Mixed Models

Júlia Volaufová

LSUHSC School of Public Health, New Orleans, USA

Abstract

Mixed models, simply put, are models of a response that involve fixed and random effects (see, e.g., Demidenko (2004)). Here we give a very superficial and brief coverage of the wide variety of models this term encompasses.

Historically, the most widely-investigated mixed model is the *linear* mixed model. For an n-dimensional response vector $\boldsymbol{Y}$, the model can be expressed as

$$\boldsymbol{Y} = X\beta + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1}$$

where X and Z are fixed and known matrices of covariates with β a fixed vector parameter and $\boldsymbol{\gamma}$ a vector of random effects. The often-invoked distributional assumptions are $\boldsymbol{\gamma} \sim N_l(0, G)$ and $\boldsymbol{\epsilon} \sim N_n(0, R)$. The matrices G (nnd) and R (pd) are modeled as members of chosen classes (e.g., compound symmetry, AR(1), unstructured), which involve further unknown parameters. In special cases when the matrices G and R depend linearly on a set of unknown scalars, the covariance matrix of the response can be expressed as $\mathrm{Cov}(\boldsymbol{Y}) = \sum_{i=1}^{p} \vartheta_i V_i$ where the parameters ϑ_i are interpreted as *variance-covariance components*. $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are assumed to be mutually independent, which implies that $\mathrm{Cov}(\boldsymbol{Y}) = ZGZ' + R$.

This class of models covers a broad range of situations. Here is a partial list.

- In *repeated measures* models (see e.g., Reinsel (1982)), also called *longitudinal* models (see, e.g., Laird and Ware (1982)), multiple observations are carried out, say over time, on each individual sampling unit.
- In *cluster randomized settings* (see, e.g., Laird (2004)), dependencies between observations on sampling units are introduced due to clustering in the randomization process.
- In *hierarchical* or *multilevel* settings, a subset of parameters on a given level is considered to be a random vector whose distribution depends on an additional set of unknown parameters.
- In some situations it is possible to partition the response vector into independent subvectors, as in longitudinal models, but in many cases such partitioning is not straightforward, e.g., in some geodetic or geophysical applications (see e.g., Kubáček et al. (1995) or Fišerová et al. (2007)) when combining experiments

with different precisions, each relating to the same mean parameter. In these models it is not obvious and it is not even necessary to identify the latent random effects - the model for the response vector $\boldsymbol{Y}$ is parametrized by (unknown) fixed vector parameters of the mean and variance-covariance components.

The class of linear mixed models can be viewed within the broader context of *nonlinear* mixed models. There, the response variable Y can be modeled in general as

$$Y = f(\boldsymbol{x}, \boldsymbol{z}, \beta, \boldsymbol{\gamma}) + \epsilon. \tag{2}$$

The function $f(.)$ is a nonlinear function of fixed (β) and random ($\boldsymbol{\gamma}$) (vector) parameters as well as vectors of covariates ($\boldsymbol{x}$ and $\boldsymbol{z}$). Mostly it is assumed that $f(.)$ is differentiable with respect to β and $\boldsymbol{\gamma}$. The distributional assumptions regarding $\boldsymbol{\gamma}$ and ϵ may be the same as in the linear case.

An example of such a model (2) is a *random coefficients* model (see, e.g., Vonesh and Chinchilli (1997)) that can be set up in two stages. For stage 1, the model takes the form

$$Y_{ij} = f(t_{ij}, \beta_i) + \varepsilon_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, T_i, \tag{3}$$

where Y_{ij} is the response for subject i at time j, $f(.)$ is a nonlinear function of the p-vector of subject-specific vector parameters β_i and time (t_{ij}), and ε_{ij} is the error term, which we assume follows a normal distribution with mean zero and variance σ^2. The second stage is at the population level. At this stage the subject-specific parameters are defined by the model:

$$\beta_i = A_i\beta + B_i\boldsymbol{\gamma}_i, \qquad i = 1, \ldots, n. \tag{4}$$

In this model, β is a vector of fixed population parameters and $\boldsymbol{\gamma}_i$ is a vector of random effects for subject i. In most cases the matrix A_i takes the form $A_i = I_p \otimes a_i'$ (see, e.g., Vonesh and Carter (1987)), where the vector a_i is the vector of covariates. The matrix B_i is used to determine which elements of β_i have random components and which are fixed. A well-known example of a random coefficient model is a *growth curve model*, a special case of which is when, among other assumptions, the dimension of each subject specific response is the same. It can be expressed in terms of a *multivariate model* for a response vector $\boldsymbol{Y}$, which in general may have expectation $E(\boldsymbol{Y}) = \sum_{i=1}^{p}(C_i \otimes D_i)\boldsymbol{\beta}_i$ and covariance matrix $\mathrm{Cov}(\boldsymbol{Y}) = \Sigma_1 \otimes \Sigma_2$ with the matrices Σ_1 and Σ_2 and the vector β unknown.

Using the *generalized* mixed linear model, we model the transformed mean of Y via a *link* function as a linear function of covariates (see, e.g., McCulloch et al. (2008)). Typically, the conditional distribution of the response belongs to the exponential family, and often there is a functional relationship between the parameters of the mean and the variance-covariance components. The conditional mean of the ith observation, μ_i, is linked via a function, say $g(\mu_i)$, to the covariates and

random effects in terms of additive effects as

$$g(\mu_i) = \boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{\gamma}, \tag{5}$$

where the meaning of β and $\boldsymbol{\gamma}$ is as above. We note that the nonlinear random coefficients model can be perceived as a special case of the generalized linear mixed model.

