group difference in the slopes. On the basis of data from the Muscatine cohort study (e.g. Lauer, Lee, and Clarke, 1988), the following preliminary estimates of the covariance parameters are obtained: $\sigma^2_{\beta_1} = 0.00137$, $\sigma^2_{\beta_2} = 0.000175$, $\mathrm{Cov}(\beta_1, \beta_2) = 0.000037$, and $\sigma^2_\epsilon = 0.000242$. The projected mean slope in the control group is 0.0100. It is desired to detect a 35% reduction in the mean slope, from 0.0100 to 0.0065, corresponding to a mean slope difference of 0.0035. The required two-sided Type I error level is 0.05, and the desired power is 80%.

We assume first that all subjects will have complete data. In this case, applying (5.14), we find that the variance of an individual subject's OLS slope is 0.0001851. Accordingly, using the normal sample size formula (EQN), we find that the required total sample size is

$$N = 4(1.96 + 0.84)^2 0.0001851/(0.0035)^2 = 474.$$

Next, we consider the case where random dropout will occur at a projected rate of 10% every two years, with no intermittent missingness. A crude dropout adjustment would be to inflate the sample size to ensure that 474

subjects will complete the study, thus disregarding the partial information provided by the dropouts. The probability that a subject will complete the study is $(0.9)^8 = 0.43$. Thus, the sample size would be inflated by a factor of $(0.43)^{-1} = 2.32$, yielding a total sample size of 1,102. A more sophisticated adjustment can be carried out based on (5.20). There are 9 possible observation patterns, with the s-th pattern being $\{\tau_0, \tau_1, \ldots, \tau_{s-1}\}$ (for $s = 1$, we get the degenerate pattern with a baseline measurement only). The corresponding probabilities are $\pi_{s|r} = (0.1)(0.9)^{s-1}$ for both the treatment ($r = 2$) and control ($r = 1$) groups. Under each of these patterns, the structure of $\mathbf{X}$ for treated subjects, the structure of $\mathbf{X}$ for control groups subjects, and the structure of $\mathbf{Z}$ is as described earlier in this section, with the t_{jk}'s now being given by $\{\tau_0, \tau_1, \ldots, \tau_{s-1}\}$ for subjects with observation pattern s. Applying (5.20) (using a simple SAS IML program), we obtain $H_{11} = H_{33} = 37.21$, $H_{22} = H_{44} = 5055.28$, and $H_{12} = H_{21} = H_{34} = H_{43} = 129.66$, with all other elements of $\mathbf{H}$ equal to zero. The parameter λ of interest is the mean difference in slopes between the two groups, given by $\lambda = 2\alpha_{12}$, which can be expressed in the form $\lambda = \mathbf{L}\boldsymbol{\beta}$ with $\mathbf{L} = [0\ 0\ 0\ 2]$. We thus get $\Gamma = \mathbf{L}\mathbf{H}^{-1}\mathbf{L}^T = 0.0008689$. The required total sample size is then computed as

$$N = (1.96 + 0.84)^2 0.0008689/(0.0035)^2 = 556.$$

This sample size is 1.18 times the sample size of 474 that is required when there are no missing data, as compared with 2.32 times 474 with the crude missing data adjustment. We thus see that adjusting for missing data has a significant effect on the sample size, and that a sophisticated sample size calculation that accounts for the various possible observation patterns yields a much lower sample size than a crude complete-case sample size adjustment.

General repeated measurements data, both continuous and discrete, can also be analyzed using a nonlinear mixed model. In this type of model, the expectation of $\mathbf{Y}_j$ is modelled as

$$E[Y_{jk}|\mathbf{X}_j, \mathbf{Z}_j, \mathbf{b}_j] = g([\mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j]_k),$$

where g is a known "link" function and $\mathbf{X}_j$, $\mathbf{Z}_j$, and $\mathbf{b}_j$ are as in the linear mixed model. As there, the $\mathbf{b}_j$ are typically modelled as being multivariate normal, though other distributions are possible. The model formulation is completed by specifying either the full distribution of $\mathbf{Y}_j$ or its covariance structure as a function of $E[\mathbf{Y}_j|\mathbf{X}_j, \mathbf{Z}_j, \mathbf{b}_j]$ and some additional parameters. These models are discussed by Davidian and Giltinan (1995, 2003) and by Vonesh and Chinchilli (1997). In view of the complexities of the nonlinear mixed model, sample size calculation under the setting of this model is probably too complicated for practical application in most studies. An alternative is the generalized estimating equation (GEE) setting discussed in the next section. As Breslow and Clayton (1993) note, the nonlinear mixed model setting and the GEE setting are related in that a nonlinear mixed model analysis can

be approximated to first order by a GEE analysis. Thus, the GEE framework is appropriate for sample size calculation.

