41. Latent Variable Models for Longitudinal Data with Flexible Measurement Schedule

This chapter provides a survey of the development of latent variable models that are suitable for analyzing unbalanced longitudinal data. This chapter begins with an introduction, in which the marginal modeling approach (without the use of latent variable) for correlated responses such as repeatedly measured longitudinal data is described. The concepts of random effects and latent variables are introduced at the beginning of Sect. 41.1. Section 41.1.1 describes the linear mixed models of Laird and Ware for continuous longitudinal response; Sect. 41.1.2 discusses generalized linear mixed models (with latent variables) for categorical response; and Sect. 41.1.3 covers models with multilevel latent variables. Section 41.2.1 presents an extended linear mixed model of Laird and Ware for multidimensional longitudinal responses of different types. Section 41.2.2 covers measurement error models for multiple longitudinal responses. Section 41.3 describes linear mixed models with latent class variables—the latent class mixed model that can be useful for either a single or multiple longitudinal responses. Section 41.4 studies the relationships between multiple longitudinal responses through structural equation models. Section 41.5 unifies all the above varieties of latent variable models under a single multilevel latent variable model formulation.

Longitudinal data consists of variables that are measured repeatedly over time. Longitudinal data can be collected either prospectively or retrospectively. The defining feature of longitudinal data is that the set of observations on one subject are likely to be correlated, and this within-subject correlation must be taken into account in order to make valid scientific inferences from the data. A frequently encountered problem in longitudinal studies is data that are missing due to missed visits or dropouts. As a result subjects often do not have a common set of visit times or they visit at nonscheduled times, thus longitudinal data may be highly unbalanced.

Except in the Introduction, the remaining sections are devoted to the study of the models along the lines of the Laird and Ware-style mixed model [41.1] and models with latent variables since they naturally handle unbalanced longitudinal data and of course these models are also useful for regularly spaced, balanced, repeatedly measured responses. Models suitable only for balanced longitudinal data as well as missing data models are not discussed in this chapter. All the models discussed in this chapter have been proved successful in practice. However, the models covered in this chapter only reflect the choice of illustration by the author and are by no means inclusive of all the variants and extensions of latent variable models.

Before moving onto the next section, let us first look at the marginal models that do not involve latent variables. Let Y_{ij} denote the longitudinal response for subject i $(i = 1, \ldots, N)$ at the j-th time point

$(j = 1, \ldots, n_i)$. For example, in the study of the evolution of the CD4+ lymphocyte in human immunodeficiency virus (HIV) positive subjects, longitudinal CD4+ cell counts can be modeled with the following *general linear model* for continuous response as:

$$Y_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{ij} \,, \tag{41.1}$$

where $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$ is a vector of fixed p covariates at j-th time point that includes the intercept of one, the linear term of time in months, the indicator of AZT (an anti-retrovirus drug) usage, the Karnovsky score, anemia, etc., $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^{\mathrm{T}}$ are the coefficients for the intercept and the partial slopes, and ϵ_{ij} is an error term that has a zero mean and variance of σ^2. The correlation between ϵ_{ij} and $\epsilon_{ij'}$ is $\rho_{jj'}$ for $j \neq j'$. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}$ and $\mathbf{X}_i = (\boldsymbol{x}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_{in_i}^{\mathrm{T}})^{\mathrm{T}}$, then the above model can be written in matrix notation as

$$\boldsymbol{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \,.$$

So $\boldsymbol{Y}_i$ has mean vector $\mathbf{X}_i\boldsymbol{\beta}$ and variance matrix $\sigma^2\mathbf{R}_i$, where $\mathbf{R}_i$ is the correlation matrix for $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{i,n_i})^{\mathrm{T}}$.

For discrete responses, marginal models that extend the *generalized* linear models GLMs can be applied [41.2]. Marginal models model a *link* function of the population-average response, $E(Y_{ij})$, as a function of a common set of explanatory variables $\boldsymbol{x}$. The mean of the longitudinal response is modeled separately from the within-subject correlation that is usually assumed to be a function of the modeled marginal means and possibly additional parameters $\boldsymbol{\nu}$. The marginal model is specified as:

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} \,, \tag{41.2}$$

where g is the monotone *link* function, $\mu_{ij} = E(Y_{ij})$. For example, g is an identity link for continuous Gaussian response, g can be a *logit* link for binary response. An attractive feature of the marginal model is that within-subject correlation does not have to be modeled explicitly, rather a class of generalized estimating equations (GEE) that gives consistent estimates of the $\boldsymbol{\beta}$ and their variance is used under some assumed working correlation matrices for within-subject dependence without specifying a multivariate distribution for $\boldsymbol{Y}_i$ [41.3].

The regression coefficients $\boldsymbol{\beta}$ in marginal models have population-average interpretation but any heterogeneity beyond the recorded covariates cannot be accounted for in marginal models. In models with latent variables, that are studied in following sections, heterogeneity among subjects in a subset of the regression coefficients, e.g., the intercept, are taken into account via subject-specific regression coefficients and/or covariates. In latent variable models, the covariate effects and within-subject association are modeled simultaneously. The concept of latent variables is a convenient way to represent statistical variation in terms of measurement error, random coefficients and variance components. Modeling the heterogeneity of a subset of regression coefficients not only reduces the extent of unexplained variation beyond the recorded explanatory variables but may be of interest in its own right. Estimates of the parameters in the latent variable models studied in this chapter can be obtained via the likelihood method with either the EM (expectation-maximization) algorithm or (adaptive) Gaussian quadrature.

41.1 Hierarchical Latent Variable Models for Longitudinal Data

One may consider longitudinal data as having a two-level structure, with repeated measurements (level 1) of a response variable being nested within subjects (level 2). Traditional fixed-effect analytical methods (e.g., analysis of variance) are limited in their treatment of the technical difficulties presented by nested designs, and in the questions they are able to address. Models that include random regression coefficients are more suited to the hierarchical data structure generally found in longitudinal data.

Latent variables are unobservable individual regression coefficients, predictors/covariates or response variables in regression models. Latent variables here are thus divided into the following three types, and sometimes a latent variable qualifies for more than one of the three types. The first type is called *random effects*, which model heterogeneity among subjects in a subset of the regression coefficients that vary from one subject to the next. A defining feature of random effects is that the individual regression coefficients are assumed to be a random sample from a common distribution so that a few parameters for the distribution characterize the behavior of the entire random coefficients. The second type is called *latent covariates*; these unobservable latent covariates have their own fixed regression coefficients that are called *factor loadings*. The third type is

called *latent responses*, which are further modeled on other fixed covariates and/or latent covariates.

