Chapter 2
Longitudinal Data Analysis

Wan Tang, Naiji Lu, Rui Chen, and Hui Zhang

2.1 Introduction

Longitudinal studies are quite common in modern clinical trials and cohort studies. Unlike cross-sectional designs, where observations from study subjects are available only at a single time point, individuals in longitudinal or cohort studies are assessed repeatedly over time. By taking advantages of multiple snapshots of a group over time, data from longitudinal studies captures both between-individual differences and within-individual dynamics, affording the opportunity to study more complicated biological, psychological, and behavioral hypotheses than their cross-sectional counterparts.

For example, if we want to test whether exposure to some chemical agent can cause some type of cancer, the between-subject difference observed in cross-sectional data can only provide evidence of an association or correlation between the exposure and disease. The within-individual dynamics in longitudinal data allows for inference of a causal nature for such a relationship.

Longitudinal data presents multiple methodological challenges in study designs and data analyses. The primary problem is the correlation among the repeated responses of the same subject. Classic models for cross-sectional data analysis such as multiple linear and logistic regressions are based on the independence of observations and thus in general do not apply to longitudinal data. For example, in

W. Tang (✉) • N. Lu • R. Chen
Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA
e-mail: wan_tang@urmc.rochester.edu; naiji_lu@urmc.rochester.edu; rui_chen@urmc.rochester.edu

H. Zhang
Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
e-mail: hui.zhang@stjude.org

W. Tang and X. Tu (eds.), *Modern Clinical Trial Analysis*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-4322-3_2,
© Springer Science+Business Media New York 2013

studies to compare quality of life (QOL) between two treatment conditions, patients are repeatedly assessed for their quality of life over time. Since such repeated assessments are unavoidably correlated, special handling of such correlation is required in order to compare the QOL measures between the treatment conditions.

Another issue common to longitudinal studies is missing values. Since patients are followed up for a period of time in longitudinal studies, it is common that some patients will drop out of the studies. This is of particular importance for cancer research because of the high mortality rate and severely deteriorated health condition of the patients. One naive approach would be to simply delete the subjects with any missing value during the study period. However, as it does not utilize all available data, such an approach is not efficient. But a more serious issue is the bias in the estimate arising from "listwise" deletions, yielding misleading findings and even wrong conclusions, since the data is obviously biased towards those who survive at the end of the study.

Note that in traditional cancer studies, the primary interest is often the time to some significant event such as death. In such situations, each patient is followed for a certain period of time to record the occurrence as well as the timing of the event, and the interest is to estimate the distribution of such *survival times*. As the event may not occur for some patients during the study period, the event times are *censored* for these subjects, requiring specialized methods in survival analysis to handle such censored event times. Although longitudinal data also involves time, it is only used to index the temporal assessment of the subject, rather than the primary focus as in survival analysis. Thus, longitudinal models in general bear little resemblance to those in survival analysis methodology.

2.2 Parametric Models

If the data is well behaved, we may be able to model it using mathematical distributions such as the multivariate normal. Such parametric models have been and still are used for modeling longitudinal data, although semi-parametric or distribution-free alternatives have become increasingly popular. We start with the classic normal-based multivariate linear regression model.

2.2.1 Multivariate Linear Regression Model

Consider a longitudinal study with n subjects assessed at m different time points. Let y_{it} denote some continuous response, dependent variable, and $\mathbf{x}_{it}$ a vector of independent variables, also known as covariates or predictors, from the ith subject at assessment time t. Let

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top, \quad \mathbf{x}_i = \left(\mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{im}^\top\right)^\top.$$

Then, we model the linear relation between y_{it} and $\mathbf{x}_{it}$ as follows:

$$y_{it} = \mathbf{x}_{it}^{\top}\beta + \varepsilon_{it}, \quad \text{or} \quad \mathbf{y}_i = \mathbf{x}_i^{\top}\beta + \varepsilon_i,$$
$$\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})^{\top} \sim \text{i.i.d. } N(0, \Sigma), \tag{2.1}$$

where ε_{it} (ε_i) denotes the model error and i.i.d. stands for independently and identically distributed. Note that if different β_t are used for different times, the model in (2.1) is known as the *seemingly unrelated regression*.

In many controlled, randomized longitudinal trials, we are interested in treatment difference at a posttreatment or a follow-up time ($1 < t \leq m$). In this case, we can apply the Analysis of Variance (ANOVA) model for comparing different treatment conditions or the Analysis of Covariance (ANCOVA) if we want to control for covariates at pre-treatment or baseline ($t = 1$). These methods yield valid inference in the absence of missing data. However, when there is missing data, both generally gives rise to biased estimates. We address this issue in Sect. 2.4.

Although multivariate normality-based models are widely used in biomedical and psychosocial research, many studies have indicated problems with estimates derived from such models because of the strong distribution assumption imposed. For example, most instruments used for QOL measures are based on item scores, which are intrinsically discrete. If a variable has a relatively large range such as the total score of the popular QOL instrument SF-36, the conventional approach by treating such a variable as a continuous outcome is sensible. However, because these variables are inherently discrete, normal-based parametric models are fundamentally flawed and, in some cases, such distribution assumptions may be severely violated (Lu et al. 2009). Thus, whenever possible, distribution-free alternatives should also be considered.

2.2.2 Linear Mixed-Effects Model

The two major limitations of the classic multivariate linear model are (1) its limited ability to deal with missing data, and (2) its requirement of common assessment times for all subjects. The mixed-effects (or latent variable) modeling approach and the distribution-free models provide an effective solution to both issues. We start with the linear mixed-effects model (LMM) for continuous responses, which is a direct extension of the classic multivariate linear models.

The LMM is a general class of models widely used to model the linear relationship between a (continuous) response and a set of independent variables within a longitudinal data setting. LMM addresses the correlated responses by modeling the between-subject variability using random effects, or latent variables, rather than directly correlating the responses as in the classic multivariate linear model. As a result, this approach enables one to address the difficulty in modeling correlated responses arising from varying assessment times as in some longitudinal cohort studies.

As data clustering arises in research studies employing longitudinal study designs and multi-level sampling strategies (e.g., sampling subjects from classes nested within schools) across a wide range of disciplines, various applications of LMM are found under different guises such as random coefficient models, random regression, hierarchical linear models (HLM), latent variable models, mixed models, and multilevel linear models (Goldstein 1987; Bryk et al. 1996; Laird and Ware 1982; Strenio et al. 1983).

Consider a longitudinal study with n subjects. Assume first a set of fixed assessment times for all subjects, $t = 1, 2, \ldots, m$. If the mean of y_{it} is a linear function of time t, then the classic linear model has the form:

$$\begin{aligned} &y_{it} = \beta_0 + \beta_1 t + \varepsilon_{it}, \quad \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})^\top \sim \text{i.i.d. } N(\mathbf{0}, \Sigma), \\ &1 \le i \le n, \quad 1 \le t \le m. \end{aligned} \tag{2.2}$$

In the above, Σ is the variance of $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top$, which contains both the between- and within-subject variation. The idea behind the LMM is to break up the two sources of variation by modeling each separately.

For each subject i, let b_{i0} and b_{i1} denote the intercept and slope of the response y_{it} and let $\mathbf{b}_i = (b_{i0}, b_{i1})^\top$. Then, for each subject i, we model the within-subject variation as follows:

$$\begin{aligned} &y_{it} \mid \mathbf{b}_i = \beta_0 + \beta_1 t + b_{i0} + b_{i1} t + \varepsilon_{it}, \\ &\varepsilon_{it} \sim \text{i.i.d. } N\left(0, \sigma^2\right), \quad 1 \le t \le m. \end{aligned} \tag{2.3}$$

In other words, for the ith individual, the response y_{it} is modeled as a linear function of time with intercept, $\beta_0 + b_{i0}$, and slope, $\beta_1 + b_{i1}$. Thus, by modifying the population mean $\beta = (\beta_0, \beta_1)^\top$ using the individual specific $\mathbf{b}_i$ to account for between-subject differences in the linear predictor, the error terms $\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{im}$, in (2.3) can be assumed to be i.i.d. This is in stark contrast to the assumption of the error term for the classic multivariate linear regression in (2.2), in which the model error ε_{it} are correlated over t to account for individual differences. By letting $\mathbf{b}_i$ vary across the subjects, we obtain a LMM that accounts for both between- and within-subject variations. In many applications, $\mathbf{b}_i$ are assumed to follow a multivariate normal $N(\mathbf{0}, D)$.

By combining the two-level specifications, we can express this LMM in a hierarchical form:

$$\begin{aligned} &y_{it} = \beta_0 + t\beta_1 + b_{i0} + b_{i1} t + \varepsilon_{it} = \mathbf{x}_{it}^\top \beta + \mathbf{z}_{it}^\top \mathbf{b}_i + \varepsilon_{it}, \\ &\mathbf{y}_i = \mathbf{x}_i^\top \beta + \mathbf{z}_i^\top \mathbf{b}_i + \varepsilon_i, \\ &\varepsilon_{it} \sim \text{i.i.d. } N\left(0, \sigma^2\right), \quad \mathbf{b}_i = N(b_{i0}, b_{i1})^\top \sim \text{i.i.d. } N(\mathbf{0}, D), \end{aligned} \tag{2.4}$$

where $\mathbf{x}_{it} = \mathbf{z}_{it} = (1, t)^\top$ and $\mathbf{x}_i = \mathbf{z}_i = \left(\mathbf{x}_{i1}^\top, \mathbf{x}_{i2}^\top, \ldots, \mathbf{x}_{im}^\top\right)^\top$. The linear predictor of the LMM in (2.4) has two parts; the first part $\mathbf{x}_{it}^\top \beta$ describes the change of the population mean over time, while the second $\mathbf{z}_{it}^\top \mathbf{b}_i$ models the deviation of each individual subject from the population mean. Since $\mathbf{b}_i$ is random, $\mathbf{b}_i$ (or $\mathbf{z}_{it}^\top \mathbf{b}_i$) is called the *random effect*. The vector of population-level parameters β (or $\mathbf{x}_{it}^\top \beta$) is called the *fixed effect* and hence the name of the LMM. The hierarchical form and the latent nature of $\mathbf{b}_i$ explain the alternative names of the LMM mentioned earlier.