5.9 GENERALIZED ESTIMATING EQUATIONS (GEE) ANALYSIS

The generalized estimating equations (GEE) method of Liang and Zeger (1986) is a popular approach to the analysis of repeated measures and clustered data, for both continuous and discrete outcomes. We review the theory of GEE analysis and then discuss sample size calculation in the GEE setting. The presentation in this section follows Rochon (1998). There is a close resemblance with the theory of the preceding section. Indeed, when the expected response is modelled as a linear function of the covariates, the linear mixed model approach and the GEE analysis approach are roughly equivalent.

We again denote by Y_{jk} the value of the k-th observation on the j-th unit ($j = 1, \ldots, N$; $k = 1, \ldots, K_j$), and by $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jK_j})^T$ the vector of all the observations on the j-th unit. Associated with unit j is a matrix $\mathbf{X}_j$ involving explanatory variables. The expectation vector $\boldsymbol{\mu}_j = E[\mathbf{Y}_j|\mathbf{X}_j]$ is modelled as $\mu_{jk} = g([\mathbf{X}_j\boldsymbol{\beta}]_k)$, where g is a known "link" function. The covariance matrix $\mathrm{Cov}(\mathbf{Y}_j|\mathbf{X}_j)$ is modelled as $\mathbf{V}_j = \psi\mathbf{A}_j^{1/2}\mathbf{R}_j(\boldsymbol{\zeta})\mathbf{A}_j^{1/2}$, where $\mathbf{A}_j = \mathrm{diag}(a(\mu_{j1}), \ldots, a(\mu_{jK_j}))$ for some known function a, $\mathbf{R}_j(\boldsymbol{\zeta})$ is the correlation matrix of $\mathbf{Y}_j$, which is modelled as a function of a parameter vector $\boldsymbol{\zeta}$, and ψ is a scale parameter. Various models for $\mathbf{R}$ are possible, including compound symmetry structure, autoregressive-like structure, and moving average-like structure. See Diggle et al. (2002) for details. See also Oman and Zucker (2001), and Qaqish (2003), for a discussion of covariance structures for clustered binary data. In GEE theory, it is assumed that the model for $E[\mathbf{Y}_j|\mathbf{X}_j]$ is correct, but the model for $\mathrm{Cov}(\mathbf{Y}_j|\mathbf{X}_j)$ may be misspecified.

For a classical linear regression model, g is the identity function, a is identically equal to 1 (this can be modified for heteroscedastic models), and ψ is the error variance σ^2. For a logistic regression model with binary Y_{jk}, g is the logistic function $g(u) = e^u/(1 + e^u)$, $a(\mu) = \mu(1 - \mu)$, and $\psi = 1$. For the Poisson regression model, the classical link function is $g(u) = \log u$, and we have $a(\mu) = \mu$ and $\psi = 1$. Often, for integer-valued data, a binomial-like or Poisson-like model is fit with general ψ to allow for "underdispersion" or "overdispersion" relative to the binomial or Poisson distribution.