Softwares for fitting latent variable models are abundant, although they may only handle a limited number of the models discussed in this chapter. Many researcher-written computer codes and softwares are also available for fitting complicated latent variable models. One omnipass software for fitting the latent variable models studied in this chapter is the STATA module generalized linear latent and mixed models (GLLAMM) developed over years by *Rabe-Hesketh* et al. [41.4]. For some models, it may take quite some computer and real time to fit. The computing time usually depends on the size of data, the number of structural levels in the data, and more critically on the number of latent variables involved. Software development will not be further discussed in this chapter.

41.1.1 Linear Mixed Model with a Single-Level Latent Variable

Since repeated measurements are obtained from each individual at different times, there may be considerable variation among individuals in the number and timing of observations. The resulting unbalanced data are typically not amenable to analysis using a general multivariate model such as (41.1) above, mainly due to the difficulty in specifying the covariance for $\boldsymbol{\epsilon}_i$ without the aid of latent variables. Although marginal models like (41.2) can handle unbalanced longitudinal data they do not model the heterogeneity across individual subjects beyond the recorded covariates. The model with random effects originally proposed by *Laird* and *Ware* [41.1] readily accommodates both the unbalanced nature of the data and the heterogeneity across subjects.

For continuous responses, their model is specified as:

$$\boldsymbol{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \,. \tag{41.3}$$

Here, $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{in_i})^{\mathrm{T}}$, is the n_i-vector of longitudinal readings for subject i. Fixed covariates including the intercept and possibly deterministic functions of time (such as a linear term of time) from subject i are represented by the $n_i \times p$ matrix $\mathbf{X}_i = (\boldsymbol{x}_{i1}^{\mathrm{T}}, \cdots, \boldsymbol{x}_{in_i}^{\mathrm{T}})^{\mathrm{T}}$, with associated p-vector of coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^{\mathrm{T}}$. The j-th row of $\mathbf{X}_i$, denoted $\boldsymbol{x}_{ij}^{\mathrm{T}}$, is thus a p-vector of covariate values measured at the j-th occasion. Covariates for random effects are denoted by the $n_i \times q$ matrix $\mathbf{Z}_i$, which is often a subset of $\mathbf{X}_i$ although does not have to be. The q-vector of individual regression coefficients, the random effects $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})^{\mathrm{T}}$, are taken to be independent, multi-normally distributed with mean $\mathbf{0}$ or non-zero (see the example of random intercept only that follows) and variance–covariance $\mathbf{D}$. The error $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^{\mathrm{T}}$ is an n_i-vector that is uncorrelated with $\boldsymbol{b}_i$, independent normals with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_i$, which is often assumed to be a n_i-diagonal matrix of $\sigma^2\mathbf{I}_{n_i}$. Given the random effects, the timings of covariates and $\boldsymbol{Y}$ are assumed to be non-informative.

Marginally, the $\boldsymbol{Y}_i$ are independent normals with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathrm{T}}$. For a single time point response Y_{ij}, the above model can be rewritten as:

$$Y_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}_i + \epsilon_i \,. \tag{41.4}$$

It can be seen that the covariance between two responses measured at different times points j and j' within a subject is $z_{ij}\mathbf{D}\boldsymbol{z}_{ij'}^{\mathrm{T}} = \mathrm{cov}(Y_{ij}, Y_{ij'})$.

The covariates that have random effects in this model have the means of their effects absorbed into the fixed effects so that the mean of $\boldsymbol{b}_i$ can be conveniently assumed to be zero. For a linear growth model with random intercept only, the above model becomes

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + \epsilon_{ij} \,,$$

where t_{ij} is the time of j-th repeated measure for subject i, the random intercept b_{i0} is assumed to have independent normal distribution with mean zero and variance σ_b^2, and the error term ϵ_{ij} is assumed to be independent of b_{i0} and to be normally distributed with mean zero and variance σ^2. The same model can also be written as:

$$Y_{ij} = b_0^* + \beta_1 t_{ij} + \epsilon_{ij}$$

in which the random intercept b_0^* is assumed to have a non-zero mean of β_0 and variance σ_b^2. Notice that a random effect can either be represented as having a mean or as being deviations from the mean.

For this random intercept model, the within-subject correlation coefficient is $\sigma_b/\sqrt{\sigma^2 + \sigma_b^2}$. It should be pointed out that, for the linear mixed model (41.4), the population-average inference can be made readily from the fixed-effects part, that is, $E\boldsymbol{Y}_{\overline{X}} = \overline{\mathbf{X}}\boldsymbol{\beta}$, where $\overline{\mathbf{X}}$ is the population-average covariate values. Although the number of individual random regression coefficients is large, the additional parameters that need to be estimated beyond the fixed regression coefficients are only those involved in the variance of the individual regression coefficients, which are called *variance components*. The use of random effects not only allows individualized

growth trajectories, but also conveniently accommodates within-subject correlation. The model (41.4) can be fit with the readily available software such as the SAS proc MIXED, as well as many other commercial softwares through the (restricted) maximum-likelihood method.

41.1.2 Generalized Linear Model with Latent Variables

Harville and *Mee* [41.5] extended the aforementioned linear mixed model for continuous response to clustered ordinal data using a threshold probit model, which also turned out to be suitable for longitudinal ordinal response. Their model was motivated by a study in which cattle breeders were interested in comparing sire with respect to the difficulty experienced in the birth of their offspring. There are five ordinal difficulty levels for the response variable: no problem, slight difficulty, needed assistance, considerable force needed and extreme difficulty. Let Y_{ij} be the ordinal response for the j-th birth by sire i ($i = 1, \ldots, N$; $j = 1, \ldots, n_i$) that take a value from one of the ordered difficulty categories $1, \ldots, M$, where here $M = 5$. For the threshold probit model, it is often assumed that there is an unobserved latent variable η relating to the actual observed ordinal response Y. Here, a response occurs in category m ($Y_{ij} = m$) if the latent variable η_{ij} exceeds the threshold value θ_{m-1} but does not exceed the threshold value θ_m. The ordinal response is related to the latent variable via the following probit model:

$$P(Y_{ij} = m|\eta_{ij}) = \Phi\left(\frac{\theta_m - \eta_{ij}}{\sigma}\right) - \Phi\left(\frac{\theta_{m-1} - \eta_{ij}}{\sigma}\right), \tag{41.5}$$

where Φ is the cumulative distribution of a standard normal and σ is the standard error of the residual ϵ_{ij} from the following linear mixed model for the latent response η:

$$\eta_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}_i + \epsilon_{ij} .$$

This model is similar to the model (41.4) except Y_{ij} is now replaced by η_{ij}. It can be seen that η serves both as a latent covariate and a latent response.

