12

Marginal Models: Introduction and Overview

12.1 INTRODUCTION

In the previous chapter we reviewed generalized linear models for a single response variable. A straightforward application of these models to longitudinal data is not appropriate, owing to the lack of independence among repeated measures obtained on the same individual. There are, however, a number of ways to extend generalized linear models to handle longitudinal data. All of these procedures account for the within-subject correlation among the repeated measures, though they differ in approach. We will see in Chapters 12 through 16 that the method of accounting for the within-subject association has important ramifications for the interpretation of the regression coefficients in models for discrete longitudinal data. For the linear regression models for continuous responses considered in Part II, the interpretation of the regression coefficients is independent of assumptions made about the correlation among the repeated measures. With discrete longitudinal data this is no longer necessarily the case. Instead, different assumptions about the source of the within-subject association can lead to regression coefficients with quite distinct interpretations. The need to distinguish models according to the interpretation of their regression coefficients has led to the use of the terms "marginal models" and "mixed effects models"; the former are often referred to as "population-average models," the latter as "subject-specific models." For the former the target of inference is the population, for the latter the target of inference is the individual. In this chapter, we introduce the main features of marginal models for longitudinal data; the meaning of the term "marginal," as used in this context, will soon be apparent. In Chapter 13, we discuss estimation of marginal models and present three case studies that illustrate the application of marginal mod-

Applied Longitudinal Analysis, Second Edition. By Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware.

els to longitudinal data. Mixed effects models, specifically, generalized linear models with random effects, are the focus of Chapters 14 and 15.

Because the method of accounting for the within-subject association has consequences for the interpretation of the regression model parameters, the choice of method for analyzing discrete longitudinal data cannot be made through any automatic procedure. Rather, the choice must be made on subject-matter grounds. Different models for discrete longitudinal data have somewhat different targets of inference and thereby address subtly different scientific questions. We return to this important issue in Chapter 16.

In this chapter we consider an approach for extending generalized linear models to longitudinal data that leads to a class of regression models that are known as *marginal models*. The term *marginal* in this context indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or previous responses. That is, the term marginal is used to emphasize that the model for the mean response at each occasion does not incorporate dependence on any random effects or previous responses. This is in contrast to *mixed effects models*, where the mean response depends not only on covariates but also on a vector of random effects. Marginal models provide a very natural way of extending generalized linear models to longitudinal data, and they have frequently been applied in the biomedical and health sciences. Marginal models do not require distributional assumptions for the observations, only a regression model for the mean response. That is, marginal models provide a unified method for analyzing diverse types of longitudinal responses, which avoids making assumptions about the distribution of the vector of responses; the method relies solely on assumptions about how the mean response is related to the covariates. The avoidance of distributional assumptions leads to a method of estimation known as *generalized estimating equations* (GEE). The generalized estimating equations approach is a general method of estimation for marginal models that will be described in detail in Chapter 13.

In our discussion of marginal models in this and the subsequent chapter, the main focus is on discrete response data, for example, binary responses and counts. However, we also point out connections between marginal models for a continuous response and the methods for longitudinal data analysis presented in Part II. In doing so, we can provide some rationale for why the multivariate normal distributional assumption made in Part II often can be relaxed.

12.2 MARGINAL MODELS FOR LONGITUDINAL DATA

We begin our discussion of marginal models by introducing some notation similar to that used in Part II. We assume that N subjects are measured repeatedly over time. We let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion. The response variable can be continuous, binary, ordinal, or a count. The nature of the response variable does have important implications for model specification; however, the notation does not distinguish among the different types of responses.

