REPEATED MEASURES AND MULTILEVEL MODELING

Geert Verbeke
Katholieke Universiteit Leuven, Leuven, Belgium

Geert Molenberghs
Limburgs Universitair Centrum Hasselt University, Belgium

Keywords: Covariance model, longitudinal data, marginal model, conditional model, random effect, linear mixed model, generalized linear mixed model, non-linear mixed model, generalized estimating equations

Contents

Summary

A general framework is sketched for model formulation and inference when outcome measurements are of a repeated nature. Emphasis is placed on the linear mixed model for Gaussian outcomes, as well as on its non-linear extension. In the non-Gaussian setting, a distinction is made between marginal, conditional, and random-effects models, whence key members of each family are discussed in detail. In particular, generalized linear mixed models and generalized estimating equations receive ample treatment.

1. Introduction

Repeated measures are obtained whenever a specific response is measured repeatedly in a set of units. Examples are hearing thresholds measured on both ears of a set of subjects, birthweights of all litter members in a toxicological animal experiment, or weekly blood pressure measurements in a group of treated patients. The last example is different from the first two examples in the sense that the time dimension puts a strict ordering on the obtained measurements within subjects. The resulting data are therefore often called longitudinal data. Obviously, a correct statistical analysis of repeated measures or longitudinal data can only be based on models which explicitly take into account the clustered nature of the data. More specifically, valid models should account for the fact that repeated measures within subjects are allowed to be correlated. For this reason, classical (generalized) linear regression models are not applicable in this context. An additional complication arises from the highly unbalanced structure of many data sets encountered in practice. Indeed, the number of available measurements per unit is often very different between units, and, in the case of longitudinal data, measurements may have been taken at arbitrary time points, or subjects may have left the study prematurely, for a number of reasons (sometimes known but mostly unknown).

A large number of models has been proposed in the statistical literature, during the last few decades. Most of them can be viewed as special cases of one general model which will be presented in this contribution. The interpretation of the different components of the model and methods for model fitting will be discussed first. Afterwards, some frequently used special cases will be presented for the analysis of continuous and for discrete data in turn. Although emphasis will be put on longitudinal data, the methods discussed are immediately applicable to other types of repeated measurements.

2. General Model

Let y_{ij} denote the *j*th measurement available for the *i*th unit, $i=1,\dots,N$, $j=1,\dots n_i$ and let y_i denote the vector of all measurements for the *i*th unit, i.e., $y_i' = (y_{i1},\dots,y_{in_i})$. Our general model assumes that y_i (possibly appropriately transformed) satisfies

$$y_i \mid b_i \sim F_i(\theta, b_i), \qquad (1)$$

i.e., conditional on b_i , y_i follows a pre-specified distribution F_i , possibly depending on covariates, and parameterized through a vector θ of unknown parameters, common to all subjects. Further, b_i is a *q*-dimensional vector of subject-specific parameters, called random effects, assumed to follow a so-called mixing distribution *G* which may depend on a vector ψ of unknown parameters, i.e., $b_i \sim G(\psi)$. The b_i reflect the between-unit heterogeneity in the population with respect to the distribution of y_i . Different factorizations of F_i will lead to different models. For example, considering the factors made up of the outcomes y_{ij} given its predecessors $(y_{i1},\dots,y_{i,j-1})'$ leads to a so-called transitional model. A model without any random effects b_i is called a marginal model for the response vector y_i . In the presence of random effects, conditional independence is often assumed, under which the components y_{ij} in y_i are independent, conditional on

b_i. The distribution function F_i in Eq. (1) then becomes a product over the n_i independent elements in y_i.

In general, unless a fully Bayesian approach is followed, inference is based on the marginal model for y_i which is obtained from integrating out the random effects, over their distribution $G(\psi)$. Let $f_i(y_i \mid b_i)$ and $g(b_i)$ denote the density functions corresponding to the distributions F_i and *G*, respectively, we have that the marginal density function of y_i equals

$$f_i(y_i) = \int f_i(y_i \mid b_i) g(b_i) db_i, \tag{2}$$

which depends on the unknown parameters θ and ψ. Assuming independence of the units, estimates of $\hat{\theta}$ and $\hat{\psi}$ can be obtained from maximizing the likelihood function built from Eq. (2), and inferences immediately follow from classical maximum likelihood theory.

Obviously, the random-effects distribution *G* is crucial in the calculation of the marginal model Eq. (2). One approach is to leave *G* completely unspecified and to use non-parametric maximum likelihood (NPML) estimation, which maximizes the likelihood over all possible distributions *G*. The resulting estimate $\hat{\theta}$ is then always discrete with finite support. Depending on the context, this may or may not be a realistic reflection of the true heterogeneity between units. One therefore often assumes *G* to be of a specific parametric form, such as a (multivariate) normal. Depending on F_i and *G*, the integration in Eq. (2) may or may not be possible analytically. Proposed solutions are based on Taylor series expansions of $f_i(y_i \mid b_i)$, or on numerical approximations of the integral, such as (adaptive) Gaussian quadrature.