The ordinal response can also be specified by a threshold cumulative probit model [41.6] as:

$$P(Y_{ij} \le m|\eta_{ij}) = \Phi\left(\frac{\theta_m - \eta_{ij}}{\sigma}\right) . \tag{41.6}$$

For binary, nominal or count data, Y_{ij}, can be modeled using the generalized linear mixed model (GLMM) [41.7], which is a direct generalization of the linear mixed model (41.4):

$$g(\mu_{ij}|\boldsymbol{b}_i) = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}_i , \tag{41.7}$$

where g is the *link* function of μ. g can be probit or logit binary Y_{ij}; g can be log for count Y_{ij}; g can be logit and $\mu_{ij} = \{P(Y_{ij} = m), m \in [1, \ldots, M]\}$ for nominal Y_{ij}. It should be noted that the marginal means of Y_{ij} are no longer $\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$, but a more complicated function depending on the form of the link function g and on the mean of the response itself. Both SAS proc NLMIXED and the STATA add-on function GLLAMM [41.4] can fit the nonlinear mixed model with logit and probit links for binomial data, logit link for polychotomous data, probit and cumulative probit for ordinal data, and log link for Poisson data.

41.1.3 Model with Hierarchical Latent Variables

Consider, the example of the National Youth Survey data analyzed by *Duncan* et al. [41.8], where repeated measures were taken from individuals nested within households nested within geographical areas. The resulting data structure, therefore, consists of four levels: repeated observations within a subject (level 1), subjects (level 2), households (level 3), and geographical areas (level 4). The response variable of interest is recorded as a scale of substance use. A question that naturally arises is whether each level in the data structure has its own submodel representing the structural relations and variability occurring at that level.

Since there was no evidence of variation among the geographical areas, a three-level model without the fourth level is presented here. Let $j = 1, \ldots, n_i$, $i = 1, \ldots, N$ and $k = 1, \ldots, K$ index respectively the repeated observations within a subject, subjects within a household, and households. Let t_{kij} denote the j-th time point when the measurements for subject i in household k are taken. The level-1 within-individual growth model can be expressed as:

$$Y_{kij} = \boldsymbol{x}_{kij}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{kij}^{\mathrm{T}}\boldsymbol{\eta}_{ki} + \boldsymbol{\epsilon}_{kij} , \tag{41.8}$$

where Y_{kij} is the response for subject i in household k at the j-th time point, $\boldsymbol{x}_{kij} = \boldsymbol{z}_{kij} = (1, t_{kij})^{\mathrm{T}}$ is a vector of covariates including the intercept and the time. The fixed coefficients vector $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ includes the intercept and the slope of time for Y_{kij}. The errors $\boldsymbol{\epsilon}_{ki} = (\epsilon_{ki1}, \ldots, \epsilon_{kin_i})^{\mathrm{T}}$ are assumed to be normally distributed with mean zerp and covariance matrix $\mathbf{R}_i$,which is a diagonal matrix with diagonal elements σ^2.

The individual random effects vector $\boldsymbol{\eta}_{ki} = (\eta_{ki0}, \eta_{ki1})^{\mathrm{T}}$ including the intercept and the slope is modeled in the following level-2 model for individuals within the same household as:

$$\boldsymbol{\eta}_{ki} = \boldsymbol{b}_k + \boldsymbol{\xi}_{ki} \,, \tag{41.9}$$

where the entries in the household-level random effects vector $\boldsymbol{b}_k = (\boldsymbol{b}_{0k}, \boldsymbol{b}_{1k})^{\mathrm{T}}$ at level 3are assumed to be bivariate normal distributions with mean zero and covariance matrix $\mathbf{D}$, and where $\boldsymbol{\xi}_{ki}$ is the residual vector, which is independent across different subjects, uncorrelated with $\boldsymbol{b}_k$, and bivariate normal distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

Models (41.8) and (41.9) can be combined into the following single model:

$$Y_{kij} = \boldsymbol{x}_{kij}^{\mathrm{T}}\boldsymbol{\beta} + z_{kij}^{\mathrm{T}}\boldsymbol{b}_k + z_{kij}^{\mathrm{T}}\boldsymbol{\xi}_{ki} + \boldsymbol{\epsilon}_{ki}, \tag{41.10}$$

in which, there are coefficients of fixed effects $\boldsymbol{\beta}$, random individual effects $\boldsymbol{\xi}_{ki}$ nested within a household and random household effects $\boldsymbol{b}_k$. The covariance between two repeated measures at different time points j and j' for a subject i is $z_{kij}\mathbf{D}z_{kij'}^{\mathrm{T}} + z_{kij}\boldsymbol{\Sigma}z_{kij'}^{\mathrm{T}}$ and the covariance between measurements from two different subjects i and i' within a household is $z_{kij}\mathbf{D}z_{kij'}^{\mathrm{T}}$. This example showed that the multilevel model can be formulated and estimated within the linear mixed model framework [41.9].

41.2 Latent Variable Models for Multidimensional Longitudinal Data

Longitudinal studies offer us an opportunity to develop detailed descriptions of the process of growth and development or of the course of progression of chronic diseases. Most longitudinal analyses focus on characterizing change over time in a single outcome variable and identifying predictors of growth or decline. Both growth and degenerative diseases, however, are complex processes with multiple markers of change, so that it may be important to model more than one outcome measure and to understand their relationship over time.

41.2.1 Extended Linear Mixed Model for Multivariate Longitudinal Responses

Lin et al. modeled multiple continuous longitudinal responses by using a mixed effects model for each of the longitudinal responses; the correlations among the different longitudinal responses were modeled through intercorrelated random effects across the mixed effects models; and the model naturally allows different measurement schedules for different types of longitudinal responses even within a same subject [41.10]. The model was illustrated with the data example from a trial of chemoprevention of cancer with β-carotene [41.11]. The trial was a randomized double-blind placebo-controlled trial with 264 patients whose primary objective was to determine whether supplemental β-carotene (50 mg/d) reduced recurrence of the primary tumors in patients cured from a recent early-stage head and neck cancer. The trial concluded that the β-carotene supplementation had no significant effect on second head and neck cancers. During the trial, blood samples of the patients were collected at about 0, 3 months, 12 months and yearly thereafter until 60 months. Several plasma nutrient concentrations were determined from the available blood samples. Analysis was focused on plasma concentrations of lycopene and lutein +zeaxanthin.