We do not require that subjects have the same number of repeated measures or that they are measured at a common set of occasions. To accommodate unbalanced data (i.e., repeated measurements that are not obtained at a common set of occasions), we assume that there are n_i repeated measurements of the response on the i^{th} subject and that each Y_{ij} is observed at time t_{ij}. Both the longitudinal data structure and the notation are the same as that used in Chapter 8; the only difference is that the response variable is no longer assumed to be continuous. The response variables for the i^{th} subject can be grouped into an $n_i \times 1$ vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad i = 1, ..., N;$$

where the vectors of responses, Y_i, are assumed to be independent of one another (but the repeated measures on the same subject are emphatically not assumed to be independent). Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, ..., N; \quad j = 1, ..., n_i.$$

Each individual has a vector of covariates, X_{ij}, associated with the response at each occasion, Y_{ij}. Note that X_{ij} may include covariates whose values do not change throughout the duration of the study and covariates whose values change over time. The former are referred to as time-invariant or between-subject covariates (e.g., gender and fixed experimental treatments), whereas the latter are referred to as time-varying or within-subject covariates (e.g., time since baseline, current smoking status, and environmental exposures). In the former case, the same values of the covariates are replicated in the corresponding rows of X_{ij}, for $j = 1, ..., n_i$. In the latter case, the values taken by the covariates can vary over time (for at least some individuals) and the values in the corresponding rows of X_{ij} can be different at each occasion.

We can group the vectors of covariates into an $n_i \times p$ matrix of covariates

$$X_i = \begin{pmatrix} X'_{i1} \\ X'_{i2} \\ \vdots \\ X'_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix}, \quad i = 1, ..., N,$$

where the rows of X_i correspond to the covariates associated with the responses at the n_i different measurement occasions, and the columns of X_i correspond to the p distinct covariates. So far we have assumed that each subject has a vector of repeated

responses, denoted by Y_i, and associated with each repeated measure there is a vector of p covariates which can be grouped into a matrix, X_i.

Marginal models are primarily used to make inferences about population means. As a result marginal models for longitudinal data separately model the mean response and the within-subject association among the repeated responses. In a marginal model the goal is to make inferences about the former, whereas the latter is regarded as a nuisance characteristic of the data that must be accounted for to make correct inferences about changes in the population mean response.

A marginal model for longitudinal data has the following three-part specification:

1. The conditional expectation or mean of each response, $E(Y_{ij}|X_{ij}) = \mu_{ij}$, is assumed to depend on the covariates through a known link function

$$g(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta.$$

2. The conditional variance of each Y_{ij}, given the covariates, is assumed to depend on the mean according to

$$\text{Var}(Y_{ij}|X_{ij}) = \phi\, v(\mu_{ij}),$$

where $v(\mu_{ij})$ is a known "variance function" (i.e., a known function of the mean, μ_{ij}) and ϕ is a scale parameter that may be known or may need to be estimated. For balanced longitudinal designs, a separate scale parameter, ϕ_j, could be estimated at each occasion; alternatively, the scale parameter could depend on the times of measurement, with $\phi(t_{ij})$ being some parametric function of t_{ij}.

3. The conditional within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of an additional set of association parameters, α (and also depends on the means, μ_{ij}). For example, the components of α might represent the pairwise correlations or log odds ratios among the repeated responses. The within-subject association among the responses is described in more detail below.

This three-part specification of a marginal model makes the extension of generalized linear models to longitudinal data more transparent. The first two parts of the marginal model correspond to the standard generalized linear model, albeit with no distributional assumptions about the responses (see Section 11.2). It is the third component, the incorporation of the association among the repeated responses from the same individual, that represents the main extension of generalized linear models to longitudinal data. In principle, this three-part specification of a marginal model can be extended by making full distributional assumptions about the vector of responses, Y_i. However, in Section 12.4 we discuss why assumptions about the joint distribution of Y_i are not necessary for estimation of the parameters of the marginal model.