Let $\mathbf{D}_j$ be the matrix defined by $[\mathbf{D}_j]_{kr} = \partial\mu_{jk}/\partial\beta_r$. We can write $\mathbf{D}_j = \mathbf{Q}_j\mathbf{X}_j$, where $\mathbf{Q}_j = \mathrm{diag}[g'([\mathbf{X}_j\boldsymbol{\beta}]_1), \ldots, g'([\mathbf{X}_j\boldsymbol{\beta}]_{K_j})]$. The parameter vector $\boldsymbol{\beta}$ is estimated as the solution to the estimating equation

$$\sum_{j=1}^{N} \mathbf{D}_j^T\mathbf{V}_j^{-1}(\mathbf{Y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta})) = \mathbf{0}.$$

As indicated by McCullagh and Nelder (1989, Secs. 2.5 and 9.2.3), the estimator can be obtained by iteratively reweighted least squares based on the

formula

$$\hat{\boldsymbol{\beta}}^{(q+1)} = \left[\sum_{j=1}^{N} \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(q)}}^{-1} \left[\sum_{j=1}^{N} \mathbf{D}_j^T \mathbf{V}_j^{-1} (\mathbf{Y}_j - \boldsymbol{\mu}_i(\boldsymbol{\beta}) + \mathbf{D}_j \boldsymbol{\beta})\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(q)}}.$$

Estimation of $\boldsymbol{\zeta}$ and ψ is described in Liang and Zeger (1986). Under suitable conditions, Liang and Zeger show that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal.

If the covariance structure is correctly specified, the asymptotic covariance matrix $\boldsymbol{\Upsilon}$ of $\hat{\boldsymbol{\beta}}$ conditional on the $\mathbf{X}_j$ is given by $\mathbf{G}^{-1}$, with (cf. (5.18)):

$$\mathbf{G} = \sum_{j=1}^{N} \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{D}_j = \sum_{j=1}^{N} \mathbf{X}_j^T \mathbf{Q}_j^T \mathbf{V}_j^{-1} \mathbf{Q}_j \mathbf{X}_j. \tag{5.21}$$

For the case where the covariance may be misspecified, Liang and Zeger present a more complex "robust" covariance formula (commonly called the "sandwich" formula), given by

$$\boldsymbol{\Upsilon}^* = \mathbf{G}^{-1} \left[\sum_{j=1}^{N} \mathbf{D}_j^T \mathbf{V}_j^{-1} \mathbf{V}_j^* \mathbf{V}_j^{-1} \mathbf{D}_j\right] \mathbf{G}^{-1},$$

where $\mathbf{G}$ is as in (5.21) and $\mathbf{V}_j^*$ is the *true* value of $\mathrm{Cov}(\mathbf{Y}_j|\mathbf{X}_j)$. In practice, $\boldsymbol{\Upsilon}^*$ is replaced by the empirical estimator

$$\hat{\boldsymbol{\Upsilon}} = \mathbf{G}^{-1} \left[\sum_{j=1}^{N} \mathbf{D}_j^T \mathbf{V}_j^{-1} (\mathbf{Y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta}))(\mathbf{Y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta}))^T \mathbf{V}_j^{-1} \mathbf{D}_j\right] \mathbf{G}^{-1},$$

with the values of $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ replaced by estimates (ψ cancels out).

At the analysis stage, the robust covariance estimator will usually be preferable (except perhaps if number of units is very small). But for sample sample size calculation, Rochon argues cogently that trying to take account of possible differences between the modelled covariance structure of $\mathbf{Y}_j$ and the true covariance structure is unduly complex. Hence, following Rochon, we will describe sample size calculation under the assumption that the assumed covariance model is correct.

In this case, the theory parallels that presented for the linear mixed model in the preceding section. The unconditional asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $N^{-1}\mathbf{H}^{-1}$, with (cf. (5.19))

$$\mathbf{H} = E[\mathbf{X}_j^T \boldsymbol{\Omega}_j \mathbf{X}_j], \tag{5.22}$$

where the expectation is taken over the distribution of $\mathbf{X}_j$ and $\boldsymbol{\Omega}_j$ is defined by $\boldsymbol{\Omega}_j = \mathbf{Q}_j^T \mathbf{V}_j^{-1} \mathbf{Q}_j$.

