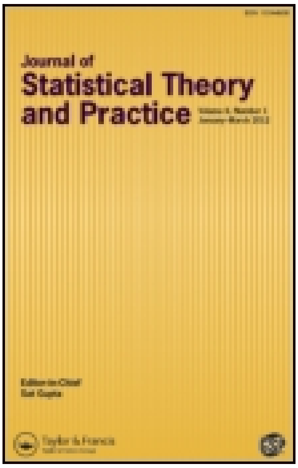

Journal of Statistical Theory and Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ujsp20

Linear Mixed Models: A Practical Guide Using Statistical Software by Brady T. West, Kathleen B. Welch, and Andrzej T. Galecki

Carlos A. Coelho [a] & Abel M. Rodrigues [b]

[a] Departamento de Matemática and Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

[b] Instituto Nacional dos Recursos Biológicos, Unidade de Silvicultura e Produtos Florestais (INRB/USPF), Oeiras, Portugal

Version of record first published: 10 Aug 2012.

To cite this article: Carlos A. Coelho & Abel M. Rodrigues (2012): Linear Mixed Models: A Practical Guide Using Statistical Software by Brady T. West, Kathleen B. Welch, and Andrzej T. Galecki, Journal of Statistical Theory and Practice, 6:3, 590-595

To link to this article: http://dx.doi.org/10.1080/15598608.2012.695708

Journal of Statistical Theory and Practice, 6:590–595, 2012

ISSN: 1559-8608 print / 1559-8616 online
DOI: 10.1080/15598608.2012.695708

Book Review

CARLOS A. COELHO[1] AND ABEL M. RODRIGUES[2]

[1]Departamento de Matemática and Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
[2]Instituto Nacional dos Recursos Biológicos, Unidade de Silvicultura e Produtos Florestais (INRB/USPF), Oeiras, Portugal

Brady T. West, Kathleen B. Welch, and Andrzej T. Gałecki, *Linear Mixed Models: A Practical Guide Using Statistical Software*, Chapmann & Hall/CRC Press (2006)

The main purpose of this book is to provide an in-depth and detailed treatment of several examples of statistical modeling of different data sets, to which linear mixed models (LMMs) may be adequately fitted. Moreover, all this is done on a quite sound theoretical basis.

The data sets are chosen in such a way that they may illustrate several typical cases with different levels of complexity (two-level clustering, three-level clustering, repeated measures, random coefficients, and clustered longitudinal data) and as such provide a good general overview of the main types of data to which LMMs may be fitted, thus providing a good overview of the main types or variants of LMMs available. Furthermore, the book also gives a very well-balanced treatment to five mainstream software packages (SAS, SPSS, R, STATA, and HLM), used to fit the LMMs used in each example.

The treatment of the examples is done in a rather thorough way, very useful for practitioners. This book is thus, beyond any doubt, highly recommended to all those who are mainly interested in learning how to fit a specific LMM to their data or willing to learn what kinds of data and for what kind of situations LMMs may be fit and adequate. It may also be a very good starting point for those willing to get a more in-depth knowledge of LMMs (see the Appendix for a very brief introduction to LMMs).

The first chapter of the book, the Introduction, adequately states the aim of the book and establishes the terminology used throughout the book regarding LMMs.

Chapter 2 is an extended summary on LMMs, which gives a rather brief but also rather sharp and quite thorough overview of LMMs in their many facets. The exposition of the material is done with an admirable brevity and logical organization, addressing issues like matrix formulation, estimation and computational issues, model building, and model selection strategies. Yet issues not commonly found in other books, like the use of several model diagnostics and the issue of marginal versus subject-specific models, may also be found in this chapter of the book.

Address correspondence to: Prof. Carlos A. Coelho, Mathematics Department and Centro de Matemática e Aplicações, Faculdade de ciencas e Tecnologia, Quinta da Torre, Caparica 2829-516, Portugal. Email: cmac@fct.unl.pt

Although this overview in Chapter 2 is, as already said, quite thorough, owing to the brevity of the exposition, the reader not only needs to be previously familiar with some of the terminology commonly used in modeling (we refer to terms like "covariate," "factor," "level," "subject," "unit of measurement," "cluster," "repeated measure," "longitudinal data") but also may find useful some previous knowledge on issues related to general linear models, variance components, or even some further details on LMMs, which may be acquired from books like the ones by Searle (1971), Searle, Casella, and McCulloch (1992), McCulloch and Searle (2001), McCulloch, Searle, and Neuhaus (2008), or Verbeke and Molenberghs (2000). Some further attention should, however, have been given to some details like the issues raised around the treatment of model matrices, which are not indeed full rank.