The covariance between the sth and tth assessment points for the classic (2.2) and linear mixed-effects (2.3) models are given by

$$\text{Cov}\,(y_{is}, y_{it}) = \sigma_{st}, \quad \text{Cov}\,(y_{is}, y_{it}) = \mathbf{z}_{is}^\top D \mathbf{z}_{it} + \sigma^2, \quad 1 \le s, t \le m.$$

If assessment times vary across individuals, σ_{ist} will depend on i and become inestimable except in special cases with some particular covariance structures. For example, if the covariance between y_{is} and y_{it} follows the uniform compound symmetry assumption, $\sigma_{st} = \sigma$ and is estimable. However, this issue does not arise for the LMM, since D and σ^2 are well defined regardless of the spacing structure of assessment. Further, the variance parameters for the random effect in LMM have well-defined interpretations; the diagonals of D measure the variability in individual intercepts and slopes among different subjects in the study population. Thus, in addition to β, inference about D is also often of interest to assess the variability of individual intercepts or slopes or both.

Suppose that each subject has a varying number as well as times of assessments. An LMM for this general setting has the following form:

$$\begin{aligned} y_{it_{ij}} &= \mathbf{x}_{it_{ij}}^\top \beta + \mathbf{z}_{it_{ij}}^\top \mathbf{b}_i + \varepsilon_{it_{ij}}, \quad \mathbf{y}_i = \mathbf{x}_i \beta + \mathbf{z}_i \mathbf{b}_i + \varepsilon_i, \quad \mathbf{b}_i \perp \varepsilon_i, \\ \mathbf{b}_i &\sim \text{i.i.d. } N\,(0, D)\,, \quad \varepsilon_i \sim \text{i.i.d. } N\left(0, \sigma^2 \mathbf{I}_{m_i}\right), \quad 1 \le j \le m_i, \end{aligned} \tag{2.5}$$

where $\mathbf{x}_{it}^\top \beta$ is the fixed and $\mathbf{z}_{it}^\top \mathbf{b}_i$ the random effect, $\perp$ denotes stochastic independence, $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, and $\mathbf{u}_i = (\mathbf{u}_{it_{i1}}^\top, \ldots, \mathbf{u}_{it_{im_i}}^\top)^\top$. For growth-curve analysis (change of y_{it} over time as in longitudinal studies), $\mathbf{z}_{it}$ is often equal to $\mathbf{x}_{it}$. It follows from the assumptions of the LMM that

$$E\,(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}_i^\top \beta, \quad \text{Var}\,(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i) = \mathbf{z}_i^\top D \mathbf{z}_i + \sigma^2 \mathbf{I}_{m_i}. \tag{2.6}$$

Clustered data also often arises from nested studies. For example, in a multicenter trial, subjects are nested within each center or site, causing clustered responses even when analyzing the data at a single time point. For simplicity, consider modeling treatment differences at a posttreatment assessment time in a multi-site, randomized trial with two treatment conditions. Let y_{ij} denote some response of interest from the jth subject within the ith site and x_{ij} be a binary indicator for the treatment received by the j subject at the ith site. Then, an appropriate LMM is given by:

$$y_{ij} = \beta_0 + x_{ij} \beta_1 + b_i + \varepsilon_{ij}, \quad b_i \sim \text{i.i.d. } N\left(0, \sigma_b^2\right), \quad \varepsilon_{ij} \sim \text{i.i.d. } N\left(0, \sigma^2\right),$$

where b_i denotes the (random) effect of site. We can assess whether there is a significant site effect by testing the null: $H_0 : \sigma_b^2 = 0$. If this null is not rejected, we can simplify the model by dropping b_i.

When the number of sites is small such as two or three, we may want to model potential site difference using a fixed effect. In this case, site difference is interpreted with respect to the particular sites in the study. If there are a large number of randomly selected sites, it is sensible to model site differences using random effects. A significant random effect implies differences not only among the participating sites but also across similar sites not included in the study.

2.2.3 Generalized Linear Mixed-Effects Models

A major limitation of the LMM is that it only applies to continuous response. To model other types of responses such as binary and count data, we must use the generalized linear mixed-effects model (GLMM). We first review the generalized linear model (GLM), the premise underlying GLMM, and then discuss how this class of models is extended to GLMM by adding random effects.

The GLM frames a wide range of seemingly disparate problems of statistical modeling and inference under a unified framework. GLM extends linear regression for a continuous response to models for other types of response such as binary and categorical outcomes. Examples of GLMs include linear regression, logistic regression for binary outcomes, and log-linear regression for count data. We give a brief review of GLM.

The classic multiple linear regression model has the form

$$\begin{aligned} y_i \mid \mathbf{x}_i &\sim \text{i.d.} N\left(\mu_i, \sigma^2\right), \\ \mu_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \beta_0 + \beta^\top \mathbf{x}_i, \end{aligned} \tag{2.7}$$

where i.d. means independently distributed. The response y_i conditional on the covariates $\mathbf{x}_i$ is assumed to have a normal distribution with mean μ_i and common variance σ^2. In addition, μ_i is a linear function of $\mathbf{x}_i$. Since the right side of the model, $\eta_i = \beta_0 + \beta^\top \mathbf{x}_i$, has a range in the real line R, concurring with the range of μ_i on the left side, the linear model is not appropriate for modeling other types of noncontinuous responses. For example, if y_i is binary, the conditional mean of $y_i \mid \mathbf{x}_i$ is

$$\mu_i = E\left(y_i \mid \mathbf{x}_i\right) = \Pr\left(y_i = 1 \mid \mathbf{x}_i\right). \tag{2.8}$$

Since μ_i is a value between 0 and 1, it is not sensible to model μ_i directly as a linear function of $\mathbf{x}_i$ as in (2.7). In addition, the normal distribution assumption does not apply to binary responses.

To generalize the linear model to accommodate other types of response, we must modify (1) the normal distribution assumption; and (2) the relationship between the

conditional mean μ_i in (2.8) and the linear predictor η_i in (2.7). GLM addresses both issues by extending (2.7) in the respective directions:

1. *Random component.* This part specifies the conditional distribution of the response y_i given the dependent variables $\mathbf{x}_i$.
2. *Deterministic component.* This part links the conditional mean of y_i given $\mathbf{x}_i$ to the linear predictor $\mathbf{x}_i$ by a one-to-one *link* function g:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \beta, \quad \text{or} \quad \mu_i = h\left(\mathbf{x}_i^\top \beta\right), \quad h = g^{-1}. \tag{2.9}$$

Thus, the linear regression is obtained as a special case if y given $\mathbf{x}$ follows a normal distribution, and $g(\mu)$ is the identity function, $g(\mu_i) = \mu_i$. By varying the distribution function for the random component and the link function $g(\cdot)$ in the deterministic part, we can use GLM to model a variety of response types with different distributions. For example, if a binary response y given $\mathbf{x}$ follows a Bernoulli distribution $\text{Bern}(\mu)$ with the probability of success given by: $E(y \mid \mathbf{x}) = \pi(\mathbf{x}) = \pi$ and the conditional mean μ is linked to the linear predictor η by the logit function, $\eta = g(\pi) = \log(\pi/(1-\pi))$, we obtain the logistic regression model. For a count response y, GLM yields a Poisson log-linear model, if y given $\mathbf{x}$ follows a Poisson distribution $\text{Poisson}(\mu)$, with the mean μ linked to the linear predictor by the log function.

To extend GLM to a longitudinal data setting, we simply add random effects, akin to LMM. Consider a study with n subjects and m assessments. For notational brevity, we assume a set of fixed assessment times, $1 \le t \le m$. Let y_{it} denote some response and $\mathbf{x}_{it}$ a vector of covariates from the ith subject at time t ($1 \le i \le n$, $1 \le t \le m$). The principle of extending GLM to a longitudinal data setting is the same as in generalizing the classic univariate linear regression to LMM.

For each subject, we first model the within-subject variability using a GLM:

$$y_{it} \mid \mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{b}_i \sim \text{i.d. } f(\mu_{it}), \quad g(\mu_{it}) = \mathbf{x}_{it}^\top \beta + \mathbf{z}_{it}^\top \mathbf{b}_i, \tag{2.10}$$

where $g(\cdot)$ is a link function, $\mathbf{z}_{it}$ is a sub-vector of $\mathbf{x}_{it}$, $\mathbf{b}_i$ denotes the random effects, and $f(\mu)$ some probability distribution with mean μ. Note that the model specification in (2.10) is quite similar to the cross-sectional GLM except for the added individual effect $\mathbf{b}_i$. By adding a distribution for $\mathbf{b}_i$ to explain the between-subject variation, we obtain from (2.10) the class of *GLMM*. As in the case of LMM, $\mathbf{b}_i$ is often assumed to follow a multivariate normal $\mathbf{b}_i \sim N(\mathbf{0}, D)$ (a common assumption in most studies), although other types of more complex distributions such as mixtures of normals may also be specified.

To use GLMM for modeling a binary y_{it}, we set $f(\mu_{it}) = \text{Bern}(\mu_{it})$, a Bernoulli with mean μ_{it}. The most popular link for modeling such a response is the logit function. If we model the trajectory of y_{it} over time as a linear function of t with a bivariate normal random effect for the mean and slope, the GLMM becomes

$$y_{it} \sim \text{i.d. BI}\left(\mu_{it}; 1\right), \quad \text{logit}\left(\mu_{it}\right) = \mathbf{x}_{it}^{\top}\beta + \mathbf{z}_{it}^{\top}\mathbf{b}_i,$$
$$\mathbf{b}_i \sim \text{i.i.d. } N\left(\mathbf{0}, D\right), \quad 1 \le i \le n, 1 \le t \le m, \tag{2.11}$$

where $\mathbf{x}_{it}^{\top} = \mathbf{z}_{it}^{\top} = (1, t)^{\top}$ and $\text{logit}(\mu_{it}) = \log(\frac{\mu_{it}}{1-\mu_{it}})$. As in the LMM case, $\beta_0 + \beta t$ describes the change over time for the population average, while the random effect $b_{i0} + b_{i1}t$ accounts for individual deviations.

For a count response y_{it}, we may assuming a Poisson distribution $\text{Poisson}\left(\mu_{it}\right)$ and a log link $\log\left(\mu_{it}\right)$ to obtain a random-effects based the log-linear model for the trajectory of a count response y_{it} over time:

$$y_{it} \sim \text{i.d. Poisson}\left(\mu_{it}\right), \quad \log\left(\mu_{it}\right) = \mathbf{x}_{it}^{\top}\beta + \mathbf{z}_{it}^{\top}\mathbf{b}_i,$$
$$\mathbf{b}_i \sim \text{i.i.d. } N\left(\mathbf{0}, D\right), \quad 1 \le i \le n, 1 \le t \le m. \tag{2.12}$$

The fixed and random effects have the same interpretation as in the GLMM for binary data.