As noted above, the first two components of a marginal model specify the mean and variance of Y_{ij} following the standard generalized linear model formulation described in Chapter 11, the only difference being that we have a common vector-valued

link function relating the vector of mean responses to the covariates. The third component recognizes the characteristic lack of independence among longitudinal data by modeling the association among the repeated responses from the same individual. In describing the third component we have been careful to avoid the use of the term *correlation* for two reasons. First, with a continuous response variable, the correlation is a very natural measure of the linear dependence among the repeated responses. Also the correlations are independent of the mean response, in the sense that the correlations are free to vary from -1 to 1, regardless of the values of the vector of mean responses. However, this is not the case with discrete responses. With discrete responses, the correlations are constrained by the mean responses, and vice versa. The most extreme example arises when the response variable is binary. For binary responses the correlations are restricted to ranges that are determined by the means of the responses (or the probabilities of success). For example, in the bivariate case, if $\mu_1 = E(Y_1) = \Pr(Y_1 = 1) = 0.2$ and $\mu_2 = E(Y_2) = \Pr(Y_2 = 1) = 0.8$, then $\rho_{12} = \text{Corr}(Y_1, Y_2) \leq 0.25$. That is, the correlation can be no larger than 0.25 when the probabilities of success are 0.2 and 0.8. As a result, with discrete responses, the correlation is not the most natural measure of within-subject association. Instead, the odds ratio (or the log odds ratio) is a preferable metric for association among pairs of binary responses. Second, for a continuous response that has a multivariate normal distribution, the correlations, along with the variances and the means, completely specify the joint distribution of the vector of longitudinal responses. This is not the case with discrete data. That is, the vector of means and the covariance matrix (the variances and correlations) do not, in general, completely specify the joint distribution of discrete longitudinal responses. Instead, the joint distribution requires specification of pairwise (e.g., pairwise odds ratios) and higher-order associations among the responses; this feature of discrete data is discussed in greater detail in Section 12.4.

In a certain sense marginal models are a very natural way to extend generalized linear models, developed for the analysis of independent observations, to the setting of correlated longitudinal responses. Marginal models specify a generalized linear model for the longitudinal responses but also include a model for the within-subject association among the responses. A crucial aspect of marginal models is that the mean response and within-subject association are modeled separately. This separation of the modeling of the mean response and the association among responses has important implications for interpretation of the regression parameters in the model for the mean response. In particular, the regression parameters, β, in the marginal model have *population-averaged* interpretations. That is, they describe features of the mean response in the population and how those features relate to covariates. For example, regression parameters in a marginal model might have interpretation in terms of contrasts of the changes in the mean responses in sub-populations (e.g., different treatment or exposure groups or other population strata defined by covariate values). The interpretation of β is not altered in any way by the assumptions made about the nature or magnitude of the within-subject association. We will return to this point in Chapter 16.

Of note, marginal models do not require distributional assumptions for the observations, only a regression model for the mean response. The avoidance of distributional

assumptions can be advantageous, since there is no convenient specification of the joint multivariate distribution of Y_i for marginal models when the responses are discrete. To avoid distributional assumptions for Y_i we would apply the method of estimation known as *generalized estimating equations* (GEE). The GEE approach provides a convenient alternative to maximum likelihood estimation; the GEE approach for estimating the parameters of marginal models is described in Chapter 13. In Section 12.4 we present a more detailed discussion of how assumptions about the joint distribution of Y_i are not required for estimation of the marginal model parameters and why it can be advantageous to avoid making distributional assumptions. The material in Section 12.4 is somewhat technical and can be omitted at first reading without loss of continuity.

Finally, we note that there is an implicit assumption in the first component of a marginal model that is often overlooked. Marginal models assume that the conditional mean of the j^{th} response, given $X_{i1}, ..., X_{in_i}$, depends only on X_{ij}

$$E(Y_{ij}|X_i) = E(Y_{ij}|X_{i1}, ..., X_{in_i}) = E(Y_{ij}|X_{ij}). \tag{12.1}$$