Let $i = 1, \ldots, N$ index the i-th subject, $k = 1, \ldots, K$ index the k-th type of the longitudinal response variables and $j = 1, \ldots, n_{ki}$ index the j-th time point when type k-th response is measured in subject i. For the above example $K = 2$ with $k = 1$ and 2 indexing lycopene and lutein+zeaxanthin, respectively. Let $\boldsymbol{x}_{kij}$ denote a p_k-vector of fixed effects covariates for the type k longitudinal response measured at the j-th time in subject i, which includes the intercept, the baseline plasma cholesterol concentration, the treatment assignment indicator of β-carotene, site (0 for Connecticut and 1 for Florida), age, sex, smoking status ($\{0, 0\}$ for nonsmoker, $\{1, 0\}$ for transient smoker and $\{0, 1\}$ for steady smoker), and linear and quadratic terms of time. The vector of random effects z_{kij} is a q_k-subvector of $\boldsymbol{x}_{kij}$ that includes the intercept and linear and quadratic terms of time. The model for the multiple longitudinal responses is specified as:

$$Y_{kij} = \boldsymbol{x}_{kij}^{\mathrm{T}}\boldsymbol{\beta}_k + z_{kij}^{\mathrm{T}}\boldsymbol{b}_{ki} + \epsilon_{kij} \,. \tag{41.11}$$

In the above specification, $\boldsymbol{x}_{kij}$ and z_{kij} are associated with fixed regression coefficients $\boldsymbol{\beta}_k$ and random coefficients $\boldsymbol{b}_{ki}$, respectively. The q_k-vector of random effects $\boldsymbol{b}_{ki}$ is assumed to be independent across the i and multinormally distributed with mean $\mathbf{0}$ and covariance $\mathbf{D}_{kk}$. The n_{ki}-vector of error term $\boldsymbol{\epsilon}_{ki} = (\epsilon_{ki1}, \ldots, \epsilon_{kin_{ki}})^{\mathrm{T}}$ is uncorrelated with $\boldsymbol{b}_{ki}$, independent normal with mean $\mathbf{0}$ and variance–covariance matrix $\mathbf{R}_{ki}$. The correlations

between different types of longitudinal responses within a same subject are built through the covariance of the random effects by assuming $\mathrm{cov}(\boldsymbol{b}_{ki}, \boldsymbol{b}_{k'i}) = \mathbf{D}_{kk'}$ for $k \neq k'$ and therefore $\mathrm{cov}(Y_{kij}, Y_{k'ij'}) = \boldsymbol{z}_{kij}\mathbf{D}_{kk'}\boldsymbol{z}_{k'ij'}^{\mathrm{T}}$. The covariance between the repeated measures of same type at two different time points j and j' is $\boldsymbol{z}_{kij}\mathbf{D}_{kk}\boldsymbol{z}_{kij'}^{\mathrm{T}}$

Let $\mathbf{X}_{Ki}$ and $\mathbf{Z}_{Ki}$ denote the covariate matrices for fixed and random effects, respectively.The model (41.11) can be re-expressed exactly as (41.3) if the covariates matrices and the variance–covariance matrices are rearranged in the following way:

$$\mathbf{X}_{\mathbf{i}} = \begin{pmatrix} \mathbf{X}_{1i} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{X}_{Ki} \end{pmatrix}, \quad \mathbf{Z}_{\mathbf{i}} = \begin{pmatrix} \mathbf{Z}_{1i} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{Z}_{Ki} \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \dots & \mathbf{D}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{K1} & \dots & \mathbf{D}_{KK} \end{pmatrix}, \quad \mathbf{R}_{\mathbf{i}} = \begin{pmatrix} \mathbf{R}_{1i} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{R}_{Ki} \end{pmatrix}.$$

Using these expressions, the model (41.11) becomes

$$\boldsymbol{Y}_{\mathbf{i}} = \mathbf{X}_{\mathbf{i}}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{i}}\boldsymbol{b}_{\mathbf{i}} + \boldsymbol{\epsilon}_{\mathbf{i}} ,$$

where $\boldsymbol{Y}_{\mathbf{i}} = (\boldsymbol{Y}_{1i}^{\mathrm{T}}, \dots, \boldsymbol{Y}_{Ki}^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{Y}_{ki} = (Y_{ki1}, \cdots, Y_{kin_{ki}})^{\mathrm{T}}$ being a long vector of all K longitudinal response for subject i, $\boldsymbol{\epsilon}_{\mathbf{i}} = (\boldsymbol{\epsilon}_{1i}^{\mathrm{T}}, \dots, \boldsymbol{\epsilon}_{Ki}^{\mathrm{T}})^{\mathrm{T}}$ is a long vector of error terms, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \dots, \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$ is a long vector of fixed coefficients and $\boldsymbol{b}_{\mathbf{i}} = (\boldsymbol{b}_{1i}^{\mathrm{T}}, \dots, \boldsymbol{b}_{Kij}^{\mathrm{T}})^{\mathrm{T}}$ is a long vector of random coefficients. This model has an identical expression and meaning as (41.3).

It is straightforward to see that by using proper link functions the model (41.11) can be further extended for mixed continuous and categorical longitudinal responses.

41.2.2 Measurement Error Model

Multiple outcomes are sometimes needed to jointly characterize an effect of interest properly. *Roy* considered the situation where multiple longitudinal outcomes are assumed to measure an underlying quantity of main interest from different perspectives [41.12]. Although separate linear mixed models can be fitted for each outcome, this approach is limited by the fact that it fails to borrow strength across the outcome variables. By exploiting the correlation structure with a multivariate longitudinal model, efficiency and power could be greatly increased. Since different outcomes are often measured on different scales and different units, it is of substantial interest to develop a statistical model to account for this special feature of the data. Correlation within each outcome over time and between outcomes on the same unit must be taken into account.

Roy analyzed the methadone treatment practices data in which methadone treatment is important in reducing illicit drug use and preventing HIV transmission and is effective when certain critical treatment practices are followed. The sampling unit is the treatment practice unit. The three longitudinal outcomes measuring the effectiveness of methadone treatment, including the maximum dose level [$\boldsymbol{Y}_1 = \log(\text{maximum dose})$], unit-average length of treatment ($\boldsymbol{Y}_2$), and percentage of clients receiving decreasing doses [$\boldsymbol{Y}_3 = \log(\text{percentage})$], were collected at three follow-up times. Analysis of this data set is challenging due to the fact that the outcome of major interest, the effective treatment practices level, is not observable, although several surrogates are available. In the following illustration, the same notation as for model (41.11) are used.