In addition to binary and count response, we can similarly generalize the generalized logit and proportional odds models for multi-level nominal and ordinal responses. Interested readers may consult Kowalski and Tu (2008); Tang et al. (2012) for details.

2.2.4 Maximum Likelihood Inference

As the multivariate linear regression in (2.1) is not widely used for longitudinal data analysis, we discuss inference only for the LMM and the GLMM. Readers interested in inference for the classic linear models may consult Seber (1984); Kowalski and Tu (2008). We start with LMM.

Consider inference for the LMM in (2.5). Let $\theta = \left(\beta^{\top}, \text{vec}^{\top}\left(D\right), \sigma^2\right)^{\top}$, where $\text{vec}^{\top}\left(D\right)$ denotes the vector operator to convert the symmetric $q \times q$ matrix D into a column $\frac{1}{2}q\left(q+1\right) \times 1$ vector consisting of the $\frac{1}{2}q\left(q+1\right)$ distinct elements of D. The log-likelihood is given by

$$l_n\left(\theta\right) = \sum_{i=1}^{n} \log\left[\int f_{y|x,z,b}\left(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{b}_i\right) f_b\left(\mathbf{b}_i\right) \mathrm{d}\mathbf{b}_i\right], \tag{2.13}$$

where $f_{y|x,z,b}\left(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i, \mathbf{b}_i\right)$ denotes the density function of a multivariate normal $N\left(\mu_i, \Sigma_i\right)$ with mean $\mu_i = \mathbf{x}_i^{\top}\beta + \mathbf{z}_i^{\top}\mathbf{b}_i$ and variance $\Sigma_i = \mathbf{z}_i^{\top} D\mathbf{z}_i + \sigma^2\mathbf{I}_{m_i}$, and $f_b\left(\mathbf{b}_i\right)$ the density function of the normal random effect $\mathbf{b}_i$. The integral in (2.13) is the result of integrating out the latent $\mathbf{b}_i$.

A major technical problem with inference for mixed-effect models is how to deal with such an integral. Fortunately for the normal–normal based LMM, this

integral can be completed in closed form. Since the marginal of $\mathbf{y}_i$ is again normal, $N\left(\mathbf{x}_i^\top \beta, V_i = \left(\mathbf{z}_i^\top D \mathbf{z}_i + \sigma_b^2 \mathbf{I}_{m_i}\right)\right)$, the log-likelihood function is given by

$$l_n(\theta) = -\frac{1}{2}\left[N \ln \sigma^2 + \sum_{i=1}^{n} \log |V_i| + \sigma^2 \sum_{i=1}^{n} \mathbf{e}_i^\top V_i \mathbf{e}_i\right], \tag{2.14}$$

where $N = \sum_{i=1}^{n} m_i$ and $\mathbf{e}_i = \mathbf{y}_i - \mathbf{x}_i^\top \beta$. Note that although the dimension of D is fixed, V_i and $\mathbf{e}_i$ both have a varying dimension from subject to subject. The maximum likelihood estimate (MLE) of θ is obtained by maximizing the log-likelihood $l_n(\theta)$.

Although straightforward in principle, it is actually quite difficult to maximize $l_n(\theta)$ in (2.14), since obtaining the derivatives of $l_n(\theta)$ in closed form is quite a daunting task. Because of the analytic complexity in calculating the derivatives, several algorithms have been proposed. Among them, a popular approach is to use the expectation/maximization (EM) algorithm and its various enhanced versions such as the expectation/maximization either (ECME) algorithm (Dempster et al. 1977; Meng and Van Dyk 1997). These approaches maximize the alternative, expected log-likelihood, allowing one to obtain the MLE in closed form. However, such algorithms are notorious for their slow convergence. In addition, they require additional methods to compute its asymptotic variance, the latter of which is often quite complex and a problem in its own right. Most software packages use the Newton–Raphson algorithm based on the analytic first- and second-order analytic derivatives of $l_n(\theta)$, which has a much faster convergence rate than EM type algorithms (Lindstrom and Bates 1988; Wolfinger and O'connell 1993; Demidenko 2004).

Inference for GLMM with noncontinuous responses such as binary and count outcomes is more difficult, since unlike LMM integration cannot even be completed in closed form for models with the simplest normal random effects. For example, consider the GLMM for binary response in (2.11). Let $\alpha = \text{vec}^\top(D)$ and $\theta = \left(\beta^\top, \alpha^\top\right)^\top$. Since $\mathbf{b}_i$ is latent, the log-likelihood is given by:

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^{n} \log\left[\int f_{y|x,z,b}(\mathbf{y}_i \mid \mathbf{b}_i, \beta)\, f_b(\mathbf{b}_i \mid \alpha)\, \mathrm{d}\mathbf{b}_i\right] \\ &= \sum_{i=1}^{n} \log\left[\int \prod_{t=1}^{\mathrm{m}} \mu_{it}^{y_{it}} (1-\mu_{it})^{1-y_{it}}\, f_b(\mathbf{b}_i \mid \alpha)\, \mathrm{d}\mathbf{b}_i\right], \end{aligned} \tag{2.15}$$

where $f_{y|x,z,b}(\mathbf{y}_i \mid \mathbf{b}_i, \beta) = \prod_{t=1}^{\mathrm{m}} \mu_{it}^{y_{it}} (1-\mu_{it})^{1-y_{it}}$ is the conditional joint density of m independent binary components of $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top$ given the random effect $\mathbf{b}_i$ with mean $\mu_{it} = \exp(\beta_0 + \beta t + b_{i0} + b_{i1} t)$ and $f_b(\mathbf{b}_i \mid \alpha)$ is the joint density

of the binary normal $\mathbf{b}_i = (b_{i0}, b_{i1})^\top$. Even in this simplest setting, $l_n(\theta)$ cannot be expressed in closed form, making it impossible to obtain $\frac{\partial}{\partial\theta} l_n(\theta)$ in closed form.

Different algorithms have been proposed to tackle the computational challenges. Because of the lack of consensus, different approaches have been adopted by and implemented in major statistical software packages such as SAS and R. As a result, it is common to obtain different estimates for the same model, when different packages are used. Understanding some basics of the computational aspects may be helpful in dealing with the discrepancies in real study applications.

The two most popular approaches for estimation in GLMM as implemented in major software packages such as SAS and R are: (1) approximating the log-likelihood function, and (2) approximating the model. Algorithms in the second category are developed using linearization of or Laplace approximation to the model. They employ some types of expansions such as the Taylor series expansion to approximate the model by one based on pseudo-data with fewer nonlinear components. The process of computing the linear approximation must be repeated several times until some criteria indicate lack of further improvement (Pinheiro and Bates 2000; Schabenberger and Gregoire 1996). The fitting methods based on such a linearization process typically involve two levels of iterations. The GLMM is first approximated by a LMM based on the current values of the covariance parameter estimates, and the resulting LMM is then fit, which itself is an iterative process. On convergence, the new parameter estimates are used to update the linearization, resulting in a new LMM. The iterative process terminates when the difference in parameter estimates between successive LMM fits falls within a specified tolerance level.

The advantages of the linearization approach include a relatively simple form of the linearized model that typically can be fit based only on the mean and variance in the linearized form. This approach can fit models with a large number of random effects, crossed random effects, multiple types of subjects, and even correlated y_{it}'s after conditioning on $\mathbf{x}_{it}$, $\mathbf{z}_{it}$, and $\mathbf{b}_i$. However, as it does not maximize the underlying log-likelihood function, this approach produces estimates with unknown asymptotic properties, except in the special LMM case for which this approach does produce MLE (Demidenko 2004). A recent study based on simulation show that this linearization approach generally provides invalid inference (Zhang et al. 2011b). In addition, algorithms implementing this approach can fail at both levels of the double iteration scheme.

In contrast, the integral approximation approach aims to directly approximate the log-likelihood in (2.15) and maximize the approximated function (Davidian and Gallant 1993; Demidenko 2004; Breslow and Clayton 1993). Since the accuracy of estimates and inference is determined by the quality of the approximation to the log-likelihood, various techniques have been proposed to compute the approximation, with the Newton and Gauss–Hermite quadratures being the most popular. The advantage of this alternative approach is to provide an actual objective function for optimization, albeit it is an approximated log-likelihood. Nonetheless, this enables likelihood-based fit statistics such as the likelihood ratio test to be used for inference.

Further, unlike its linearization counterpart whose approximation accuracy and thus the asymptotic properties of the estimates are limited by the type of models fit (e.g., linear, logistic, etc.), the approximation to the log-likelihood under this approach can be improved to any degree by increasing the precision of numerical integration, at least in principle.

Thus, algorithms based on the integral approximation are expected to provide better estimates than those based on linearization. In particular, since they can get arbitrarily close to the MLE as the precision of numerical integration increases, estimates obtained under this approach have the same nice large sample properties as the MLE such as consistency, asymptotic normality and efficiency. However, the quality of approximation seems to vary tremendously across the different packages. For example, as shown by Zhang et al. (2011a) using simulated data, none of the available functions in R provides valid inference. Thus, if this approach is sought, the NLMIXED procedure in SAS should be used unless improvements have been made to the others, as it is the only software to provide correct inference at the moment of this writing among a number of available packages evaluated based on simulated data by Zhang et al. (2011a).

2.2.5 *Composite Hypothesis Testing*

Hypotheses concerning the parameter vector $\theta = (\beta, \alpha)$ for most applications can be expressed in the following form:

$$H_0 : C\theta = \mathbf{b}, \quad \text{vs.} \quad H_a : C\theta \neq \mathbf{b}, \tag{2.16}$$

where $\mathbf{b}$ is a known constant vector, C is some known full rank $k \times p$ matrix with p ($\geq k$) denoting the dimension of θ. If $\mathbf{b} = \mathbf{0}$, H_0 becomes a linear contrast. If θ only consists of parameters for the fixed-effect, i.e., $\theta = \beta$, we can re-express the linear hypothesis in terms of a linear contrast by performing the transformation: $\gamma = \beta - C^\top \left(C^\top C\right)^{-1} \mathbf{b}$. When expressed in the new parameter vector γ, the linear predictor will contain an offset term.