This assumption implies that given X_{ij}, there is no dependence of Y_{ij} on X_{ik} for $k \neq j$. With time-invariant covariates, this assumption poses no difficulties; it necessarily holds since $X_{ij} = X_{ik}$ for all occasions $k \neq j$. Also with time-varying covariates that are fixed by design of the study (e.g., time since baseline, treatment group indicator in a crossover trial), the assumption also holds since values of the covariates at any occasion are determined a priori by study design and in a manner completely unrelated to the longitudinal response. However, when a time-varying covariate varies randomly over time, the assumption made in (12.1) may not hold. For example, the assumption will be violated when the current value of the response, say Y_{ij}, given the current covariates X_{ij}, predicts the subsequent value of X_{ij+1}. This might arise, for example, in a longitudinal observational study designed to assess the effects of physical exercise on reducing blood glucose levels. If study participants with elevated blood glucose levels, Y_{ij}, at the j^{th} occasion subsequently increase their amount of physical activity, X_{ij+1} (while those with normal blood glucose levels continue to maintain their usual level of physical activity), then the assumption made in (12.1) does not hold. As a result somewhat greater care is required when fitting marginal models with time-varying covariates that are not fixed by design of the study. A more detailed discussion of this issue is postponed until Chapter 13 (see Section 13.5).

12.3 ILLUSTRATIVE EXAMPLES OF MARGINAL MODELS

In the previous section we described how marginal models for longitudinal data have a three-part specification in terms of assumptions concerning (1) the mean response at each occasion, (2) the variance of the response at each occasion, and (3) the pairwise within-subject association among the responses. In this section we consider some examples of marginal models using this three-part specification.

Example 1: Marginal Model for a Continuous Response

The linear regression model for longitudinal data described in Part II is a special case of the marginal model. It is useful to consider its formulation within the framework and terminology of marginal models. By doing so, the extensions to other types of response variables will become more apparent.

Suppose that Y_{ij} is a continuous response and that it is of interest to relate changes in the mean response over time to the covariates. An example of a marginal model for Y_{ij} is given by the following three-part specification:

1. The mean of Y_{ij} is related to the covariates by an identity link function,

$$\mu_{ij} = \eta_{ij} = X'_{ij}\beta.$$

2. The variance of each Y_{ij}, given the effects of the covariates, is ϕ and does not depend on the mean response. That is,

$$\text{Var}\,(Y_{ij}|X_{ij}) = \phi\, v\,(\mu_{ij}) = \phi,$$

where $v\,(\mu_{ij}) = 1$ and ϕ is a scale parameter that needs to be estimated. This model makes the strong, and often unrealistic, assumption that the variance is homogeneous over time. Alternatively, a separate scale parameter, ϕ_j, could be estimated at the j^{th} occasion if the longitudinal design is balanced on time.

3. The within-subject association among the vector of repeated responses is modeled by assuming a first-order autoregressive correlation pattern

$$\text{Corr}\,(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha^{|k-j|},$$

where $0 \leq \alpha \leq 1$. In this example it is assumed that the within-subject associations do not depend on the means but only on a single correlation parameter, α. That is, α is used to model the pairwise correlations among the responses (which are assumed to be approximately equally separated in time).

This illustration of a marginal model for a continuous response is a special case of the linear regression models for longitudinal data considered in Part II. However, marginal models provide a much broader class of models for continuous responses. For example, the means can be related to the covariates by a link function other than the identity or the variances can be allowed to depend on some known function of the means. Also in this illustration the correlations among the components of Y_i have been specified as a function of the parameter α via a first-order autoregressive correlation pattern. The correlations can take on many alternative structures, and this example is but one possible structure; other models for the correlation (e.g., unstructured or equicorrelated correlation patterns) can be adopted.

Example 2: Marginal Model for Counts

Next suppose that Y_{ij} is a count and we wish to relate changes in the expected count (or expected rate) to the covariates. Counts are often modeled as Poisson random

variables, using a log link function and a Poisson variance function. This motivates the following illustration of a marginal model for Y_{ij}:

1. The mean of Y_{ij} is related to the covariates through a log link function,

$$\log(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta.$$

2. The variance of each Y_{ij}, given the effects of the covariates, depends on the mean response,

$$\text{Var}(Y_{ij}|X_{ij}) = \phi\,\mu_{ij},$$

where ϕ is a time-invariant scale parameter that needs to be estimated.

3. The within-subject association among the vector of repeated responses is assumed to have an unstructured pairwise correlation pattern

$$\text{Corr}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha_{jk}.$$

Here a balanced longitudinal design is assumed and the vector of parameters α represents the pairwise correlations among the responses.