The parameter of interest is taken to be $\boldsymbol{\lambda} = \mathbf{L}\boldsymbol{\beta}$, where $\mathbf{L}$ is a pre-specified matrix. The asymptotic covariance matrix of $\hat{\boldsymbol{\lambda}}$ is then given by $N^{-1}\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma} = \mathbf{L}\mathbf{H}^{-1}\mathbf{L}^T$. The null hypothesis of interest is taken to be $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}_0$, where $\boldsymbol{\lambda}_0$ is a pre-specified null value (such as $\mathbf{0}$). We define $\boldsymbol{\Delta} = \boldsymbol{\lambda} - \boldsymbol{\lambda}_0$ and $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0$. We denote by $\hat{\mathbf{G}}$ the estimate of $\mathbf{G}$ obtained by substituting into the expression for $\mathbf{G}$ the estimated values of $\boldsymbol{\beta}$, $\boldsymbol{\zeta}$, and ψ, and define $\hat{\boldsymbol{\Gamma}} = N\mathbf{L}^T\hat{\mathbf{G}}^{-1}\mathbf{L}$. The null hypothesis H_0 can then be tested using the Wald test statistic $W = N\hat{\boldsymbol{\Delta}}^T\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\Delta}}$, whose asymptotic null distribution is χ^2_l, where l is the number of rows in $\mathbf{L}$. Under a general alternative, the statistic W has an asymptotic noncentral χ^2_l distribution with noncentrality parameter $\eta = N\boldsymbol{\Delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}$. Hence, once $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ are specified, power calculations can be carried out using methods for the noncentral chi-square distribution described in Sec. x.x. In the case of testing a scalar parameter ($l = 1$), we can work instead with the test statistic $Z = \sqrt{N}\hat{\Delta}/\sqrt{\Gamma}$, and apply methods for a normally-distributed test statistic.

It remains to discuss the specification of $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$. The specification of $\boldsymbol{\Delta}$ will be based on the usual considerations of practical significance and plausibility. The parameters $\boldsymbol{\beta}$, $\boldsymbol{\zeta}$, and ψ that enter into the expression for $\boldsymbol{\Omega}_j$ can be estimated or guessed from prior studies.

The behavior of $\mathbf{X}_j$, as before, can be described through consideration of the anticipated design structure. Rochon (1998) assumes that there is a known collection $\mathbf{X}^{(s)}, s = 1, \ldots, S$, of possible $\mathbf{X}$ matrices, corresponding to a known specified set of possible covariate configurations in different strata. Here we absorb the treatment group indicators, if any, directly into the $\mathbf{X}^{(s)}$ matrices rather than distinguishing treatment groups explicitly as in the preceding section. Let π_s denote the probability of obtaining covariate configuration s. Under this setup, the matrix $\mathbf{H}$ defined by (5.22) is given by the following analogue of (5.20):

$$\mathbf{H} = \sum_{s=1}^{S} \pi_s \mathbf{X}^{(s)^T} \boldsymbol{\Omega}^{(s)} \mathbf{X}^{(s)}. \tag{5.23}$$

Dropout and missing data can be accounted for by expanding the collection of possible $\mathbf{X}$ matrices, along the lines described in the preceding section. Rochon (1998, Sec. 4) describes this in detail.

Example: As an illustration, we present a modified (and simplified) version of the sample size calculation for the Post Coronary Artery Bypass (Post-CABG) trial (Post-CABG Investigators, 1997). This was an NIH-sponsored factorial trial designed to test the efficacy of extreme cholesterol-lowering (as compared with usual-care cholesterol lowering) and warfarin (as compared with placebo) for maintaining patency of saphenous vein bypasses in patients who had undergone heart bypass surgery. For purposes of this example, we will set aside the factorial aspect and assume a simple two-group trial of aggressive versus standard cholesterol lowering with equal allocation. Based on prior studies, the projected per-graft 5-year rate of substantial graft narrow-

ing in the standard care group is taken to be $p_0 = 0.15$, with a between-graft correlation (i.e. Cohen's kappa) between two grafts in the same patient of $\rho = 0.40$. We specify the treatment effect desired to be detected as being a 33% reduction in the per-graft progression rate, to a level of $p_1 = 0.10$. Thus, the difference to be detected is $p_0 - p_1 = 0.05$.