Then, each of the five following chapters, chapters 3 through 7, features the application of a specific LMM to a data set, highlighting the use of a different software package for the fit of the model, teaching both the basics as well as some details related to the fitting of the specific LMM model being considered. Each chapter is prefaced by an introduction to the specific LMM being applied, and while specific attention is given to one of the software packages, the same model is fitted using all five software packages.

In general there is an invigorating agreement among the estimates obtained using the different software packages for the parameters related to both the fixed and random effects. Only for the examples in chapters 6 and 7 were some convergence problems found, and these were possible to easily overcome.

Chapter 3 deals with a model for two-level clustered data. The data reports on weights of rat pups nested into litters, subjected to three doses (high, low, and control) of an experimental compound. Each litter was randomly assigned to a treatment level, being the pups nested in the corresponding litter. The fixed effects are the ones corresponding to the treatments, sex, and corresponding interactions. The random effects parameters in the model are the ones corresponding to each litter. A top-down approach was taken for model selection, starting with the complete model, where the hypotheses of nullity of the interaction between treatment and sex, nullity of the variance of the random effects associated with each litter, and homogeneity of the variances of the errors were tested. The chosen final model has all the fixed effects, without interaction and the random parameters for each litter. This final model also has a covariance matrix for the errors where the variances associated with the high- and low-dose treatments are equal but different from the variances for the control treatment.

In Chapter 4 a model for three-level clustered data is fitted to a problem that tries to analyze student performance in mathematics. Data were collected at three hierarchical levels, which are the student, nested into the class, itself nested into the school. The covariates used were covariates associated with the school neighborhood poverty level, the teacher's class mathematical preparation, students socio-economic status, scores in math for kindergarten, and sex. In this chapter the model-building strategy used was, opposite to the other four middle chapters, a step-up procedure, starting with the model with only the fixed intercept and the random parameters for school and class, nested into school. This model was then augmented with fixed effects corresponding to several of the mentioned covariates. However, as the authors say, likelihood ratio tests for some fixed effects associated with classroom covariates could not be implemented owing to the existence of missing values for those covariates.

In Chapter 5 LMMs are used for repeated-measures analysis, where multiple measurements are made on the same subject under different conditions. The data set used refers to a study on three brain regions of five rats, where the optical density of the tissue is measured for two treatments (carbachol and saline solution) administered in each of the three brain regions. Brain region and treatment were considered as fixed effects at the within-subject level. However, the results show distinct variation patterns among animals and a greater variability of response for the carbachol treatment. For this reason, besides the fixed effect for treatment in the model, a random coefficient is added at the animal or between-subject level. Also a random intercept is used at this level. The joint distribution of the random effects is assumed to be bivariate-normal. The final selected model assumes homoscedasticity for the errors, having the brain regions, the treatments, and their interactions as fixed effects and as random parameters the ones associated with the random intercepts and coefficients for the treatment with carbachol versus saline solution, for each rat.

Chapter 6 addresses the longitudinal repeated-measures case, that is, repeated measures taken on the subjects over time, using a data set relating to a study on socialization capabilities of children with autism. A previous study of the data has shown that although at the age of two years there was not much variability in the socialization capabilities of children, at the age of 13 years, there were large variations among children, concerning their socialization capabilities, as measured by the VSAE (Vineland Socialization Age Equivalent). This pattern of variability led to the consideration of random intercepts and the use of random coefficients associated to age and its square, by child. For this model there were some problems with the estimation of the variance–covariance matrix for the random parameters by child, given that some of these matrices were not positive-definite, since some of the variance components estimates were negative. This way the random intercept terms by child were excluded from subsequent models.