The marginal model specified above is a log-linear regression model, with an extra-Poisson variance assumption. The within-subject association is specified in terms of an unstructured pairwise correlation pattern. Of course, other choices for the link and variance functions are possible; similarly other models for the correlation (e.g., first-order autoregressive correlation pattern) are also possible. In this example the extra-Poisson variance assumption allows the variance to be inflated by a factor ϕ (when $\phi > 1$). In many applications count data have variability that far exceeds that predicted by the Poisson distribution; this phenomenon is referred to as *overdispersion*. Indeed, many statisticians believe that overdispersion is the rule, not the exception, when dealing with count data. The excess variability can be accounted for by including the scale factor ϕ in the specification of the variance.

Example 3: Marginal Model for a Binary Response

Suppose that Y_{ij} is a binary response, taking values of 0 (denoting "failure") or 1 (denoting "success"), and it is of interest to relate changes in $E(Y_{ij}|X_{ij}) = \Pr(Y_{ij} = 1|X_{ij})$ to the covariates. With a binary response the distribution of each Y_{ij} is Bernoulli, and the probability of success is often modeled using a logit or probit link function. Recall that for a Bernoulli random variable, the variance is a known function of the mean. This motivates the following illustration of a marginal model for Y_{ij}:

1. The mean of Y_{ij}, or probability of success, is related to the covariates by a logit link function,

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij} = X'_{ij}\beta.$$

2. The variance of each Y_{ij}, given the effects of the covariates, depends on the mean response,

$$\text{Var}\,(Y_{ij}|X_{ij}) = \mu_{ij}(1-\mu_{ij})$$

and $\phi = 1$.

3. The within-subject association among the vector of repeated responses is assumed to have an unstructured pairwise log odds ratio pattern,

$$\log \text{OR}\,(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha_{jk},$$

where

$$\text{OR}(Y_j, Y_k) = \frac{\Pr(Y_j = 1, Y_k = 1)\Pr(Y_j = 0, Y_k = 0)}{\Pr(Y_j = 1, Y_k = 0)\Pr(Y_j = 0, Y_k = 1)}.$$

The marginal model specified above is a logistic regression model, with a Bernoulli variance assumption, $\text{Var}\,(Y_{ij}|X_{ij}) = \mu_{ij}(1-\mu_{ij})$, and an unstructured within-subject association specified in terms of pairwise log odds ratios rather than pairwise correlations (recall the discussion in Section 12.2 on why the odds ratio is a preferable metric for association among pairs of binary responses).

Example 4: Marginal Model for an Ordinal Response

Finally, suppose that Y_{ij} is an ordinal response with K categories $(1, \ldots, K)$ and it is of interest to relate changes in the ordinal response to the covariates. To do so, we can specify a marginal model for the *cumulative response probabilities*. For example, a natural extension of the proportional odds model to longitudinal data is

$$\log\left\{\frac{\Pr(Y_{ij} \le k|X_{ij})}{\Pr(Y_{ij} > k|X_{ij})}\right\} = \alpha_k + X'_{ij}\beta, \quad (k = 1, ..., K-1).$$

In this model, changes in the $K-1$ cumulative logits over time are related to the covariates. Although the model includes $K-1$ intercepts (α_k's), it assumes that the effects of covariates are the same across the $K-1$ cumulative logits; this is equivalent to assuming the covariate effects on the cumulative odds are proportional.

Recall from Section 11.4 that the construction of a generalized linear model for the cumulative probabilities requires treating the ordinal response as a set of $K-1$ binary variables. Therefore, with repeated measures of an ordinal response, Y_{ij} can be replaced by $K-1$ binary responses,

$$U_{ijk} = \begin{cases} 1 & \text{if } Y_{ij} \le k, \\ 0 & \text{if } Y_{ij} > k, \end{cases}$$

for $k = 1, ..., K-1$. That is, the ordinal response at each occasion, Y_{ij}, is replaced by a vector of binary variables, $(U_{ij1}, ..., U_{ij,K-1})'$ for the $K-1$ dichotomizations of the ordinal response. As before, we index subjects by i and occasions by j; however,

we now require a third index k to distinguish the $K-1$ dichotomizations of the ordinal response.