Suppose that the K longitudinal outcomes attempt to characterize a latent outcome of major interest, η_{ij}, e.g., the treatment practices level in unit i at the j-th follow-up time in the methadone example. One way to view this problem is that each type of observed outcome (Y_{kij}, $k = 1, \dots, K$) measures the latent variable η_{ij} with error. It is likely that the measurement error for each outcome from the same unit is correlated over time. A linear mixed model is assumed to relate Y_{kij} to η_{ij}:

$$Y_{kij} = \beta_{k0} + \beta_{k1}\eta_{ij} + b_{ki0} + \epsilon_{kij} , \tag{41.12}$$

where the measurement error term ϵ_{kij} is independent normal with mean zero and variance σ_k^2 and the type-specific random intercept b_{ki0} is independent between different units and assumed to have a normal distribution with mean zero and variance σ_{kb}^2. Correlation between the different types of the outcomes in an unit is due to the shared latent variable η_{ij}. The observed outcome Y_{kij} then measures the underlying true treatment practice evolution with error. The factor loading β_{k1} and the type-specific intercept β_{k0} are used to accommodate the fact that different types of outcomes have different scales. Each unit has its random intercept b_{ki0} for the type-k outcome, which is a random deviation from the type-specific intercept β_{k0}. For the sake of identifiability, β_{11} is set to one.

A linear mixed model is assumed to describe the effects of covariates on the latent variable η_{ij} of the underlying treatment practice:

$$\eta_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{a}_i + \xi_{ij} , \tag{41.13}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{a}_i$ are defined similarly to $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ in the linear model (41.4), $\boldsymbol{a}_i$ is a $q-$vector with normal($\mathbf{0}$, $\mathbf{D}$) and the residual ξ_{ij} is distributed with independent normal$(0, \sigma^2)$. $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are vectors of fixed and random covariates at the j-th time point for unit i that are defined similarly to those in the linear model (41.4) except that $\boldsymbol{x}_{ij}$ does do not contain the intercept.

It is often of substantial interest to estimate the unit-specific latent variables η_{ij}. The estimates of the latent effective practices score via the posterior mean $E(\boldsymbol{\eta}_i|\mathbf{Y_i})$ can be used to identify the units whose treatment practices effectiveness are well below those of a typical unit. The model also provides a straightforward way to estimate and test for global covariate effects since the parameters $\boldsymbol{\alpha}$ represent the effects of the covariates on the overall effective treatment practices level in the methadone data. Estimates of the parameters in (41.12) and (41.13) can be obtained via the EM algorithm as described by Roy or using the STATA add-on function GLLAMM of *Rabe-Hesketh* et al. [41.4].

Extension of the model to allow mixed discrete and continuous outcomes is straightforward, e.g. the model (41.12), can be modified to allow discrete outcomes through the GLM formulation:

$$g_k(\mu_{kij}) = \beta_{k0} + \beta_{k1}\eta_{ij} + b_{ki0} \,,$$

where g_k is a link function specific to the type-k outcome and $\mu_{kij} = E(Y_{kij})$.

41.3 Latent Class Mixed Model for Longitudinal Data

The linear mixed model is a well-known method for incorporating heterogeneity (for example, subject-to-subject variation) into a statistical analysis for continuous responses. However heterogeneity cannot always be fully captured by the usual assumptions of normally distributed random effects. Latent class mixed models offer a way of incorporating additional heterogeneity which can be used to uncover distinct subpopulations, to incorporate correlated non-normally distributed outcomes and to classify individuals into *risk* classes. *Lin* et al. and *McCulloch* et al. used a latent class mixed model [41.13, 14] to model the trajectory of longitudinal prostate specific antigens (PSAs) before diagnosis of prostate cancer from a retrospective study of nutritional prevention of cancer (NPC) trials in which subjects were randomized to either selenium-supplement groups or the control group [41.15, 16]. Serial PSA levels were determined retrospectively from frozen blood samples that had been collected from all patients at successive clinic visits. The PSA data set that was analyzed consists of 1182 subjects with a highly variable number of readings (range 120, median 4) per subject at irregularly spaced intervals.

These latent class mixed models assume that there are K latent classes, with each class representing a subpopulation that has its own trajectories of longitudinal responses. Suppose we have N subjects indexed by $i = 1, \ldots, N$, and K latent classes labeled by $k = 1, \ldots, K$. We define $C_{ik} = 1$ if subject i is member of class k and 0 otherwise. The probability that subject i belongs to latent class k is described through the multinomial distribution of the class membership vector for subject i, $\boldsymbol{C}_i = (C_{i1}, \cdots, C_{iK})^{\mathrm{T}}$, modeled via a logit model with covariate vector $\boldsymbol{v}_i = (v_{i1}, \cdots, v_{im})^{\mathrm{T}}$ and associated class-specific coefficient vector $\boldsymbol{\phi}_k$:

$$\pi_{ik} = P(C_{ik} = 1) = \frac{\exp\left(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\phi}_k\right)}{\sum_{s=1}^{K}\exp\left(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\phi}_s\right)}, \qquad (41.14)$$

where π_{ik} denotes the probability that subject i belongs to latent class k and $\boldsymbol{\phi}_k$ is the coefficient vector for class k with $\boldsymbol{\phi}_1 = 0$.

Each subpopulation has its own model for the longitudinal response with subpopulation differences entering the mean:

$$\boldsymbol{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{b}_i + \mathbf{W}_i(\boldsymbol{\Gamma}\,\boldsymbol{C}_i) + \boldsymbol{\epsilon}_i \,. \qquad (41.15)$$

Here, $\boldsymbol{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are defined in the same way as those in model (41.3). Covariates for class-specific effects are denoted by the $n_i \times p_w$ matrix $\mathbf{W}_i$, which has a similar structure to $\mathbf{X}_i$. There may be overlap of the covariate effects in $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{W}_i$. The class-specific regression parameters are in the $p_w \times K$ matrix $\boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_K)$, with $\boldsymbol{\gamma}_k$ being a p_w-dimensional column vector containing the parameters specific to class k. Given $C_{ik} = 1$, we have $\boldsymbol{\Gamma}\,\boldsymbol{C}_i = \boldsymbol{\gamma}_k$ for $k = 2, \ldots, K$ and we take $\boldsymbol{\gamma}_1 = 0$ to assure identifiability.

The model (41.15) captures common characteristics of the longitudinal trajectories within a subpopulation through latent classes while accommodating the variability among subjects in the same class through random effects. The use of a mixture of multivariate normal distributions for the longitudinal response $\boldsymbol{Y}$ provides flexibility that allows non-normal distributions for random effects.