For example, the linear predictor for the GLMM in (2.10) under this transformation becomes

$$\eta_{it} = \mathbf{x}_{it}^\top \beta + \mathbf{z}_{it_i}^\top \mathbf{b}_i = c_{it} + \mathbf{x}_{it}^\top \gamma + \mathbf{z}_{it_i}^\top \mathbf{b}_i,$$

where $c_{it} = \mathbf{x}_{it}^\top \left(C^\top \left(C^\top C\right)^{-1} \mathbf{b}\right)$ is the offset. Except for LMM, where c_{it} can be absorbed into the response by redefining the dependent variable as $z_{it} = y_{it} - c_{it}$, the offset must be specified when fitting GLMM using software packages.

The most popular tests for linear hypothesis are the Wald, score, and likelihood ratio statistics. We briefly review these tests below.

If $\widehat{\theta} \sim N\left(\theta, \frac{1}{n}\Sigma_\theta\right)$, then it follows from the properties of multivariate normal distribution that $K\widehat{\theta} \sim N\left(K\theta, \frac{1}{n}K\Sigma_\theta K^\top\right)$. Thus, under the null (2.16), $K\widehat{\theta} \sim N\left(\boldsymbol{b}, \frac{1}{n}K\Sigma_\theta K^\top\right)$. The Wald statistic,

$$Q_n^2 = n\left(K\widehat{\theta} - \boldsymbol{b}\right)^\top \left(K\Sigma_\beta K^\top\right)^{-1}\left(K\widehat{\theta} - \boldsymbol{b}\right),$$

follows asymptotically a chi-square distribution with l degrees of freedom $\left(\chi_l^2\right)$, where l is the rank of K. Note that because K is full rank, $K\Sigma_\theta K^\top$ is invertible. The likelihood-ratio statistic is defined as

$$LR = 2\log R\left(\widetilde{\theta}\right) = 2\left[\log L\left(\widehat{\theta}\right) - \log L\left(\widetilde{\theta}\right)\right],$$

where $L(\theta)$ denotes the likelihood function, $\widehat{\theta}$ the MLE of θ, and $\widetilde{\theta}$ the MLE of the constrained model under the null hypothesis. The likelihood-ratio statistic also follows asymptotically a chi-square distribution with l degrees of freedom.

Both the Wald and likelihood ratio tests require the existence and estimate of the MLE of θ. A third popular alternative, the score test, only requires the existence and the MLE of θ under the restricted model. In general, we can reparameterize θ through a linear transformation so that θ can be decomposed as $\theta = \left(\theta_1^\top, \theta_2^\top\right)^\top$, with the null in (2.16) expressed as $\theta_2 = \mathbf{c}$ (a constant vector). Let $l_i(\theta)$ be the log-likelihood associated with the ith subject. The score equation is given by

$$\mathbf{w}_n^{(1)}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\theta)}{\partial\theta_1} = \mathbf{0}, \quad \mathbf{w}_n^{(2)}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\theta)}{\partial\theta_2} = \mathbf{0}.$$

Let $\widetilde{\theta}_1$ denote the MLE estimate of θ_1 under the null obtained by solving $w_n^{(1)}(\theta_1, \mathbf{c}) = \mathbf{0}$. Then the score statistic,

$$T_s\left(\widetilde{\theta}_1, \mathbf{c}\right) = n\left(\mathbf{w}_n^{(2)}\left(\widetilde{\theta}_1, \mathbf{c}\right)\right)^\top \widehat{\Sigma}_2^{-1}\left(\widetilde{\theta}_1, \mathbf{c}\right)\mathbf{w}_n^{(2)}\left(\widetilde{\theta}_1, \mathbf{c}\right),$$

follows a χ_q^2 , where $\widehat{\Sigma}_2$ is the asymptotic variance of $w_n^{(2)}$ and q is the dimension of θ_2. Since only θ_1 is estimated based on $\theta_2 = \mathbf{c}$, the score statistic does not require the existence of MLE for the full model.

As the likelihood ratio test only depends on the height, rather than the curvature of the likelihood function, it usually derives more accurate inference than the Wald statistic. Thus, although asymptotically equivalent, it is generally preferred. Note that the likelihood ratio test is only appropriate for parametric models, while the Wald and score tests also apply to distribution-free models to be discussed next.

2.3 Distribution-Free Models

A major problem with using random effects to account for correlated responses is the difficulty in empirically validating the distribution assumption for the random effects because of their latent or unobservable nature. To further exacerbate the problem, GLMM also relies on parametric assumptions for the response (conditional on the fixed- and random-effects) such as normality for inference. If either of these assumptions is violated, estimates will be biased. Unfortunately, such assumptions are often unacknowledged, laying the basis for inconsistent and spurious findings. A popular alternative is to use models that are free of such assumptions, yielding estimates that are robust to a much wide range of data distributions. We start with distribution-free models for cross-sectional data.

2.3.1 Distribution-Free Models for Cross-Sectional Data

In contrast to the parametric GLM in Sect. 2.2.3, distribution-free alternatives only posit a relationship between the response and the set of predictors without the random component that postulates an analytic distribution for the response (conditional on the predictors), thereby providing more robust estimates regardless of the complexity of the data distribution. As noted in Sect. 2.1, outcomes from most instruments such as those for assessing quality of life are prone to violations of distribution assumptions, because they are not even continuous, let alone following the normal distribution. The removal of the random component of GLM, however, entails serious ramifications for inference about model parameters; without a distribution model specified in this part of GLM, it is not possible to use maximum likelihood for parameter estimation. Thus, we need an alternative paradigm for inference about parameters.

By removing the random component in GLM and modeling the mean of response, we obtain from (2.9) the distribution-free GLM:

$$g(\mu_i) = \mathbf{x}_i^\top \beta, \quad 1 \le i \le n, \tag{2.17}$$

where $g(\cdot)$ has the same interpretation as in (2.9). For example, for a binary (count) response, $g(\mu) = \text{logit}(\mu)$ ($g(\mu) = \log \mu$). Because of the absence of a parametric distribution model, MLEs cannot be computed for distribution-free GLM. Inference is typically based on a set of estimating equations (EE):

$$\begin{aligned} \mathbf{w}_n(\beta) &= \sum_{i=1}^{n} D_i V_i^{-1} S_i = \mathbf{0}, \\ D_i &= \frac{\partial}{\partial \beta} \mu_i, \quad V_i = v(\mu_i), \quad S_i = y_i - \mu_i, \end{aligned} \tag{2.18}$$

In the above, S_i is called the *theoretical residual* (to differentiate it from the *observed residual* with estimated μ_i). The quantity V_i is assumed to be a function of μ_i. With the right selection of V_i, the estimating equations in (2.18) yield the MLE of β for the parametric GLM, when y_i is modeled by the exponential family of distributions (e.g., Kowalski and Tu 2008, Chap. 4).

For example, for a count response y_i following the Poisson, $\text{Var}(y_i \mid \mathbf{x}_i) = \mu_i$. Setting $V_i = \mu_i$ in (2.18) yields the MLE of β, if y_i is modeled by a Poisson log-linear regression. However, the advantage of the estimating equations is that even when y_i does not follow a Poisson the estimate obtained from (2.18) with $V_i = \mu_i$ is still consistent and asymptotically normal.

For example, the Poisson distribution is inappropriate for modeling count data with overdispersion, i.e., $\text{Var}(y_i \mid \mathbf{x}_i) > \mu_i$. Under parametric analysis, we must choose a different model to make allowances for such overdispersions. The negative binomial is a popular choice to address overdispersed count data, as it has the same mean as but a larger variance than the Poisson. However, since the distribution-free model only involves the mean response, the distinction between the Poisson and NB in parametric analysis does not arise for distribution-free models. In fact, such a model yields valid inference regardless of whether y_{it} follows the Poisson, NB, or any other distributions, so long as the relationship between the (conditional) mean and the covariates specified in (2.18) is correct. This feature is particularly important for longitudinal data analysis, since the conditional variance is much more complex and difficult to specify for correlated longitudinal responses.

2.3.2 Distribution-Free Models for Longitudinal Data

Consider a longitudinal study with n subjects and m assessment times, and again for notational brevity, assume a set of fixed assessment times $1 \leq t \leq m$, with y_{it} denoting a response and $\mathbf{x}_{it}$ a set of independent variables of interest from the ith subject at time t, as in the discussion of parametric models in Sect. 2.2. By applying the distribution-free GLM in (2.17) to each time t, we obtain a class of distribution-free regression models within the current context of longitudinal data:

$$E(y_{it} \mid \mathbf{x}_{it}) = \mu_{it}, \quad g(\mu_{it}) = \mathbf{x}_{it}^{\top}\beta, \quad 1 \leq t \leq m, \quad 1 \leq i \leq n, \tag{2.19}$$

where β is a $p \times 1$ vector of parameters of interest. Note that at each time t, the distribution-free GLM above is exactly the same as the cross-sectional version in (2.17). For example, if y_{it} is a continuous response, $\mu_{it} = \mathbf{x}_{it}^{\top}\beta$ models the mean of y_{it} as a function of $\mathbf{x}_{it}$ at each time t, while for a binary y_{it}, $\mu_{it} = E(y_{it})$, which relates to $\mathbf{x}_{it}$ via a link function $g(\mu_{it})$ such as the logit.

Note that y_{it} in (2.19) can be a continuous, binary, or count response. Distribution-free models are also available for multi-level categorical and ordinal responses. See Kowalski and Tu (2008, Chap. 4) for details.

2.3.3 *Inference for Distribution-Free Models*

Consider the class of distribution-free models in (2.19). At each time t, (2.19) reduces to the distribution-free model for cross-sectional data in (2.17). The generalized estimating equations (GEE) are used to provide inference for the longitudinal GLM in (2.19) by extending the estimating equations in (2.18) for a single time t to multiple times across all assessments. This is achieved by capitalizing on the fact that the use of a wrong correlation matrix has no impact on the consistency of the GEE estimate of β, just as in the univariate case that the misspecification of $V(\cdot)$ does not affect the consistency of the estimating equations estimate. Let

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top, \quad \mu_i = (\mu_{i1}, \ldots, \mu_{im})^\top,$$
$$\mathbf{x}_i = \left(\mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{im}^\top\right)^\top, \quad S_i = \mathbf{y}_i - \mu_i.$$

In analogy to (2.18), the GEE are defined by

$$\mathbf{w}_n = \sum_{i=1}^{n} G_i(\mathbf{x}_i) S_i = \sum_{i=1}^{n} G_i(\mathbf{x}_i)(\mathbf{y}_i - \mu_i) = \mathbf{0}, \tag{2.20}$$

where $G_i(\mathbf{x}_i)$ is some matrix function of $\mathbf{x}_i$. In most applications, $G_i(\mathbf{x}_i)$ has the form

$$G_i(\mathbf{x}_i) = D_i V_i^{-1}, \quad D_i = \frac{\partial}{\partial \beta}\mu_i, \quad A_i = \operatorname{diag}(v(\mu_{it})),$$
$$V_i = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}, \tag{2.21}$$

where $R(\alpha)$ denotes a *working correlation* matrix parameterized by α, $\operatorname{diag}(v(\mu_{it}))$ a diagonal matrix with $v(\mu_{it})$ on the tth diagonal. As in the case of cross-sectional data, $v(\mu_{it})$ can be set equal to $\operatorname{Var}(y_{it} \mid \mathbf{x}_{it})$ under some parametric assumptions, such as the mean μ_{it} for a count response based on the Poisson distribution. The phrase "working correlation" is used to emphasize the fact that $R(\alpha)$ is not necessarily the true correlation matrix. For example, we may simply set $R = \mathbf{I}_m$. In this case, the correlated components of $\mathbf{y}_i$ are treated as if they were independent. In addition, there is no parameter associated with this particular *working independence model*. Another popular choice is the uniform compound symmetry correlation matrix, $R(\alpha) = C_m(\rho)$, which assumes a common correlation ρ for any pair of responses y_{is} and y_{it} $(1 \le s, t \le m)$. This working correlation matrix involves a single parameter ρ.