Although these models have been notoriously studied for many decades, there is a variety of questions still to be addressed. The main aim of inference is estimation and hypotheses testing. Maximum likelihood or quasi-likelihood methods result in point estimates with good large sample properties under generally mild conditions. Point and interval estimation of variance-covariance components has kept statisticians busy for decades (see, e.g., Hartley and Rao (1967), LaMotte (1973a,b), Mathew et al. (2009), etc.). Ultimately in almost all settings one is interested in testing (linear) hypotheses about the parameter β and/or about variance-covariance components. Usually we see $H_{0_1} : H'\beta = h$ or $H_{0_2} : h'\boldsymbol{\vartheta} = 0$, or simultaneously both. Except for a few special models there is no exact test available for H_{0_1}; a variety of approximate tests has been studied for quite a time (see, e.g., Harville and Jeske (1992), Kenward and Roger (1997), Khuri et al. (1998), Volaufova and LaMotte (2008), Volaufova (2009), Livacic-Rojas et al. (2010), Volaufova and LaMotte (2012), and many others). The hypothesis H_{0_2} also has been extensively studied; however even in models with only two variance-covariance components the question of finding a test with optimal properties (in some sense) in general is still open.

In linear mixed models, the choices of the structures of G and R may be consequential, but often these choices are made arbitrarily and subjectively. Various information criteria have been developed, recommended, and modified for the purpose of informing these choices. Effects of such data-driven choices on inferential procedures for fixed effects are just beginning to be investigated and are related to the broad area of model building (see, e.g., Vaida and Blanchard (2005)).

Here we invite contributions that address pertinent questions and relate to any aspects of this broad class of mixed models.

References

Demidenko, E. (2004). *Mixed Models. Theory and Applications.* John Wiley & Sons, Inc., New York.

Fišerová, E., L. Kubáček, and P. Kunderová (2007). *Linear Statistical Models. Regularity and Singularities.* Academia, Praha.

Hartley, H.O. and J.N.K. Rao (1967). Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika 54*, 93–108.

Harville, D.A. and D.R. Jeske (1992). Mean Squared Error of Estimation or Prediction Under a General Linear Model. *J. Amer. Statist. Assoc. 87*, 724–731.

Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer.

Kenward, M.G. and J.H. Roger (1997). Small Sample Inference for Fixed Effects From Restricted Maximum Likelihood. *Biometrics 53*, 983–997.

Khuri, A.I., T. Mathew, and B.K. Sinha (1998). *Statistical Tests for Mixed Linear Models.* John Wiley & Sons, Inc., New York.

Kubáček, L., L Kubáčková, and J. Volaufová (1995). *Statistical Models with Linear Structures.* Veda, Bratislava.

Laird, N.M. and J.H. Ware (1982). Random-Effects Models for Longitudinal Data. *Biometrics 38*, 963–974.

Laird, N.(2004). Analysis of Longitudinal and Cluster-Correlated Data.*NSF-CBMS Regional Conference Series in Probability and Statistics 8*, Published by IMS, i-ii+1–155.

LaMotte, L.R. (1973). Quadratic estimation of variance components. *Biometrics 29*, 311–330.

LaMotte, L.R. (1973) On nonnegative quadratic unbiased estimation of variance components. *J. Amer. Statist. Assoc. 68*, 728–730.

Livacic-Rojas, P., G. Vallejo, and P. Fernandez (2010). Analysis of Type I Error Rates of Univariate and Multivariate Procedures in Repeated Measures Designs. *Comm. Statist. Simulation Comput. 39*, 624–640.

McCulloch, Ch.E., S.R. Searle, and J.M. Neuhaus(2008). *Generalized, Linear, and Mixed Models.* 2nd Ed. John Wiley & Sons, Inc., New York.

Mathew, T., T. Nahtman, D. von Rosen, and B.K. Sinha (2009). Nonnegative Estimation of Variance Components in Heteroscedastic One-way Random-effects ANOVA Models. *Statistics 44*, 557–569.

Rao, C.R. and S.K. Mitra (1971). *Generalized Inverse of Matrices and Its Applications.* John Wiley & Sons, New York.

Reinsel, G. (1982). Multivariate Repeated-Measurement or Growth Curve Models with Multivariate Random-Effects Structure. *J. Amer. Stat. Assoc. 77*, 190–195.

Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed-effects models. *Biometrika 92*, 351–370.

Volaufova, J. and L.R. LaMotte (2008). Comparison of approximate tests of fixed effects in linear repeated measures design models with covariates. *Tatra Mt. Math. Publ. 39*, 17-25.

Volaufova, J. and L.R. LaMotte (2012). A Simulation Comparison of Approximate Tests for Fixed Effects in Random Coefficients Growth Curve Models. *Comm. Statist. Simulation Comput.* To appear.

Volaufova, J. (2009). Heteroscedastic ANOVA: old p values, new views. *Statist. Papers 50*, 943–962.

Vonesh, E.F. and R.L. Carter (1987). Efficient Inference for Random-Coefficient Growth Curve Models with Unbalanced Data. *Biometrics 43*, 617–628.

Vonesh, E.F. and V.M. Chinchilli (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* Marcel Dekker, Inc.