The variables included in the model are as follows. The covariate vector $\boldsymbol{v}$ used to predict class membership

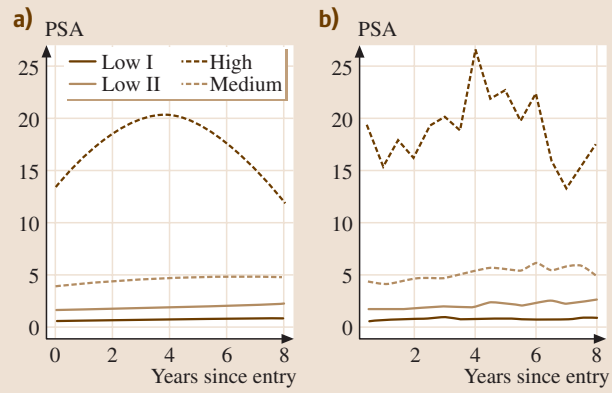

Fig. 41.1a,b Longitudinal trajectories of PSA for the four-class models. PSA values were fitted to the log-transformed data and then back-transformed to the original scale for plotting. The observed trajectories in the *right panel* are calculated by first dividing the time period into six-month intervals. For each subject, for each interval, the available PSA readings are averaged. The observed PSA trajectory for class k are calculated as averages weighted by each individual's estimated probability of class-k membership: **(a)** fitted trajectories for the four-class model; **(b)** observed trajectories for the four-class model

in (41.14) contains the treatment assignment indicator of selenium (Se) supplementation group, age at randomization, baseline PSA and Se level at randomization. The longitudinal biomarker value $\boldsymbol{Y}$ in (41.15) is the vector of log(PSA+1). The fixed effect covariate vector $\boldsymbol{X}$ contains the treatment assignment indicator, age at randomization and Se level at randomization, and linear and quadratic terms of visit time expressed in years since entry into the trial. The covariates for the random effects and class-specific effects, $\boldsymbol{Z}$ and $\boldsymbol{W}$, also contain an intercept and linear and quadratic terms of years since entry.

The four-class solution from the above latent class mixed model identifies fitted PSA trajectory classes that are labeled as "Low I","Low II", "Medium" and "High" Fig. 41.1. The majority classes "Low I" and "Low II" are characterized by a consistently low PSA level throughout the trial period. The "Medium" class has a higher PSA level than the two "Low" classes throughout the trial; the PSA level increases over time for this class. The minority class "High" has the highest PSA level at the beginning of the trial, and the predicted level of PSA increases over time until the fourth year after randomization and then decreases. In comparison the usual linear mixed model (41.4) would only be able to give one PSA trajectory that is rather flat.

Extension of the latent class mixed model to simultaneously modeling of multiple longitudinal responses is straightforward.

41.4 Structural Equation Model with Latent Variables for Longitudinal Data

Structural equation models (SEM) refer to the models that additionally specify the regression relationships among latent variables themselves. Models (41.9) and (41.13) can be regarded as SEMs.

Modeling growth within the SEM framework is a more recent approach for studying developmental trends. Because the SEM approach offers more flexibility in testing different research hypotheses about the developmental trend, many researchers have argued in favor of its superiority over some other analytic approaches [41.17, 18]. These models have provided researchers with an array of tools to interpret longitudinal data, understand developmental processes, and formulate new research questions.

Frosch et al. studied the relationship between tobacco and illicit drug use of cocaine and heroin among 166 methadone-maintained persons participating in a smoking-cessation intervention [41.19]. After completing a two-week screening period, participants were randomly assigned to one of four conditions: (a) contingency management (CM; a behavioral treatment in which participants receive increasingly valuable incentives for providing successive breath samples documenting smoking abstinence; $n = 44$); (b) relapse prevention (RP; a cognitive–behavioral group treatment providing educational and skills training information for smoking cessation; $n = 42$); (c) CM and RP combined ($n = 46$); and (d) a control condition in which participants received neither CM nor RP ($n = 43$). During the 12-week treatment period, participants provided urine and breath samples for heroin and cocaine toxicology and measurement of expired CO three times weekly (Monday, Wednesday, and Friday). The impact of use of heroin and cocaine on levels and changes in cigarette use was assessed with latent growth models in structural equations framework. The time axis is divided into two-week periods for the 12 weeks of treatment. Scales for the use of heroin, cocaine, and