Under the specification in (2.21), (2.20) can be expressed as

$$\mathbf{w}_n(\beta) = \sum_{i=1}^{n} D_i V_i^{-1} S_i = \sum_{i=1}^{n} D_i V_i^{-1}(\mathbf{y}_i - \mu_i) = \mathbf{0}. \tag{2.22}$$

The above is identical to the estimating equations in (2.18) for cross-sectional data analysis, except that D_i in (2.21) is a $p \times m$ matrix rather than a $p \times 1$ vector. Although the GEE in (2.22) is a function of both β and α, we express it explicitly as a function of β to emphasize the fact that (2.22) is used to obtain the estimate of the parameter vector of interest β. If α is known as in the case of working independence model, we can obtain the solution of β to (2.22) by the following recursive relation based on the Newton–Raphson algorithm:

$$\beta^{(k+1)} = \beta^{(k)} + \left(\sum_{i=1}^{n} D_i V_i^{-1} D_i \right)^{-1} \mathbf{w}_n \left(\beta^{(k)} \right), \tag{2.23}$$

where $\beta^{(0)}$ denotes some initial value. The GEE estimate $\widehat{\beta}$ is obtained by iterating the above until convergence.

When α is unknown, we must estimate it so that (2.22) can be used to find estimates of β. For example, consider modeling a binary response y_{it} with the logistic model:

$$E\left(y_{it} \mid \mathbf{x}_{it}\right) = \mu_{it}, \quad \log\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = \mathbf{x}_{it}^{\top}\beta, \quad 1 \le i \le n, 1 \le t \le m. \tag{2.24}$$

Since $\mathrm{Var}\left(y_{it} \mid \mathbf{x}_{it}\right) = \mu_{it}\left(1-\mu_{it}\right)$, $A_i = \mathrm{diag}(\mu_{it}\left(1-\mu_{it}\right))$. If α is known, $\widehat{\beta}$ can be obtained from (2.23). Otherwise, an estimate of α is needed to compute β using the Newton–Raphson algorithm. As an example, consider the completely unstructured working correlation matrix, i.e., $R\left(\alpha\right) = [\rho_{st}]$. We can estimate ρ_{st} by the Pearson correlation:

$$\begin{aligned}
\widehat{\rho}_{st} &= \frac{\sum_{i=1}^{n}\left(\widehat{e}_{is} - \overline{r}_{\cdot s}\right)\left(\widehat{e}_{it} - \overline{\widehat{e}}_{\cdot t}\right)}{\sqrt{\sum_{i=1}^{n}\left(\widehat{e}_{is} - \overline{\widehat{e}}_{\cdot s}\right)^2 \sum_{i=1}^{n}\left(\widehat{e}_{it} - \overline{\widehat{e}}_{\cdot t}\right)^2}}, \quad \widehat{e}_{it} = y_{it} - \mathbf{x}_{it}^{\top}\widehat{\beta}, \\
\overline{\widehat{e}}_{\cdot t} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_{it}, \quad 1 \le i \le n, \quad 1 \le t \le m.
\end{aligned} \tag{2.25}$$

Note that since the correlation $\rho_{i,st} = \mathrm{Corr}\left(y_{is}, y_{it} \mid \mathbf{x}_{is}, \mathbf{x}_{it}\right)$ for binary response generally varies across subjects, a constant $R\left(\alpha\right)$ is hardly ever the true correlation matrix except for some special cases. Further, $\rho_{i,st}$ must also satisfy an additional set of *Frechet bounds* (e.g., Shults et al. 2009):

$$\max\left\{-\left[\frac{\mu_{is}\mu_{it}}{(1-\mu_{is})(1-\mu_{it})}\right]^{1/2}, -\left[\frac{(1-\mu_{is})(1-\mu_{it})}{\mu_{is}\mu_{it}}\right]^{1/2}\right\}$$

$$\leq \rho_{i,st} \leq \min \left\{ \left[\frac{\mu_{is}(1-\mu_{it})}{(1-\mu_{is})\mu_{it}} \right]^{1/2}, \left[\frac{(1-\mu_{is})\mu_{it}}{\mu_{is}(1-\mu_{it})} \right]^{1/2} \right\}. \tag{2.26}$$

Because of these constraints, α is typically selected by some ad hoc rules rather than estimated. For example, under the uniform compound symmetry assumption, $R(\rho) = C_m(\rho)$, we may select ρ to satisfy the *Frechet bounds* in (2.26).

While primary interest lies in β, α must be estimated to proceed with the computation of the GEE estimate of β. Although the consistency of the GEE estimate $\widehat{\beta}$ is independent of how α is estimated, judicious choices of the type of estimates of α not only ensure the asymptotic normality but also simplify the asymptotic variance of $\widehat{\beta}$. To this end, we require that $\widehat{\alpha}$ be $\sqrt{n}$-consistent, i.e., $\widehat{\alpha}$ converges to some point α and $\sqrt{n}(\widehat{\alpha} - \alpha)$ is bounded in probability. Most popular estimates of $\widehat{\alpha}$ such as the moment estimate are asymptotically normal and thus are $\sqrt{n}$-consistent. Given such an estimate of $\widehat{\alpha}$, the GEE estimate $\widehat{\beta}$ is consistent and asymptotically normal, with the asymptotic variance given by

$$\Sigma_\beta = B^{-1} E\left(G_i S_i S_i^\top G_i^\top\right) B^{-\top}, \quad B^{-\top} = \left(B^{-1}\right)^\top. \tag{2.27}$$

A consistent estimate of Σ_β is given by

$$\widehat{\Sigma}_\beta = \frac{1}{n} \widehat{B}^{-1} \sum_{i=1}^{n} \left(\widehat{G}_i \widehat{S}_i \widehat{S}_i^\top \widehat{G}_i^\top\right) \widehat{B}^{-\top}. \tag{2.28}$$

where $\widehat{A}$ denotes the estimated A obtained by replacing β and α with their respective estimates.

Note that in some cases $\sqrt{n}$-consistent estimates $\widehat{\alpha}$ and their limits α may give rise to some or all entries of the working correlation matrix that exceed one (Crowder 1995). However, such examples are rare in practice, but even when this occurs, the GEE estimate $\widehat{\beta}$ is still consistent and asymptotically normal, although efficiency may be an issue.

2.4 Missing Values

Missing data occurs in most longitudinal studies, including well-designed and carefully executed clinical trials. In longitudinal studies, subjects may simply quit the study or they may not show up at follow-up visits because of problems with transportation, weather condition, relocation, and so on. We characterize the impact of missing data on model estimates through modeling assumptions. Such assumptions allow one to ignore the multitude of reasons for missing data and focus instead on addressing their impact on inference.

The *missing completely at random* (MCAR) assumption is used to define a class of missing data that does not affect model estimates when completely ignored. For example, missing data resulting from relocation and lack of transportation typically follows this model. In clinical trials, missing data may also be the results of patients' deteriorated or improved health conditions due to treatment and treatment-related complications. For example, if some patients in a study feel that the interventions received have resulted in no change or even deteriorated health conditions, and any further treatment will only worsen their overall physical and mental health, they may simply quit the study. On the other hand, some others may feel that they have completely responded to the treatment and do not see any additional benefit in continuing the treatment, they may also choose to stop participating in the study. In both scenarios, missing data does not follow the MCAR mechanism, since the missingness of these patients' data is related to treatment effects.

The *missing at random* (MAR) assumption, which posits that the occurrence of a missing response at an assessment time depends on the observed data prior to the assessment point, attempts to model such a treatment-dependent missing data mechanism. In the two missing data scenarios above, MAR is a reasonable model, since the missingness is a function of past treatment responses.

Missing data satisfying either the MCAR or MAR model is known as *ignorable missing data*. Thus, the *Nonignorable nonresponse* (NINR) or *missing not at random* (MNAR) mechanism encompasses the remaining class of missing data whose occurrence depends on unobserved data, such as current and/or future responses as in a longitudinal study. This category of missing data is generally quite difficult to model without additional information (from other sources), because of lack of information from the study data.

Note that the term "ignorable missing" may be a misnomer. For parametric models, we can indeed ignore such missing data since MLEs are consistent. However, the validity of inference for parametric models depends on the model assumptions; if there is some serious violation of the assumed distributional models, MLEs will be biased (Lu et al. 2009). On the other hand, distribution-free models such as the distribution-free GLM discussed in Sect. 2.3 provide valid inference without any distribution assumption. But for such robust models, "ignorable missing" data may not be ignored. For example, GEE estimates discussed in Sect. 2.3 are generally biased when missing data follows MAR (Lu et al. 2009). We discuss a new class of estimating equations to provide valid inference under MAR in Sect. 2.4.2.

We focus on MCAR and MAR, which apply to most studies in biomedical and psychosocial research, but will mention some popular approaches for NINR in Discussion. In addition, we only consider missingness in the response, as missingness in the independent variables is much more complex to model.

2.4.1 Inference for Parametric Models

Let y_{it} be the response and $\mathbf{x}_{it}$ a vector of independent variables of interest for the ith subject at time t from a longitudinal study design with n subjects and m assessments. We define a missing, or rather observed, data indicator as follows:

$$r_{it} = \begin{cases} 1 \text{ if } y_{it} \text{ is observed} \\ 0 \text{ if } y_{it} \text{ is missing} \end{cases}, \quad \mathbf{r}_i = (r_{i1}, \ldots, r_{im})^\top. \tag{2.29}$$

We assume no missing data at baseline $t = 1$ so that $r_{i1} = 0$ for all $1 \leq i \leq n$.