Although in practice one is usually primarily interested in estimating the parameters in the marginal model, it is often useful to calculate estimates for the random effects b_i as well. They reflect between-subject variability, which makes them helpful for detecting special profiles (i.e., outlying individuals) or groups of individuals evolving differently in time. Also, estimates for the random effects are needed whenever interest is in prediction of subject-specific evolutions. Inference for the random effects is often based on their so-called posterior distribution $f_i(b_i \mid y_i)$, given by

$$f_i(b_i \mid y_i) = \frac{f_i(y_i \mid b_i) g(b_i)}{\int f_i(y_i \mid b_i) g(b_i) db_i} \tag{3}$$

in which the unknown parameters θ and ψ are replaced by their estimates obtained earlier from maximizing the marginal likelihood. The mean or mode corresponding to Eq. (3) can be used as point estimates for b_i, yielding empirical Bayes (EB) estimates.

3. Some Models for Continuous Data

Most longitudinal models for continuous responses assume that all y_{ij} are normally distributed, possibly after appropriate transformation, i.e., it is assumed that $y_i \sim N(\mu_i, V_i)$, for some parameterization of the mean vector μ_i and covariance matrix V_i. Many different models have been proposed for μ_i as well as for V_i, some of which will be presented in the following sections.

3.1. Multivariate Linear Regression Models

The multivariate regression model is one of the most frequently used models for the analysis of balanced data, i.e., data where a fixed number n of repeated measurements is taken at fixed time points for all units. It assumes that μ_i is of the form $X_i\beta$ for some known $(n \times p)$ design matrix X_i and associated vector β of p unknown regression coefficients, and that all matrices V_i are equal to a general unstructured covariance matrix V. The parameter vector θ then consists of the regression parameters in β and the vector α of $n(n+1)/2$ variances and covariances in V. The maximum likelihood estimators of β and V satisfy

$$\hat{\beta} = \left(\sum_{i=1}^{N} X'_i \hat{V}^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X'_i \hat{V}^{-1} y_i, \tag{4}$$

$$\hat{V} = \frac{1}{N} \sum_{i=1}^{N} (y_i - X_i\hat{\beta})(y_i - X_i\hat{\beta})', \tag{5}$$

and estimates are obtained from iterating between Equations (4) and (5) until convergence is attained.

Note that, in case of high dimensional vectors y_i, the number of parameters in α is large, which may result in inefficient inferences, also for the regression coefficients in β, which are usually of primary interest. Therefore, a parsimonious parameterization of the covariance matrix V is often looked for, i.e., V is assumed to have a specific parametric form, depending on a (relatively) small number of unknown parameters, again combined into a vector α. In general, $\hat{\beta}$ then still satisfies Eq. (4), but no analytic expression is available to replace Eq. (5) such that full numerical maximization routines are required for the joint calculation of the maximum likelihood estimates $\hat{\beta}$ and $\hat{\alpha}$.

Many parametric models for V have been proposed in the statistical literature. Examples are the Toeplitz model, the first-order autoregressive model, and the compound symmetry model. Let the (k,l) element of V be denoted by v_{kl}. A Toeplitz model assumes that v_{kl} only depends on $|k-l|$ resulting in a so-called banded covariance matrix with constant covariances in bands parallel to the main diagonal. A special case is the first-order autoregressive model, AR(1), in which the elements v_{kl} satisfy $v_{kl} = \sigma^2 \rho^{|k-l|}$, for some correlation parameter ρ. Finally, the compound symmetry model corresponds to a covariance matrix V with constant variance and constant covariance. In case the assumptions of common variance are not appropriate, straightforward heterogeneous extensions can be made to each of these parametric models.