Before constructing a marginal model for the ordinal response, we note that the components of $(U_{ij1}, ..., U_{ij,K-1})'$ are correlated, in the sense that $\text{Corr}(U_{ijk}, U_{ijk'}) \neq 0$ for $k \neq k'$. For example,

$$\text{Corr}(U_{ij1}, U_{ij2}) = \frac{F_{ij1} - F_{ij1}F_{ij2}}{\sqrt{F_{ij1}F_{ij2}(1-F_{ij1})(1-F_{ij2})}},$$

where $F_{ij1} = \Pr(U_{ij1} = 1) = \Pr(Y_{ij} \leq 1)$ and $F_{ij2} = \Pr(U_{ij2} = 1) = \Pr(Y_{ij} \leq 2)$. This correlation is a direct consequence of the fact that the K multinomial response probabilities must necessarily sum to 1. Although the components of $(U_{ij1}, ..., U_{ij,K-1})$ are correlated, the correlations can be expressed in terms of known functions of the cumulative probabilities, F_{ijk}.

An example of a marginal model for the cumulative response probabilities is given by the following three-part specification:

1. The cumulative probabilities are related to the covariates through a logit link function,

$$\text{logit}\{\Pr(Y_{ij} \leq k|X_{ij})\} = \alpha_k + X'_{ij}\beta, \quad (k = 1, ..., K-1).$$

   Letting $F_{ijk} = \Pr(U_{ijk} = 1|X_{ij}) = \Pr(Y_{ij} \leq k|X_{ij})$, this model can be expressed by relating the mean of U_{ijk} to the covariates through a logit link function,

$$\text{logit}\,(F_{ijk}) = \alpha_k + X'_{ij}\beta, \quad (k = 1, ..., K-1).$$

2. The variance of each U_{ijk}, given the effects of the covariates, depends only on the mean response,

$$\text{Var}\,(U_{ijk}|X_{ij}) = F_{ijk}(1 - F_{ijk}).$$

3. In specifying the within-subject association, the correlation among the components of $(U_{ij1}, ..., U_{ij,K-1})$ at the j^{th} occasion are known functions of the means at that occasion, F_{ijk}. The associations between the components at different occasions are assumed to have an unstructured pairwise pattern.

The marginal model specified above is a proportional odds model for the repeated ordinal response, with a multinomial variance (and covariance) assumption, and an unstructured within-subject association between pairs of repeated measurements.

The four examples of marginal models considered so far are purely illustrative. They demonstrate how the specification of the three components of a marginal model might differ according to the type of response variable. However, these four examples should not be considered prescriptions for constructing marginal models; in principle, any suitable link function can be chosen and other assumptions about the variances and within-subject associations can be made. The choices for the three components of a marginal model should reflect statistical and subject-matter considerations. In Chapter 13 we present three case studies that illustrate the application of marginal models to longitudinal data.

12.4 DISTRIBUTIONAL ASSUMPTIONS FOR MARGINAL MODELS*

In Section 12.2 a marginal model was defined in terms of a three-part formulation. This formulation highlights how generalized linear models have been extended to handle longitudinal data. In this section† we consider making additional distributional assumptions about the vector of responses, Y_i. Previously we mentioned that specification of the mean vector and the covariance (or the variance and pairwise associations) does not, in most cases, determine the joint distribution of discrete longitudinal data. That is, the three-part marginal model specification does not determine the joint distribution of Y_i. As a result the method of maximum likelihood cannot be used for estimation of the parameters in the marginal model without further distributional assumptions. This presents two alternative ways to proceed.