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^\top$ and $\mathbf{x}_i = (\mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{im}^\top)^\top$. Let $\mathbf{y}_i^{\mathrm{o}}$ and $\mathbf{y}_i^{\mathrm{m}}$ denote the observed and unobserved responses, respectively. Thus, $\mathbf{y}_i^{\mathrm{o}}$ and $\mathbf{y}_i^{\mathrm{m}}$ form a partition of $\mathbf{y}_i$. Under likelihood-based parametric inference, the joint density, $f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{x}_i, \theta)$, can be factored into:

$$f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{x}_i) = f(\mathbf{y}_i \mid \mathbf{x}_i) f(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{x}_i). \tag{2.30}$$

Under MAR, the distribution of $\mathbf{r}_i$ only depends on the observed responses, $\mathbf{y}_i^{\mathrm{o}}$, yielding

$$f(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{x}_i) = f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{y}_i^{\mathrm{m}}, \mathbf{x}_i\right) = f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{x}_i\right). \tag{2.31}$$

It then follows from (2.30) and (2.31) that

$$\begin{aligned} f\left(\mathbf{y}_i^{\mathrm{o}}, \mathbf{r}_i \mid \mathbf{x}_i\right) &= \int f\left(\mathbf{y}_i^{\mathrm{o}}, \mathbf{y}_i^{\mathrm{m}} \mid \mathbf{x}_i\right) f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{x}_i\right) \mathrm{d}\mathbf{y}_i^{\mathrm{m}} \\ &= f\left(\mathbf{y}_i^{\mathrm{o}} \mid \mathbf{x}_i, \theta_y\right) f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{x}_i, \theta_{y|r}\right), \end{aligned} \tag{2.32}$$

where θ_y and $\theta_{y|r}$ denote a partition of θ. If θ_y and $\theta_{y|r}$ are disjoint, it follows from (2.32) that the log-likelihood based on the joint observations $\left(\mathbf{y}_i^{\mathrm{o}}, \mathbf{r}_i\right)$ can be expressed as

$$\begin{aligned} l(\theta) &= l_1\left(\theta_y\right) + l_2\left(\theta_{y|r}\right) \\ &= \sum_{i=1}^{n} \log\left(f\left(\mathbf{y}_i^{\mathrm{o}} \mid \mathbf{x}_i, \theta_y\right)\right) + \sum_{i=1}^{n} \log\left(f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{x}_i, \theta_{y|r}\right)\right). \end{aligned} \tag{2.33}$$

Within the context of GLMM, $l_1\left(\theta_y\right)$ above is the likelihood in (2.13) based on the observed data. Thus, we can simply use $l_1\left(\theta_y\right)$ for inference about the regression relationship between $\mathbf{y}_i$ and $\mathbf{x}_i$. In other words, under MAR, missing data can be "ignored,"insofar as this regression relationship is concerned.

Note that inference based on $f\left(\mathbf{y}_i^{\mathrm{o}} \mid \mathbf{x}_i, \theta_y\right)$ may be incorrect, if θ_y and $\theta_{y|r}$ are not disjoint. In practice, it is difficult to validate this disjoint assumption. However, under MCAR, it follows from (2.31) and (2.32) that $f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{x}_i, \theta_{y|r}\right) = f\left(\mathbf{r}_i \mid \mathbf{x}_i, \theta_{y|r}\right)$, implying that θ_y and $\theta_{y|r}$ are disjoint. Thus, only under MCAR can missing data be truly ignored when making inference about the relationship between $\mathbf{y}_i$ and $\mathbf{x}_i$.

Note also that if there is some serious violation of the assumed distributional models, estimates obtained by maximizing $l_1(\theta_y)$ are generally not consistent. Some robust estimates have been adopted to improve the validity of inference (Goldstein 1995; Rasbash et al. 2009; Raudenbush and Bryk 2002). These estimates are essentially equivalent to the GEE estimates. However, as GEE estimates are consistent only under MCAR, rather the more general MAR, these robust adjustments do not provide valid inference either (Lu et al. 2009). To obtain consistent estimates under MAR and violations of distribution assumptions, we must use a new class of estimating equations.

2.4.2 *Inference for Distribution-Free Model*

Define the *probability weight* π_{it} and *inverse probability weight* Δ_{it} function as follows:

$$\pi_{it} = \Pr(r_{it} = 1 \mid \mathbf{x}_i, \mathbf{y}_i), \quad \Delta_{it} = \frac{r_{it}}{\pi_{it}}, \quad \Delta_i = \text{diag}_t(\Delta_{it}). \tag{2.34}$$

Given Δ_i, we can use the following weighted generalized estimating equations (WGEE) to estimate β:

$$\mathbf{w}_n(\beta) = \sum_{i=1}^{n} G_i \Delta_i S_i = \sum_{i=1}^{n} D_i V_i^{-1} \Delta_i (\mathbf{y}_i - \mathbf{h}_i) = \mathbf{0}, \tag{2.35}$$

where D_i, V_i, and S_i are defined the same way as those in (2.22). If the probability weight function π_{it} is known, the WGEE above can be readily solved for β using the Newton–Raphson algorithm discussed in Sect. 2.3.3. In some multi-stage studies where only a fraction of subjects move to the next stage, π_{it} is known and the WGEE can be used to provide inference for such studies. However, in most longitudinal studies, π_{it} is unknown and must be estimated. In general, it is quite difficult to model and estimate π_{it}, as it may depend on unobserved current, $\mathbf{x}_{it}$ and y_{it}, or even future, $\mathbf{x}_{is}$ and y_{is} $(t \le s \le m)$, observations. But, if missing data follows either MCAR or MAR, π_{it} can be modeled as discussed below.

Under MCAR, $\mathbf{r}_i$ is independent of $\mathbf{x}_i$ and $\mathbf{y}_i$. It follows from (2.34) that π_{it} is a constant and is readily estimated by the sample moment:

$$\widehat{\pi}_t = \frac{1}{n} \sum_{i=1}^{n} r_{it}, \quad 2 \le t \le m, \quad 1 \le i \le n.$$

Under MAR, π_{it} is not a constant, and the above is not a valid estimate. However, as π_{it} is a function of observed $\mathbf{x}_i$ and $\mathbf{y}_i$, we should be able to model π_{it}. For example,

if $m = 2$ as in a pre-post study, π_{i2} is a function of observed $\mathbf{x}_{i1}$ and y_{i1}, $\pi_{i2} = \Pr(r_{i2} = 1 \mid y_{i1}, \mathbf{x}_{i1})$, which may be modeled using logistic regression.

However, as m increases, it becomes more difficult to model π_{it}; unlike the pre-post design discussed earlier, there could be many different patterns (2^{m-1} total) in the observed data. Thus, we generally assume the monotone missing data pattern (MMDP) to facilitate modeling. Under MMDP, a subject with missing y_{it} and $\mathbf{x}_{it}$ implies that all subsequent components, y_{is} and $\mathbf{x}_{is}$ $(t \le s \le m)$, are also missing. With the help of MMDP, we can express π_{it} as

$$\pi_{it} = \Pr(r_{it} = 1 \mid \mathbf{y}_{it-}, \mathbf{x}_{it-}), \quad 1 \le i \le n, \quad 2 \le t \le T, \tag{2.36}$$

where $\mathbf{y}_{it-} = \left(y_{i1}, \ldots, y_{i(t-1)}\right)^\top$ and $\mathbf{x}_{it-} = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(t-1)}\right)^\top$. Since

$$\pi_{it} = \prod_{s=2}^{t} p_{is}, \quad p_{is} = \Pr(r_{is} = 1 \mid \mathbf{y}_{is-}, \mathbf{x}_{is-}, r_{i(s-1)} = 1). \tag{2.37}$$

we can estimate π_{it} through estimates of one-step transition probabilities p_{is} $(2 \le s \le t)$.

We can readily model p_{it} by logistic regression:

$$\text{logit}\left(p_{it}\left(\gamma_t\right)\right) = \gamma_{0t} + \gamma_{xt}^\top \mathbf{x}_{it-} + \gamma_{yt}^\top \mathbf{y}_{it-}, \quad 2 \le t \le m, \tag{2.38}$$

where $\gamma_t = \left(\gamma_{0t}, \gamma_{xt}^\top, \gamma_{yt}^\top\right)^\top$ denotes the model parameters. We can estimate each γ_t using either maximum likelihood or estimating equations. For example, if using maximum likelihood, the score equations for each γ_t are given by

$$\mathbf{Q}_{it} = \frac{\partial}{\partial \gamma_t} \left\{ r_{i(t-1)} \left[r_{it} \log\left(p_{it}\right) + \left(1 - r_{it}\right) \log\left(1 - p_{it}\right) \right] \right\},$$
$$2 \le t \le m, \quad i = 1, 2, \ldots n.$$

By combining the score equations across all γ_t's, we obtain a set of estimating equations for joint γ_t as

$$\mathbf{Q}_n(\gamma) = \sum_{i=1}^{n} \left(\mathbf{Q}_{i2}^\top, \ldots, \mathbf{Q}_{im}^\top\right)^\top = \mathbf{0},$$

where $\gamma = \left(\gamma_2^\top, \ldots, \gamma_m^\top\right)^\top$.

As in the case of GEE, the WGEE estimate $\widehat{\beta}$ is asymptotically normal if $\widehat{\alpha}$ is $\sqrt{n}$-consistent. However, the asymptotic variance is more complex, as it must reflect the additional variability in the estimated $\widehat{\gamma}$:

$$\Sigma_\beta = B^{-1} E\left(G_i \Delta_i S_i S_i^\top \Delta_i G_i^\top\right) B^{-\top} + B^{-1} \Phi B^{-\top},$$
$$B = E\left(D_i V_i^{-1} \Delta_i D_i^\top\right), \quad C = E\left[\frac{\partial}{\partial \gamma}\left(D_i V_i^{-1} \Delta_i S_i\right)\right]^\top, \quad H = E\left(\frac{\partial}{\partial \gamma} \mathbf{Q}_{ni}\right)^\top,$$
$$\Phi = -CH^{-\top} C^\top - E\left(D_i V_i^{-1} \Delta_i S_i \mathbf{Q}_{ni}^\top H^{-\top} C^\top\right) - \left[E\left(D_i V_i^{-1} \Delta_i S_i \mathbf{Q}_{ni}^\top H^{-\top} C^\top\right)\right]^\top. \tag{2.39}$$

The first term of Σ_β in the above is identical to the asymptotic variance of the GEE estimate in (2.27), while the second term accounts for the additional variability in $\widehat{\gamma}$. A consistent estimate of Σ_β is obtained by substituting consistent estimates of the respective parameters.