Depending on the context, as well as on the design of the study, some parametric models may not be appropriate while other models are particularly appropriate. For example, the Toeplitz and AR(1) models are only meaningful when there exists a natural strict ordering in the repeated measurements within units. For example, if units represent families, and repeated measurements correspond to measurements taken on different members of those families, no such ordering is present, implying that a covariance model should be selected which allows for exchangeability of the repeated measurements within units, i.e., the compound symmetry model or its heterogeneous version. Even for the analysis of longitudinal data, where the time dimension implies a natural ordering of the measurements within units, the Toeplitz or AR(1) structures may not be valid. For example, consider a longitudinal experiment, with n measurements taken at fixed time points t_j, $j = 1, \ldots, n$. Both models then implicitly assume that the covariance (and in case of the homogeneous models also the correlation) between y_{ik} and $y_{i(k+1)}$, $k = 1, \ldots, n-1$ does not depend on k, which is only fully interpretable if the time points t_j are equally spaced, i.e., if $|t_k - t_{k+1}|$ is independent of k as well. In examples with unequally spaced time points, so-called spatial covariance structures can be used, which model v_{kl} as $\sigma(t_k)\sigma(t_l)\rho\left(|t_k - t_l|\right)$, for some standard deviation function $\sigma(t)$ and some (usually monotonically decreasing) correlation function $\rho(u)$ with $\rho(0) = 1$. Common choices for $\rho(u)$ are the exponential serial correlation structure $\rho(u) = \exp(-\phi u)$ and the Gaussian serial correlation function $\rho(u) = \exp(-\phi u^2)$, with ϕ some unknown parameter to be estimated from the data. Note that, in case of equally spaced time points, the exponential serial correlation model reduces to the AR(1) model, and that the homogeneous versions of the spatial covariance structures are obtained from assuming $\sigma(t)$ to be constant.

Although multivariate linear regression models are primarily used in the case of balanced repeated measurements data, they can, strictly speaking, also be used to model unbalanced data. It is then assumed that the covariance matrices V_i are modeled through a fixed number of variance components, i.e., the number of variance and covariance parameters does not depend on the number of subjects included in the sample. The maximum likelihood estimator of β then still satisfies Eq. (4), but with V replaced by the subject-specific matrices V_i, and full numerical maximization routines are again required for the joint estimation of β and the covariance parameters in all V_i, which are still combined into a vector α.

Bibliography

Davidian, M. and Giltinan, D.M. (1995), *Nonlinear models for repeated measurement data*, Chapman & Hall [This book provides a general overview of nonlinear models].

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994), *Analysis of longitudinal data*, Clarendon Press, Oxford [Non-technical book, summarizing most frequently used approaches].

Fahrmeir, L. and Tutz, G. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models*. Heidelberg: Springer-Verlag [This book provides an extensive overview of models for non-normal data].

Goldstein, H. (1995), *Multilevel Statistical Models*. Kendall's Libary of Statistics 3. London: Arnold [This book provides an overview of models for repeated measures from a multilevel perspective].

Laird, N.M. and Ware, J.H. (1982), "Random-effects models for longitudinal data", *Biometrics*, **38**, 963-974 [Basic reference for linear mixed models].

Liang, K.Y. and Zeger, S.L. (1986), "Longitudinal data analysis using generalized linear models", *Biometrika*, **73**, 13-22 [Basic reference for GEE methodology].

Molenberghs, G. and Lesaffre, E. (1999),. "Marginal modelling of multivariate categorical data," Statistics in Medicine, **18**, 2237-2255 [Overview of methods for categorical data].

Verbeke, G. and Molenberghs, G. (2000), *Linear mixed models for longitudinal data*, Springer Series in Statistics, Springer-Verlag, New-York [Overview of linear mixed models for longitudinal data, with many examples, and a lot of emphasis on SAS procedure MIXED].

Vonesh, E.F. and Chinchilli, V.M. (1997), *Linear and nonlinear models for the analysis of repeated measurements*, Marcel Dekker, New-York [This book provides an extensive summary of nonlinear models].

Biographical Sketches

Geert Molenberghs is Professor of Biostatistics at the Limburgs Universitair Centrum in Belgium. He received the B.S. degree in mathematics (1988) and a Ph.D. in biostatistics (1993) from the Universiteit Antwerpen. Geert Molenberghs published methodological work on the analysis of non-response in clinical and epidemiological studies. He serves as an associate editor for Biostatistics and is Joint Editor of Applied Statistics. He is an officer of the Belgian Statistical Society and the Belgian Region of the International Biometric Society. He serves on the Executive Committee of the International Biometric Society. He has held visiting positions at the Harvard School of Public Health (Boston, MA). With Geert Verbeke, he is a co-author of books on longitudinal data.

Geert Verbeke is tenured Associate Professor at the Biostatistical Centre of the Katholieke Universiteit Leuven in Belgium. He received the B.S. degree in mathematics (1989) from the Katholieke Universiteit Leuven, the M.S. in biostatistics (1992) from the Limburgs Universitair Centrum, and earned a Ph.D. in biostatistics (1995) from the Katholieke Universiteit Leuven. Geert Verbeke wrote his dissertation, as well as a number of methodological articles, on various aspects of linear mixed models for longitudinal data analyses. He has held visiting positions at the Gerontology Research Center and the Johns Hopkins University (Baltimore, MD). He is President of the Quetelet Society, the Belgian Region of the International Biometric Society, and Associate Editor to several journals, such as Biometrics and Applied Statistics. With Geert Molenberghs, he is a co-author of books on continuous and categorical longitudinal data with Geert Molenberghs.