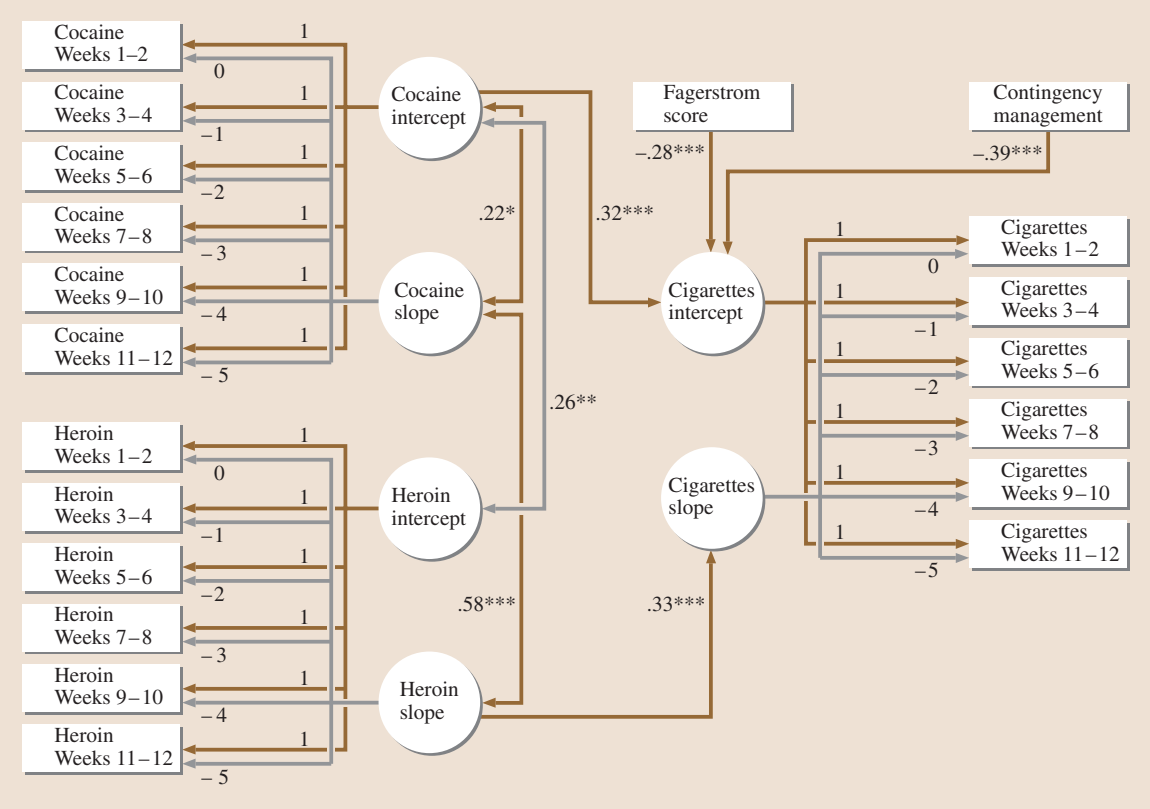

Fig. 41.2 Final latent-growth model presenting significant predictors of levels and trajectory of change in cigarette use over time. Intercepts were fixed to unity; slopes were hypothesized to be equal-interval linear trend coefficients. Double-headed arrows represent correlations; single-headed arrows represent regressions. Higher levels of cocaine use predicted higher levels of cigarette use; accelerated use of heroin predicted acceleration in use of cigarettes. *Circles* indicate latent variables; *rectangles* indicate measured variables. Parameter estimates are standardized. *Fagerstrom* = Fagerstrom test for nicotine dependence (after *Frosch* et al. [41.19])

cigarette were constructed for each of the two-week periods, and these scales were used as the longitudinal response variables.

Let $k = 1, 2, 3$ index the responses of cigarette, heroin and cocaine. For $k = 2$ or 3, an extended linear mixed model such as (41.11) is specified for heroin or cocaine response:

$$Y_{kij} = \boldsymbol{x}_{kij}^{\mathrm{T}}\boldsymbol{\beta}_k + \boldsymbol{z}_{kij}^{\mathrm{T}}\boldsymbol{\eta}_{ki} + \epsilon_{kij} \ ,$$

where the fixed covariates $\boldsymbol{x}_{kij}$ include the intercept, the linear term in time in weeks and the treatment dummy variables, and where $\boldsymbol{z}_{kij}$ is a q_k-vector that includes only the intercept and the linear term in time; the model has exactly the same definitions as those of (41.11) except that the coefficients of the random effects $\boldsymbol{b}_{ki}$ in (41.11) are replaced by $\boldsymbol{\eta}_{ki}$ here. The model describes the possible improvement in heroin or cocaine use over the course of treatment and accounts for the repeated measures of the same type within the same subject with the random effects of intercept and slope included in $\boldsymbol{\eta}_{ki}$. The correlation between the responses of the two different types of heroin and cocaine use is modeled with the covariance of $\boldsymbol{\eta}_{2i}$ and $\boldsymbol{\eta}_{3i}$. The intercept represents the individual baseline level of use of cocaine or heroin. The slope represents the trend of the growth curve.

With the additional $p_{w,1}$-vector of fixed covariates of $\boldsymbol{w}_{1i}$ including the Fagerstrom score and contingency management measure, the slopes and intercepts of cocaine use and heroin use are used as the latent predictors in the following model to ascertain the impact of their initial levels and their own dynamic changes (slopes) on

predicting the slope and intercept of cigarette use:

$$\boldsymbol{\eta}_{1i} = \boldsymbol{\Lambda}_2\boldsymbol{\eta}_{2i} + \boldsymbol{\Lambda}_3\boldsymbol{\eta}_{3i} + \boldsymbol{\Gamma}_1\boldsymbol{w}_{1i} + \boldsymbol{\xi}_{1i}\,, \tag{41.16}$$

where $\boldsymbol{\Lambda}_k$ $(k = 2, 3)$ is a $q_k \times q_k$ ($q_k = 2$ here) diagonal matrix of factor loadings for $\boldsymbol{\eta}_{ki}$ with the s-th diagonal element being $\lambda_{k,s}$, $\boldsymbol{\Gamma}_1$ is a $q_1 \times p_{w,1}$ ($q_1 = p_{w,1} = 2$ here) matrix of regression coefficients associated with fixed covariates $\boldsymbol{w}_{1i}$ and ξ_{1i} is a q_1-vector of residuals.

The model (41.16) can be written as the following structural equation for the relationships among the latent variables:

$$\boldsymbol{\eta}_{\mathbf{i}} = \boldsymbol{\Lambda}\boldsymbol{\eta}_{\mathbf{i}} + \boldsymbol{\Gamma}\boldsymbol{w}_i + \boldsymbol{\xi}_i\,, \tag{41.17}$$

where $\boldsymbol{\eta}_{\mathbf{i}} = (\boldsymbol{\eta}_{1i}^{\mathrm{T}}, \ldots, \boldsymbol{\eta}_{Ki}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Lambda}$ is a $\sum_{k=1}^{K} q_k \times \sum_{k=1}^{K} q_k$ upper-diagonal matrix of coefficients, $\boldsymbol{w}_i$ is a vector of $\sum_{k=1}^{K} p_{w,k}$ covariates, $\boldsymbol{\Gamma}$ is a $\sum_{k=1}^{K} q_k \times \sum_{k=1}^{K} p_{w,k}$ matrix of regression coefficients and $\boldsymbol{\xi}_i$ is a $\left(\sum_{k=1}^{K} q_k\right)$-vector of residuals. $\boldsymbol{\Lambda}$ is upper-diagonal, which implies that there are no simultaneous effects with latent variable 1 regressed on latent variable 2 and vice versa. The lower-level latent variables (e.g., the $\boldsymbol{\eta}_{1i}$ for cigarette use) come before the higher-level ones (e.g., the $\boldsymbol{\eta}_{2i}$ and $\boldsymbol{\eta}_{3i}$ for heroin and cocaine use) in the $\boldsymbol{\eta}$ vector, an the upper-diagonal $\boldsymbol{\Lambda}$ matrix ensures that lower-level latent variables can be regressed on higher ones but not the reverse since it would not make sense to regress a higher-level latent variable on a lower-level one. Some elements of the upper-diagonal matrix $\boldsymbol{\Lambda}$ can be additionally set to zero, which indicates that a latent variable does not depend on a corresponding higher-level one.

Using this structural equation model for the methadone-maintenance data, a significant relationship during the treatment period between rate of change in heroin and rate of change in tobacco use was revealed, with increased heroin use corresponding to increased tobacco use. Although levels of cocaine use were related to levels of tobacco use, there was no significant relationship between the rates of change of the two substances. *Frosch* et al.'s findings demonstrate the utility of latent growth models with the structural equation approach for analyzing short-term clinical trial data and strongly suggest that successful smoking cessation in this population requires a concurrent focus on reducing heroin use. The final model that *Frosch* et al. used is represented by a path diagram, as shown in Fig. 41.2.