Note that if π_{it} is known, the asymptotic variance $\widehat{\beta}$ is given by the first term of Σ_β in (2.39). In some multi-stage studies, selection of subjects for each stage may follow models such as those defined by (2.37) and (2.38), but with known γ. We may use π_{it} based on the designed values of γ when estimating β, in which case the asymptotic variance of the WGEE estimate is again given by the first term of Σ_β. However, one may also estimate γ as in the above and use the estimated version in (2.35), in which case the asymptotic variance of the WGEE estimate should include the second term. The latter approach may be preferred, since it may yield more efficient estimates (Tsiatis 2006).

2.5 Software for Fitting Longitudinal Models

The two popular parametric and distribution-free modeling approaches discussed above have been implemented in many general-purpose commercial statistical packages such as R (R Development Core Team 2011), SAS (SAS Institute 2011), SPSS (SPSS Inc. 2009), and Stata (StataCorp 2011). They are also available in some specialized packages for longitudinal data analysis such as HLM (Raudenbush et al. 2007). For example, in SAS, we can use MIXED, GLMMIX, and NLMIXED for fitting GLMM, but only the latter two can be used for fitting noncontinuous responses such as binary. As inference for distribution-free models is typically based on the GEE or WGEE, such models are commonly referred to simply as GEE (WGEE). The SAS GENMOD procedure fits the distribution-free GEE models. Currently, WGEE is not available as a SAS-supported procedure, and some user-written SAS macros may be used to facilitate such inference. In R, the primary functions (packages) for fitting GLMM are lme4, ZELIG, and glmmML, while the functions gee and geepack are used to obtain GEE estimates.

Note that as the log-likelihood for GLMM involves high-dimensional integration, due to the needs to integrate out the latent random effects $\mathbf{b}_i$, numerical approximations must be used to approximate the likelihood function, except for linear regression Zhang et al. (2011a). As the approximation is quite complex and the accuracy of estimate depends critically on the precision of such approximations,

performance varies considerably across these available procedures. For example, when fitting binary responses, none of the currently available functions and packages in R at the moment of this writing time yields correct inference, while for the two SAS procedures, only NLMIXED provides good results (Zhang et al. 2011b). Thus, before using a package for fitting GLMM in a real study, it is important to check for information about its performance to ensure valid inference.

2.6 A Real Data Example

The Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) is widely used as a significant health outcome indicator (Lubetkin et al. 2003; Wan et al. 2011). It is a multipurpose and self-reported health status measure, yielding an eight-subscale profile of scores (domains) as well as the Physical and Mental Health summary measures. The higher measurement precision, reduced floor and ceiling effects, and superior responsiveness make the SF-36 a popular measure of quality of life in research and clinical studies. As a result, the SF-36 has been translated into many foreign languages, including the simplified Chinese (Mandarin) version of SF-36 (CSF-36), and used in more than 40 different countries (Wan et al. 2011).

A recent study was conducted to evaluate the performance of CSF-36 when used to assess health-related quality of life (HRQOL) for patients with hypertension, coronary heart diseases, chronic gastritis, or peptic ulcer in mainland China. The study population consisted of inpatients with one of these four types of chronic diseases seen at the first affiliated hospital of Kunming Medical University, Kunming province, China. There were no exclusions based on age, clinical stage of disease, and treatment course, but the subjects were required to be able to read, understand, and complete the questionnaires, which led to exclusions based on illiteracy and advanced disease status.

Among the 534 patients who completed the initial survey, 40–50 % of these were randomly selected to take the questionnaire a second time, 1–2 days after hospitalization, to assess the test–retest reliability. Another sub-sample of patients (80–90 %) was selected to complete the questionnaire a third time at discharge (after about 2 weeks of treatment) to evaluate clinical change and responsiveness of the instrument to detecting such changes.

For illustration purposes, we focused on one component of the study that examined the criterion-related validity. As there was no agreed-upon gold standard, the Quality of Life Instruments for Chronic Diseases-General Module (QLICD-GM), developed by Wan et al. (2005, 2007), was used to provide the criterion for validating each of the eight subscales: Physical Function (PF), Role-Physical (RP), Bodily Pain (BP), General Health (GH),Vitality (VT), Social Function (SF), Role-Emotional (RE), and Mental-Health (MH). These subscales were used as predictors in three separate regression models with the QLICD-GM's three domain scores, Physical Function, Psychological Function, and Social Function, serving as the respective responses. To utilize all available data, we modeled each regression

Table 2.1 Estimates, standard errors (S.E), and p-values from linear mixed-effects model for the Quality of Life Study Data

Estimates for eight domains of CSF-36 from linear mixed effects model with response based on each domain score of QLICD-GM								
	PF	RP	BP	GH	VT	SF	RE	MH
Physical health from QLICD-GM								
Estimate	0.407	−0.028	0.025	0.153	0.116	−0.025	−0.005	0.063
S.E.	0.019	0.014	0.022	0.031	0.032	0.023	0.014	0.032
P-value	<0.01	0.055	0.263	<0.01	<0.01	0.275	0.711	0.047
Psychological function from QLICD-GM								
Estimate	0.143	−0.043	0.010	0.167	0.006	0.105	0.047	0.296
S.E.	0.019	0.014	0.022	0.031	0.032	0.023	0.014	0.032
P-value	<0.01	<0.01	0.650	<0.01	0.848	<0.01	<0.01	<0.01
Social function from QLICD-GM								
Estimate	0.103	−0.005	−0.001	0.093	0.063	0.028	−0.003	0.201
S.E.	0.016	0.011	0.017	0.025	0.025	0.018	0.010	0.025
P-value	<0.01	0.617	0.959	<0.01	0.012	0.123	0.732	<0.01

using the longitudinal methods discussed above based on the data from all three assessment times.

For each of the three domains of CSF-36, let y_{it} denote the domain score, and x_{ikt} the kth subscale of QLICD-GM from the ith subject at time t, with k denoting the subscales, PF, RP, BP, GH, VT, SF, RE, and MH, in order for $1 \leq k \leq 8$. Let $x_{i0t} = 1$ and $\mathbf{x}_{it} = (x_{i0t}, x_{i1t}, \ldots, x_{i8t})^\top$.

2.6.1 LMM

We first fit the LMM (LMM) in (2.4) with $\mathbf{z}_{it} = \mathbf{x}_{it}$. Shown in Table 2.1 are the estimates of parameters from the LMM, along with the standard errors (S.E.) and p-values. Overall, the estimates were larger (in absolute value) for the predictors that correspond to the same and similar domains than to different and non-similar domains of the CSF-36. The estimate for PF was the largest for the LMM with Physical Health as the response. In addition, GH, VT, and MH also significantly predicted this outcome. For the LMM with Psychological Function as the response, MH had the largest beta weight, followed by PF, GH, SF, RE, and RP, which were all related to this outcome. All these domains significantly predicted Psychological Function. The MH domain also had the largest beta weight for the LMM with Social Function as the response. Additionally, PF, GH, and VT were also significantly associated with this outcome.

Table 2.2 Estimates, standard errors (S.E), and p-values from distribution-free model with inference based on GEE WGEE for the Quality of Life Study Data

Estimates for eight domains of CSF-36 from distribution-free model with response based on each domain score of QLICD-GM								
	PF	RP	BP	GH	VT	SF	RE	MH
Physical health from QLICD-GM								
Estimate	0.400	−0.017	0.016	0.137	0.111	0.001	−0.001	0.032
S.E.	0.023	0.017	0.025	0.031	0.036	0.025	0.018	0.035
P-value	<0.01	0.293	0.519	<0.01	<0.01	0.973	0.958	0.370
Psychological function from QLICD-GM								
Estimate	0.112	−0.030	−0.012	0.128	0.033	0.122	0.059	0.322
S.E.	0.023	0.014	0.023	0.033	0.035	0.029	0.015	0.038
P-value	<0.01	0.041	0.598	<0.01	0.357	<0.01	<0.01	<0.01
Social function from QLICD-GM								
Estimate	0.134	0.011	−0.002	0.094	0.032	0.054	−0.008	0.197
S.E.	0.019	0.015	0.024	0.032	0.038	0.022	0.014	0.037
P-value	<0.01	0.441	0.932	<0.01	0.390	0.017	0.551	<0.01

2.6.2 GEE

The estimates from LMM will be valid, if there is no serious violation against the normal–normal assumption for the random effects and error terms. Although various types of residuals may be used to help describe the distribution of the response, it is difficult to formally test the normal–normal assumption. However, it is possible to conduct a test to validate the marginal normality, i.e., the sum of random effects and errors follows a multivariate normal (Tan et al. 2005; Feng et al. 2009). If this marginal normality holds true, the LMM above still yields valid inference, provided that robust variance estimates such as the sandwich variance estimates are used. In fact, inference in this case is equivalent to that based on GEE. Thus, we also applied the distribution-free model in (2.19), with the same explanatory variables $\mathbf{x}_{it}$ using GEE for inference.

Shown in Table 2.3 are the GEE estimates for this distribution-free model. By comparing these with the results for LMM in Table 2.2, it is seen that the order of magnitude of estimate is the same for the first three largest beta weights. However, there were some changes for the order of the remaining estimates. For example, as in LMM, MH, PF, and GH were also the first three largest beta weights for Social Function for the distribution-free model. However, VT, the 4th largest beta weight under LMM, drops down to the 5th place for the distribution-free model.

Table 2.3 Estimates, standard errors (S.E), and p-values from logistic regression for modeling missingness under MAR for each of the three domain scores of QLICD-GM for the Quality of Life Study Data

Estimates of parameters of logistic regression parameters for modeling missingness for each domain score of QLICD-GM at time 2/3			
Predictors	Estimates	S.E.	P-value
Prior physical health	−0.0005/0.0002	0.0091/0.0073	0.99/0.92
Prior psychological function	0.006/0.005	0.009/0.006	0.41/0.34
Prior social function	0.001/0.002	0.012/0.022	0.88/0.75

2.6.3 Testing MCAR

As discussed in Sect. 2.4, estimates from GEE may not be consistent under MAR; it is only valid under the more stringent MCAR. In principle, the missing data should follow MCAR in this study, as it is the result of random selection of a subgroup at each of the two follow-up times. But, we can also formally test this assumption to see if the random selection was successful. To this end, we modeled the MAR following the discussion in Sect. 2.4.2. Since we did not consider any covariate in the analysis, the logistic regression in (2.38) reduced to

$$\text{logit}\left(p_{it}\left(\gamma_t\right)\right) = \gamma_{0t} + \gamma_{1t} y_{it-}, \quad t = 2, 3, \tag{2.40}$$

where $y_{i2-} = y_{i1}$ and $y_{i3-1} = y_{i2}$.