Finally, in Chapter 7 a study of LMMs for clustered longitudinal data is proposed. These are models in which units of analysis are nested into clusters and measurements on the units of analysis are made across time. Each cluster of units may be formed by a different number of units, and also the time instants of measurement may be different. The example presented encompasses three levels of analysis, with level 1 being the ensemble of longitudinal measurements made across time, level 2 the units of analysis (the teeth), and level 3 the clusters of units. This kind of model includes random effects associated with either or both the clusters and the units of analysis. If, as is the case in the example presented in Chapter 7, the measurements on the units of analysis are made at the same time points, then the time factor may be crossed with the random effects. The problem analyzed consists of a study on the consequences of dental veneer applications upon gingival health, with data collected by researchers at the University of Michigan Dental School. Gingival health was evaluated by measurements of GCF (gingival crevicular fluid) at two time points (3 and 6 months) for each tooth, with teeth clustered within patients. Graphical analysis showed that the GCF values for all teeth of a given patient tend to follow the same time trend. On the other hand, the temporal evolution of GCF was distinct for different patients. Furthermore, GCF values were different among individuals as well as among teeth for the same individual. This way, the model to be used should include random coefficient associated with time and intercept terms associated with patients and teeth within patients. Using a top-down approach, the interaction terms of time with other fixed effects were excluded from the final model. The authors also test several structures for the residual variance–covariance matrices. Some estimation problems, consisting in too high values for the standard errors of some variance–covariance estimates, were found in the search for an adequate structure for these matrices. This way the authors choose to use diagonal error variance–covariance matrices with possible different variances for the errors relative to the two measurements in each tooth, although after testing for equality they do not reject this hypothesis.

These chapters would, however, have benefited from a final summary with some recommendations from the authors concerning what they would find to be the most adequate software and fitting techniques for each model under analysis, or otherwise a final chapter with some authoritative views and recommendations from the authors concerning the adequacy of softwares for a specific LMM would be useful. This is surely an issue that may be taken into consideration in future editions of the book.

In the choice of coefficients, that is, factors and intercepts to be included in the models, and the choice of the ones to be considered random and fixed, in order to circumvent an almost infinite number of possible choices, the authors followed a methodology of taking a previous data analysis in order to obtain an adequate choice. Although this methodology may be seen as having advantages related with the appropriate model specification, we think that the model specification, and the choice of the effects and intercepts to be considered as random or fixed, should be dictated by the way data were collected, the knowledge we have of the problems and phenomena under study, and the objectives of the study, namely, in terms of the extrapolation scope we want to achieve.

As a final comment, we would say that taking into account the difficulties of gathering in a not-so-long book all the topics associated with the complex subject of LMMs, we consider this book a very useful manual for the application of LMMs, which will contribute beyond any doubt to the development of work in this and related areas.

References

Chi, E. M., and G. C. Reinsel. 1989. Models of longitudinal data with random effects and AR(1) errors. *J. Am. Stat. Assoc.*, 84, 452–459.

Gibbons, R. D., and D. Hedeker. 2000. Applications of mixed effects models in biostatistics. *Sankhya, Ser. B*, 62, 70–103.

Lee, Y., and J. A. Nelder. 2001. Hierarchical generalized linear models: A synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika*, 88, 987–1006.

Lee, Y., and J. A. Nelder. 2006. Double hierarchical generalized linear models. *Appl. Stat.*, 55, 139–185.

Mansour, H., E. V. Nordheim, and J. J. Rutledge. 1985. Maximum likelihood estimation of variance components in repeated measures designs assuming autoregressive errors. *Biometrics*, 41, 287–294.

McCulloch, C. E., and S. R. Searle. 2001. *Generalized, linear, and mixed models*. New York, Wiley.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus. 2008. *Generalized, linear, and mixed models*, 2nd ed. Hoboken, NJ, J. Wiley & Sons.

Searle, S. R. 1971. *Linear models*. New York, Wiley.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. *Variance components*. New York, Wiley.

Ten-Have, T. R. 1996. A mixed effects model for multivariate ordinal response data including correlated failure times with ordinal responses. *Biometrics*, 52, 473–491.

Verbeke, G., and G. Molenberghs. 2000. *Linear mixed models for longitudinal data*. New York, Springer.