41.5 Concluding Remark: A Unified Multilevel Latent Variable Model

In the case of hierarchical data including longitudinal data, the term *level* is often used to describe the position of a unit of observation within a hierarchy of units, typically reflecting the sampling design. Here level-1 units are nested in level-2 units, which are nested in level-3 units, a typical example being patients in clinics in regions. In this context, a random effect is said to vary at a given level, e.g. at the region level, if it varies between regions but, for a given region, is constant for all clinics and patients belonging to that region. If a repeated measure is taken on the patients and the regions are ignored, then time points are the level-1 units, patients are the level-2 units and clinics are the level-3 units. The multilevel models assume that lower-level units are conditionally independent given the higher-level latent variables and the explanatory variables. The latent variables at the same level are usually assumed to be mutually correlated whereas latent variables at different levels are independent. The aim of multilevel modeling is to analyze data simultaneously from different levels of the hierarchy. All the models discussed in this chapter can be regarded as special cases of a multilevel model with latent variables for longitudinal data.

The expression for the s-th element of $\boldsymbol{\eta}$ $(s = 1, \ldots, \sum_{k=1}^{K} q_k)$ in an SEM such as (41.16) can be substituted into the expression for $(s-1)$-th element, which can be substituted into the expression for $(s-2)$-th element, and so forth. (i. e., recursive substitution). Then, using similar notational definitions to those documented by *Rabe–Hesketh* et al. [41.4], we obtain an equation of the following form for longitudinal data with constraints among the parameters:

$$g(\mu_{kij}) = \boldsymbol{x}_{kij}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{l=2}^{L}\sum_{s=1}^{q_l} \boldsymbol{z}_{kij,s}^{(l),\mathrm{T}}\boldsymbol{\lambda}_s^{(l)}\eta_{ki,s}^{(l)}\,, \tag{41.18}$$

where $l = 1, \ldots, L$ indexes the L levels, there are q_l latent variables at level l, $\eta_{ki,s}^{(l)}$ is the s-th latent variable for subject i in the type-k response at level l, $\boldsymbol{x}_{kij}$ and $\boldsymbol{z}_{kij,s}^{(l)}$ are two vectors of explanatory variables associated with fixed and latent variables and $\boldsymbol{\lambda}_s^{(l)}$ is the vector of factor loadings for the s-th latent variable $\eta_{ki,s}^{(l)}$ in level l. In the general form of equation (41.18), the latent

variables η can be continuous or discrete (e.g., latent classes).

Multilevel modeling techniques offer researchers the opportunity not only to analyze their data in a more technically appropriate manner than traditional single-level methods allow, but also to extend the range of research questions to a level with more contextual richness and complexity.

References

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

41.2 P. J. Diggle, P. Heagerty, K.-Y. Liang, S. L. Zeger: *Analysis of Longitudinal Data*, 2nd edn. (Oxford Univ. Press, Oxford 2002)

41.3 K.-Y. Y. Liang, S. L. Zeger: Longitudinal data analysis using generalized linear models, Biometrika **73**, 13–22 (1986)

41.4 S. Rabe-Hesketh, A. Skrondal, A. Pickles: *GLLAMM Manual*, U.C. Berkeley Division of Biostatistics Working Paper Series, Vol. 160 (2004) http://www.gllamm.org

41.5 D. A. Harville, R. W. Mee: A mixed-model procedure for analyzing ordered categorical data, Biometrics **40**, 393–408 (1984)

41.6 J. Catalano P. Bivariate modelling of clustered continuous and ordered categorical outcomes, Stat. Med. **16**, 883–900 (1997)

41.7 C. E. McCulloch, S. R. Searle: *Generalized, Linear, and Mixed Models* (Wiley, New York 2001)

41.8 T. E. Duncan, S. C. Duncan, H. Okut, L. A. Strycker, F. Li: An extension of the general latent variable growth modeling framework to four levels of the hierarchy, Struct. Equ. Model. **9**(3), 303–326 (2002)

41.9 A. Skrondal, S. Rabe-Hesketh: *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models* (Chapman Hall/CRC, New York 2004)

41.10 H. Q. Lin, C. E. McCulloch, S. T. Mayne: Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables, Stat. Med. **21**, 2369–2382 (2002)

41.11 S. T. Mayne, B. Cartmel, M. Baum, G. Shor-Posner, B. G. Fallon, K. Briskin, J. Bean, T. Z. Zheng, D. Cooper, C. Friedman, W. J. Goodwin: Randomized trial of supplemental beta-carotene to prevent second head and neck cancer, Cancer Res. **61**, 1457–1463 (2001)

41.12 J. Roy: Latent variable models for longitudinal data with multiple continuous outcomes, Biometrics **56**, 1047–1054 (2000)

41.13 H. Q. Lin, B. W. Turnbull, C. E. McCulloch, E. H. Slate: Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer, J. Am. Stat. Assoc. **97**, 53–65 (2002)

41.14 C. E. McCulloch, H. Lin, E. H. Slate, B. W. Turnbull: Discovering subpopulation structure with latent class mixed models, Stat. Med. **21**, 417–429 (2002)

41.15 L. C. Clark, G. F. Combs Jr., B. W. Turnbull, E. H. Slate, D. K. Chalker, J. Chow, L. S. Davis, R. A. Glover, G. F. Graham, E. G. Gross, A. Krongrad, J. L. Lesher, H. K. Park, B. B. Sanders, C. L. Smith, J. R. Taylor: Effects of selenium supplementation for cancer prevention in patients with carcinoma of the skin, J. Am. Med. Assoc. **276**, 1957–1963 (1996)

41.16 L. C. Clark, B. Dalkin, A. Krongrad, G. F. Combs Jr., B. W. Turnbull, E. H. Slate, R. Witherington, J. H. Herlong, E. Janosko, D. Carpenter, C. Borosso, S. Falk, J. Rounder: Decreased incidence of prostate cancer with selenium supplementation: results of a double-blind cancer prevention trial, Brit. J. Urol. **81**, 730–734 (1998)

41.17 J. J. McArdle: A latent difference score approach to longitudinal dynamic analysis. In: *Structural Equation Modeling: Present and Future*, ed. by R. Cudeck, S. DuToit, D. Sörbom (Scientific Software International, Lincolnwood, IL 2001) pp. 341–380

41.18 J. J. McArdle, E. Ferrer-Caja, F. Hamagami, R. W. Woodcock: Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span, Devel. Psychol. **38**, 115–142 (2002)

41.19 D. L. Frosch, J. A. Stein, S. Shoptaw: Use latent-variable models to analyze smoking cessation clinical trial data: an example among the methadone maintained, Exp. Clin. Psychopharmacol. **10**, 258–267 (2002)