We assume that, at entry, 50% of the patients will have 2 patent grafts, 30% will have 3 patent grafts, and 20% will have 4 patent grafts. We simplify the example by assuming that all patients will be followed for the full 5 years with no deaths or dropouts. We assume also that the percentage of patients who will suffer a heart attack during the trial in negligible, thus ensuring that examination of graft patency is a reasonable measure of the patient's clinical condition. The parameter of interest is taken to be the difference in per-graft progression rates between the treatment and control groups. The sample size calculation is done without accounting for covariates other than the treatment group indicator, although in the analysis it may be desired to incorporate some background covariates. The desired two-sided Type I error level is 0.05 and the desired power is 90%.

In this setting, Y_{jk} is a binary variable equal to 1 if the j-th subject's k-th graft suffered substantial progression, and zero otherwise. The number of observations K_j on a subject can be 2, 3, or 4. Although binary data are usually modelled with a logistic link, in the present setting, with the treatment indicator as the only covariate, we can use the identity link. The regression parameter vector $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = [p_0 \ p_1]^T$. The matrix $\mathbf{X}_j$ comprises two columns, with the first column being a column of 1's and the second column being a column of either 1's or -1's, according to whether the subject is in the treatment group or the control group. The number of rows in $\mathbf{X}_j$ is equal to K_j. Thus there are $S = 6$ possible $\mathbf{X}$ matrices. corresponding to the three possible values of K_j and two possible treatment indicator levels. According the assumptions stated above, these six $\mathbf{X}$ configurations have respective probabilities $\pi_1 = 0.25, \pi_2 = 0.15, \pi_3 = 0.10, \pi_4 = 0.25, \pi_5 = 0.15$, and $\pi_6 = 0.10$. The covariance matrix $\mathbf{V}_j$ has compound symmetry structure with diagonal values $p(1-p)$ and off-diagonal values $\rho p(1-p)$, where p equals p_0 for control group subjects and p_1 for treatment group subjects. The matrix $\mathbf{Q}_j$ is equal to the identity matrix, so that $\mathbf{D}_j = \mathbf{X}_j$ and $\boldsymbol{\Omega}_j = \mathbf{V}_j^{-1}$. Thus, the formula (5.23) becomes

$$\mathbf{H} = \sum_{s=1}^{S} \pi_s {\mathbf{X}^{(s)}}^T \mathbf{V}^{(s)} \mathbf{X}^{(s)},$$

where $\pi_s, \mathbf{X}^{(s)}$, and $\mathbf{V}^{(s)}$ are as just described. Substituting in the projected values of the various parameters and evaluating the above expression (which can be done with a simple SAS IML program) we obtain $H_{11} = H_{22} = 14.9542$ and $H_{12} = H_{21} = 2.5783$. We note that $\mathbf{H}$ is always symmetric, while the equality $H_{11} = H_{22}$ in this example is due to the particular structure of the example. Inverting $\mathbf{H}$, we obtain $[\mathbf{H}^{-1}]_{11} = [\mathbf{H}^{-1}]_{22} = 0.06892$ and $[\mathbf{H}^{-1}]_{12} = [\mathbf{H}^{-1}]_{21} = -0.01188$.

The parameter λ of interest is given by $\lambda = \mathbf{L}\boldsymbol{\beta}$, where $\mathbf{L} = [0\ 2]^T$. The variance of λ is given by $N^{-1}\Gamma$, with $\Gamma = 4[\mathbf{H}^{-1}]_{22} = 0.2757$. Using the normal sample size formula, we find that the required total sample size N is

$$N = (1.96 + 1.28)^2 0.2757/(0.05)^2 = 1,158.$$

If the trial were restricted to subjects with exactly 4 patent grafts at entry, the sample size could be computed by following the foregoing procedure with the π_s's defined as follows: $\pi_1 = 0, \pi_2 = 0, \pi_3 = 0.5, \pi_4 = 0, \pi_5 = 0$, and $\pi_6 = 0.5$. Alternatively, since this is a case where all subjects have the same number of observations, the formula (5.7) can be used. Both approaches, as expected, give the same result – namely, a total sample size of 1,004. The similarity between the two sample size figures is due to the high projected correlation of 0.40. If, instead, the correlation were 0.05, the sample sizes would be 742 under the original distribution for the number of grafts and 526 under the assumption that all subjects would have exactly 4 grafts. Not only does a lower correlation lead to a lower sample size, as expected, but it also causes the distribution of the number of subunits within a unit to have a greater influence on the sample size.