Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 59 (2008) 390–412
Journal of Memory and Language
www.elsevier.com/locate/jml

Mixed-effects modeling with crossed random effects for subjects and items

R.H. Baayen [a,*], D.J. Davidson [b], D.M. Bates [c]

[a] University of Alberta, Edmonton, Department of Linguistics, Canada T6G 2E5
[b] Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
[c] University of Wisconsin, Madison, Department of Statistics, WI 53706-168, USA

Received 15 February 2007; revision received 13 December 2007
Available online 3 March 2008

Abstract

This paper provides an introduction to mixed-effects models for the analysis of repeated measurement data with subjects and items as crossed random effects. A worked-out example of how to use recent software for mixed-effects modeling is provided. Simulation studies illustrate the advantages offered by mixed-effects analyses compared to traditional analyses based on quasi-F tests, by-subjects analyses, combined by-subjects and by-items analyses, and random regression. Applications and possibilities across a range of domains of inquiry are discussed.

Keywords: Mixed-effects models; Crossed random effects; Quasi-F; By-item; By-subject

Psycholinguists and other cognitive psychologists use convenience samples for their experiments, often based on participants within the local university community. When analyzing the data from these experiments, participants are treated as random variables, because the interest of most studies is not about experimental effects present only in the individuals who participated in the experiment, but rather in effects present in language users everywhere, either within the language studied, or human language users in general. The differences between individuals due to genetic, developmental, environmental, social, political, or chance factors are modeled jointly by means of a participant random effect.

A similar logic applies to linguistic materials. Psycholinguists construct materials for the tasks that they employ by a variety of means, but most importantly, most materials in a single experiment do not exhaust all possible syllables, words, or sentences that could be found in a given language, and most choices of language to investigate do not exhaust the possible languages that an experimenter could investigate. In fact, two core principles of the structure of language, the arbitrary (and hence statistical) association between sound and meaning and the unbounded combination of finite lexical items, guarantee that a great many language materials must be a sample, rather than an exhaustive list. The space of possible words, and the space of possible sentences, is simply too large to be modeled by any other means. Just as we model human participants as random variables, we have to model factors characterizing their speech as random variables as well.

* Corresponding author. Fax: +1 780 4920806.
E-mail addresses: baayen@ualberta.ca (R.H. Baayen), Doug.Davidson@fcdonders.ru.nl (D.J. Davidson), bates@stat.wisc.edu (D.M. Bates).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.12.005

Clark (1973) illuminated this issue, sparked by the work of Coleman (1964), by showing how language researchers might generalize their results to the larger population of linguistic materials from which they sample by testing for statistical significance of experimental contrasts with participants and items analyses. Clark's oft-cited paper presented a technical solution to this modeling problem, based on statistical theory and computational methods available at the time (e.g., Winer, 1971). This solution involved computing a quasi-F statistic which, in the simplest-to-use form, could be approximated by the use of a combined minimum-F statistic derived from separate participants (F1) and items (F2) analyses. In the *30+* years since, statistical techniques have expanded the space of possible solutions to this problem, but these techniques have not yet been applied widely in the field of language and memory studies. The present paper discusses an alternative known as a mixed effects model approach, based on maximum likelihood methods that are now in common use in many areas of science, medicine, and engineering (see, e.g., Faraway, 2006; Fielding & Goldstein, 2006; Gilmour, Thompson, & Cullis, 1995; Goldstein, 1995; Pinheiro & Bates, 2000; Snijders & Bosker, 1999).

Software for mixed-effects models is now widely available, in specialized commercial packages such as MLwiN (MLwiN, 2007) and ASReml (Gilmour, Gogel, Cullis, Welham, & Thompson, 2002), in general commercial packages such as SAS and SPSS (the'mixed' procedures), and in the open source statistical programming environment R (Bates, 2007). West, Welch, and Gałlechki (2007) provide a guide to mixed models for five different software packages.

In this paper, we introduce a relatively recent development in computational statistics, namely, the possibility to include subjects and items as crossed, independent, random effects, as opposed to hierarchical or multilevel models in which random effects are assumed to be nested. This distinction is sometimes absent in general treatments of these models, which tend to focus on nested models. The recent textbook by West et al. (2007), for instance, does not discuss models with crossed random effects, although it clearly distinguishes between nested and crossed random effects, and advises the reader to make use of the `lmer()` function in R, the software (developed by the third author) that we introduce in the present study, for the analysis of crossed data.

Traditional approaches to random effects modeling suffer multiple drawbacks which can be eliminated by adopting mixed effect linear models. These drawbacks include (a) deficiencies in statistical power related to the problems posed by repeated observations, (b) the lack of a flexible method of dealing with missing data, (c) disparate methods for treating continuous and categorical responses, as well as (d) unprincipled methods of modeling heteroskedasticity and non-spherical error variance (for either participants or items). Methods for estimating linear mixed effect models have addressed each of these concerns, and offer a better approach than univariate ANOVA or ordinary least squares regression.

In what follows, we first introduce the concepts and formalism of mixed effects modeling.

Mixed effects model concepts and formalism

The concepts involved in a linear mixed effects model will be introduced by tracing the data analysis path of a simple example. Assume an example data set with three participants *s*1, *s*2 and *s*3 who each saw three items *w*1, *w*2, *w*3 in a priming lexical decision task under both short and long SOA conditions. The design, the RTs and their constituent fixed and random effects components are shown in Table 1.

This table is divided into three sections. The leftmost section lists subjects, items, the two levels of the SOA factor, and the reaction times for each combination of subject, item and SOA. This section represents the data available to the analyst. The remaining sections of the table list the effects of SOA and the properties of the subjects and items that underly the RTs. Of these remaining sections, the middle section lists the fixed effects: the intercept (which is the same for all observations) and the effect of SOA (a 19 ms processing advantage for the short SOA condition). The right section of the table lists the random effects in the model. The first column in this section lists by-item adjustments to the intercept, and the second column lists by-subject adjustments to the intercept. The third column lists by-subject adjustments to the effect of SOA. For instance, for the first subject the effect of a short SOA is attenuated by 11 ms. The final column lists the by-observation noise. Note that in this example we did not include by-item adjustments to SOA, even though we could have done so. In the terminology of mixed effects modeling, this data set is characterized by *random intercepts* for both subject and item, and by *by-subject random slopes* (but no by-item random slopes) for SOA.

Formally, this dataset is summarized in (1).

$$\mathbf{y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{S}_i\mathbf{s}_i + \mathbf{W}_j\mathbf{w}_j + \boldsymbol{\epsilon}_{ij} \quad (1)$$

The vector $\mathbf{y}_{ij}$ represents the responses of subject i to item j. In the present example, each of the vectors $\mathbf{y}_{ij}$ comprises two response latencies, one for the short and one for the long SOA. In (1), $\mathbf{X}_{ij}$ is the design matrix, consisting of an initial column of ones and followed by columns representing factor contrasts and covariates. For the present example, the design matrix for each subject-item combination has the simple form