The first is to attempt to enrich the formulation of the marginal model so that full distributional assumptions about Y_i have been made. Then the likelihood can be specified and the method of maximum likelihood can be used for estimation and inference. However, this poses a number of difficulties. First, unlike the multivariate normal distribution for a continuous response, the joint distribution of Y_i is not usually specified by the mean vector and covariance matrix. That is, with discrete longitudinal data there is no simple analogue of the multivariate normal distribution. Instead, the joint distribution of Y_i requires specification of the mean vector and pairwise (or two-way) associations, as well as the three-, four-, and higher-way associations among the responses. As the number of responses increases, the number of association parameters proliferates rapidly. This is best exemplified in the case where Y_i is a vector of binary responses. When the number of repeated measures $n_i = 10$, the joint distribution of Y_i has 1013 (or $2^{10} - 10 - 1$) two-way, three-way, four-way, and higher-way association parameters. This excessive number of within-subject association parameters will often far exceed the number of subjects enrolled in a longitudinal study. As a result specification of the joint distribution for discrete longitudinal data is inherently difficult. In addition, even in cases where it might be possible to specify the joint distribution of Y_i, the likelihood is often intractable and maximum likelihood estimation is computationally infeasible. Furthermore procedures for ML estimation of marginal models are not currently incorporated in commercially available general-purpose statistical software packages.

The second alternative is to avoid distributional assumptions about Y_i altogether and specify the marginal model solely in terms of assumptions about the mean response, the variances, and the pairwise (or two-way) within-subject association. This corresponds to the three components in the formulation given in Section 12.2. This alternative approach has the following three advantages. First, it leads to a method for estimation and inference that does not require any distributional assumptions on Y_i. As a result the empirical researcher does not have to be concerned that the distribu-

†This section provides a rationale for the use of the generalized estimating equations (GEE) approach for marginal models presented in Chapter 13. The content of this section is somewhat technical and can be omitted at first reading without loss of continuity.

tion of Y_i closely approximates some multivariate distribution. Put another way, there may be a gain in robustness because distributional assumptions on Y_i are not required. Second, it circumvents the need to specify models for the three-way, four-way, and higher-way associations among the responses. Modeling three-way, four-way, and higher-way associations among the responses is conceptually very difficult, and ordinarily requires a relatively large sample size. Third, it leads to a method of estimation known as generalized estimating equations (GEE); the GEE approach is described in detail in Chapter 13. The GEE approach has become an extremely popular method for analyzing longitudinal data, and for good reasons too. It provides a flexible approach for modeling the mean and the pairwise within-subject association structure. It can handle inherently unbalanced designs and missing data with ease. Finally, the GEE approach is computationally straightforward and has been implemented in existing, widely available statistical software. The one potential drawback that must be acknowledged is that avoidance of distributional assumptions will usually result in some loss of efficiency for estimation of β relative to the optimal, but intractable, likelihood-based estimates. In addition there are some implications for the assumptions made about missing responses; the latter issue will be addressed in Chapters 17 and 18. However, given that the distinct advantages of this alternative approach far outweigh its drawbacks, this is the approach that we emphasize in Chapter 13.

12.5 FURTHER READING

A very accessible description of marginal models, and the generalized estimating equations approach, can be found in Chapter 11 of the textbook by Agresti (2002).

Bibliographic Notes

There is an extensive statistical literature on likelihood-based marginal models for discrete longitudinal data. Bahadur (1961) proposed a model for the vector of repeated responses expressed in terms of pairwise and higher-order correlations among the responses. Because of the restrictions on the correlations, alternative multinomial models for the joint distribution of the vector of discrete responses have been proposed where the within-subject association is parameterized in terms of other metrics of association. For example, Dale (1984), McCullagh and Nelder (1989), Lipsitz, Laird, and Harrington (1990), Liang, Zeger, and Qaqish (1992), Becker and Balagtas (1993), Molenberghs and Lesaffre (1994), Lang and Agresti (1994), Glonek and McCullagh (1995), and others have proposed full likelihood approaches where the higher-order moments are parameterized in terms of marginal odds ratios. In closely related work, Ekholm (1991) parameterized the association directly in terms of the higher-order marginal probabilities (see also Ekholm, Smith, and McDonald, 1995). An alternative approach parameterizes the within-subject association in terms of conditional associations, leading to so-called "mixed parameter" models (Fitzmaurice and Laird, 1993; Glonek, 1995; Molenberghs and Ritter, 1996).