Appendix: Very Brief Introduction to Linear Mixed Models

This short appendix is mostly devoted to those who, although quite familiar with linear models, may find useful a very brief introduction to linear mixed models. A linear mixed model (LMM) is a model of the form

$$\underline{Y} = X\underline{\beta} + Z\underline{\gamma} + \underline{\varepsilon} \tag{1}$$

where vectors appear underlined in bold and where we assume $\underline{\gamma}$ and $\underline{\varepsilon}$ to be uncorrelated random vectors, both with null expectation. It is thus a model where the response variable is modeled by two sets of parameters. In Eq. (1) X and Z are the design or model matrices. X is the design or model matrix for the variables associated with the fixed, that is, nonrandom effects and Z is the design or model matrix associated with the random effects; accordingly, $\underline{\beta}$ represents the nonrandom, although unknown, parameter vector associated with the variables in X and $\underline{\gamma}$ represents the random parameter vector associated with the variables in Z. In simple terms, we may say that in the long term, that is, if we would be able to repeat indefinitely our experiment under exactly the same conditions, we would expect to see much the same estimates for $\underline{\beta}$, while the estimates for $\underline{\gamma}$ would follow the distribution law assumed for $\underline{\gamma}$ itself. The distribution of the random-effects parameter vector $\underline{\gamma}$ is typically assumed to be multivariate normal, while, formally, for the simple LMMs we should assume for $\underline{\varepsilon}$ a multivariate normal distribution, and for the generalized LMMs we may assume for $\underline{\varepsilon}$ any distribution in the exponential family.

Usually, while the parameters in $\underline{\beta}$ are estimated through a minimum squares technique, the parameters in $\underline{\gamma}$ are estimated through a maximum likelihood, EM, or Newton–Raphson technique by using an iterative algorithm that alternates between a phase where $\underline{\beta}$ is taken as fixed and $\underline{\gamma}$ is estimated and another phase where $\underline{\gamma}$ is taken as fixed and $\underline{\beta}$ is estimated, until convergence is reached. LMMs, compared with the common linear models, go two steps further in that they not only allow for both nonrandom and random regression coefficients, instead of just the nonrandom, although unknown, ones in the common linear models, and they allow as well for the existence of a correlation structure in the errors, instead of just the uncorrelated errors for the common linear models. LMMs are also known as multilevel models or hierarchical linear models (see Lee and Nelder 2001; 2006) as they indeed are used to model different clusters of observations. Therefore, the model in Eq. (1) may and should be written as

$$\underset{(n_i \times 1)}{\underline{Y}_i} = \underset{(n_i \times p)}{X_i} \; \underset{(p \times 1)}{\underline{\beta}} + \underset{(n_i \times q)}{Z_i} \; \underset{(q \times 1)}{\underline{\gamma}_i} + \underset{(n_i \times 1)}{\underline{\varepsilon}_i} \quad (i = 1, \ldots, N) \tag{2}$$

for the ith cluster of n_i observations ($i \in \{1, \ldots, N\}$), where we assume the $\underline{\varepsilon}_i$ to be independent across clusters and where indeed each cluster may have one only observation, that is, where we may have $n_i = 1$ for some or all i. In Eq. (2) we assume

$$\underline{\gamma}_i \sim N_q\left(\underline{0}, \Sigma_{\underline{\gamma}}\right)$$

and

$$\underline{\varepsilon}_i \sim N_{n_i}\left(\underline{0}, \Sigma_{\underline{\varepsilon}_i}\right)$$

where $\Sigma_{\underline{\gamma}}$ is $q \times q$ and $\Sigma_{\underline{\varepsilon}_i}$ is $n_i \times n_i$ and it depends on i only through its dimension, $n_i \times n_i$, but not in terms of number of parameters. The structure for the variances and covariances for each $\underline{\varepsilon}_i$ has to be defined from structural assumptions or conventions. In the case of independent errors we will have $\Sigma_{\underline{\varepsilon}_i} = \sigma^2 I_{n_i}$, while in other cases several other different structures may be more adequate, as for example the ones that originated from AR(1), MA(1) or ARMA(1,1) models (see, e.g., Chi and Reinsel [1989] or Mansour and Nordheim [1985]). A typical assumption is, as Gibbons and Hedeker (2000) state, the assumption of stationarity, or some relaxed form of it, that is, the assumption that for repeated-measures designs the variance of the errors is constant over time points and that the covariance of the errors from different time points depends only on the time interval between these time points and not on the starting time point.

LMMs thus extend the common linear models so that not only both fixed and random effects are allowed to model the response variable, but also the error terms and the random effects are allowed to exhibit correlated and nonconstant variability, providing in this way a great flexibility to model not only the mean of the response variable but also its covariance structure.