We applied the logistic regression in (2.40) to each of the three domain scores of QLICD-GM. Shown in Table 2.3 are the estimates of parameters from the logistic model, their standard errors, and corresponding p-values. As neither the response at time 1 nor at 2 significantly predicted the missingness at time 2 and 3, the results confirmed the random selection process, i.e., the missing data follows the MCAR mechanism. Since the missingness in our study followed MCAR, GEE provides consistent estimates, regardless of the data distribution, and inference based on this distribution-free would be preferred over its parametric counterpart LMM.

2.7 Discussion

We discussed longitudinal designs and associated advantages of such models in clinical trial and observational studies. Unlike cross-sectional studies that only indicate associations between outcomes, longitudinal designs provide temporal changes of such associations, thereby offering a framework for investigating the causal mechanisms of such relationships. However, adding the time dimension in longitudinal data creates several methodological issues, which in particular preclude applications of standard statistical models and procedures for cross-sectional data

analysis. As a result, longitudinal study data must be analyzed by methods that address the unique data-analytic problems associated with such data.

The two most popular approaches for longitudinal data are the GLMM and the distribution-free GLMs. As inference for the latter is often based on the GEE or WGEE, this class of models is typically referred to simply as GEE (WGEE). GLMM explicitly models the two sources of between- and within-subject variability using random effects, while GEE (WGEE) tackles the correlated responses over time directly by basing inference based on the marginal distribution of the subject's responses. Although it carries no effect on linear models for continuous responses, this conceptual difference does have significant implications for interpretations of estimates when modeling noncontinuous responses such as binary and count data. In the validation study of CSF-36, the estimates are quite similar, as they estimate the same underlying quantities. However, when modeling the binary or count response using GLMM and GEE (WGEE), the parameters for the same explanatory variable in GLMM (fixed-effect) and GEE are generally on a different scale (Diggle et al. 2002; Zeger et al. 1988; Zhang et al. 2011a,b). For example, when using the logit link for model binary responses, the parameter may not have the familiar log odds ratio interpretation for the GLMM (Zhang et al. 2011a). Thus, we must be mindful when applying these approaches to real study data, as such differences are likely the source for discrepant findings.

We focused on MCAR and MAR when addressing missing data, as these are the most likely missing data mechanisms for real studies. Although MAR takes into account the dependence of missing visit on the response, such a dependent relationship relates only the *observed* response to the missing visit. Thus, if the missingness relates to the subject's future response, i.e., following NINR, both GLMM and WGEE will yield biased inference. In cancer studies, the subject may drop out of the study either because of death or intolerance of the medication. Thus, the two causes of missingness may be associated with different responses of disease progression, invalidating the MAR mechanism. A number of approaches have been proposed to address this issue. For example, in *pattern mixture* models, different regression relationships are posited based on the different missing data patterns (Birmingham and Fitzmaurice 2002). The most popular approach, however, is to model the relationship between missingness and future responses through the random effects, $\mathbf{b}_i$ in (2.10). For example, in the cancer study example, the potential differential effects of the two causes of missingness on future responses are accommodated by adding a survival model to relate $\mathbf{b}_i$ as a predictor to the outcome of time of death (Tsiatis and Davidian 2004). If $\mathbf{b}_i$ significantly predicts the time to death, then $f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{y}_i^{\mathrm{m}}, \mathbf{x}_i\right) \neq f\left(\mathbf{r}_i \mid \mathbf{y}_i^{\mathrm{o}}, \mathbf{x}_i\right)$, thereby invalidating the MAR condition in (2.31) (Degruttola and Tu 1994; Hogan and Laird 1997). As the two log-likelihood functions, $l_1\left(\theta_y\right)$ and $l_2\left(\theta_{y|r}\right)$ in (2.33), share a set of parameters for the variance of $\mathbf{b}_i$, this joint modeling approach is also known as the *shared parameter* models (Follmann and Wu 1995).

References

Birmingham J, Fitzmaurice G (2002) A pattern-mixture model for longitudinal binary responses with nonignorable nonresponse. Biometrics 58(4):989–996

Breslow N, Clayton D (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc, 88:9–25

Bryk A, Raudenbush S, Congdon R (1996) HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs. Scientific Software International, Lincolnwood, IL

Crowder M (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika 82(2):407–410

Davidian M, Gallant A (1993) The nonlinear mixed effects model with a smooth random effects density. Biometrika 80(3):475–488

Degruttola V, Tu XM (1994) Modeling progression of cd4-lymphocyte count and its relationship to survival-time. Biometrics 50(4):1003–1014

Demidenko E (2004) Mixed models: theory and applications, vol 518. LibreDigital, Austin, TX

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc B 39(1):1–38

Diggle PJ, Heagerty PJ, Liang K-Y, Zeger SL (2002) Analysis of longitudinal data, 2nd edn. Oxford statistical science series, vol 25. Oxford University Press, Oxford

Feng C, Su H, Wang H, Tang W, Yu Q, Tu XM (2009) Testing normality for longitudinal studies with missing data. Far East J Theor Stat 28(2):133–155

Follmann D, Wu M (1995) An approximate generalized linear model with random effects for informative missing data. Biometrics, 51:151–168

Goldstein H (1987) Multilevel models in education and social research. Charles Griffin & Co, London, England; Oxford University Press, New York

Goldstein H (1995) Hierarchical data modeling in the social sciences. J Educ Behav Stat 20(2): 201–204

Hogan J, Laird N (1997) Mixture models for the joint distribution of repeated measures and event times. Stat Med 16(3):239–257

Kowalski J, Tu XM (2008) Modern applied U-statistics. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, NJ

Laird N, Ware J (1982) Random-effects models for longitudinal data. Biometrics, 38:963–974

Lindstrom M, Bates D (1988) Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. J Am Stat Assoc, 83:1014–1022

Lu NJ, Tang W, He H, Yu Q, Crits-Christoph P, Zhang H, Tu X (2009) On the impact of parametric assumptions and robust alternatives for longitudinal data analysis. Biom J 51(4):627–643

Lubetkin E, Jia H, Gold M (2003) Use of the SF-36 in low-income Chinese American primary care patients. Medical Care 41(4):447–457

Meng XL, Van Dyk D (1997) The em algorithm—an old folk-song sung to a fast new tune. J Roy Stat Soc B Stat Meth 59(3):511–567

Pinheiro J, Bates D (2000) Mixed-effects models in S and S-PLUS. Springer, Berlin

R Development Core Team (2011) R: A language and environment for statistical computing. R doundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0

Rasbash J, Steele F, Browne W, Goldstein H (2009) A User's Guide to MLwiN, v2.10. Centre for Multilevel Modelling, University of Bristol

Raudenbush S, Bryk A (2002) Hierarchical linear models: Applications and data analysis methods, vol 1. Sage Publications, Thousand Oaks, CA

Raudenbush S, Bryk A, Congdon R (2007) Hlm for windows (version 6.04)[computer software]. Scientific Software International, Lincolnwood, IL

SAS Institute (2011) Base SAS 9.3 procedures guide. SAS Institute, Cary, NC

Schabenberger O, Gregoire T (1996) Population-averaged and subjectspecific approaches for clustered categorical data. J Stat Comput Simulat 54(1-3):231–253

Seber G (1984) Multivariate observations, vol 41. Wiley Online Library, Hoboken, NJ

Shults J, Sun WG, Tu X, Kim H, Amsterdam J, Hilbe A, Ten-Have T (2009) A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. Stat Med 28(18):2338–2355

SPSS Inc. (2009) SPSS base 17.0 for Windows User's Guide. SPSS Inc., Chicago IL

StataCorp (2011). Stata statistical software: release 12. StataCorp LP: StataCorp, College Station, TX

Strenio J, Weisberg H, Bryk A (1983) Empirical bayes estimation of individual growth-curve parameters and their relationship to covariates. Biometrics, 39:71–86

Tan M, Fang H-B, Tian GL, Wei G (2005) Testing multivariate normality in incomplete data of small sample size. J Multivariate Anal 93(1):164–179

Tang W, He H, Tu X (2012) Applied categorical and count data analysis. Chapman & Hall/CRC, London/Boca Raton

Tsiatis A, Davidian M (2004) Joint modeling of longitudinal and time-to-event data: an overview. Statistica Sinica 14(3):809–834

Tsiatis AA (2006) Semiparametric theory and missing data. Springer series in statistics. Springer, New York

Wan C, Gao L, Li X et al. (2005) Development of the general module for the system of quality of life instruments for patients with chronic disease: items selection and structure of the general module[j]. Chin Ment Health J 11:444–447

Wan C, Tu X, Messing S, Li X, Yang Z, Zhao X, Gao L, Yang Y, Pan J, Zhou Z (2011) Development and validation of the general module of the system of quality of life instruments for chronic diseases and its comparison with sf-36. J Pain Symptom Manag 42:93–104

Wan C, Yang Z, Yang Y (2007) Development of the general module of the system of quality of life instruments for patients with chronic disease: Evaluation of the general module. Chinese J Behavior Medical Sci 16:559–561

Wolfinger R, O'connell M (1993) Generalized linear mixed models a pseudo-likelihood approach. J Stat Comput Simulat 48(3–4):233–243

Zeger S, Liang K, Albert P (1988) Models for longitudinal data: a generalized estimating equation approach. Biometrics, 44:1049–1060

Zhang H, Lu N, C, Feng S, Thurston Y, Xia L, Zhu, and Tu X (2011a). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. Stat Med 30, 2562–2572

Zhang H, Xia Y, R, Chen D, Gunzler W, Tang, and Tu X (2011b). Modeling longitudinal binomial responses: implications from two dueling paradigms. J Appl Stat 38:2373–2390

http://www.springer.com/978-1-4614-4321-6

Modern Clinical Trial Analysis
Tang, W.; Tu, X. (Eds.)
2013, X, 254 p., Hardcover
ISBN: 978-1-4614-4321-6