crossed random e ects and items

Davidson [b], D.M. Bates [c]

Department of Linguistics, Canada T6G 2E5
cs, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
n, Department of Statistics, WI 53706-1685, USA

2007; revision received 13 December 2007

mixed-e ects models for the analysis of repeated measurement data with sub-
evelopmental, envi-
factors are mod-
t random e ect.

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 materialsmeans. Just as we model human participants as rando variables, we have to model factors characterizing their

$$X_{ij} \quad \begin{matrix} 1 & 0 \\ 1 & 1 \end{matrix} \qquad 2$$

and is the same for all subjects i and items j. The design matrix is multiplied by the vector of population coe - cients $\boldsymbol{\beta}$. Here, this vector takes the form

$\boldsymbol{\beta}$

shorter latencies, for both SOA conditions, across all subjects.

$W_j w_1$	1 1	28.3	28.3 28.3	6

The model specification

Turning to the subtable of random e ects, we observe that the first column lists the main grouping factors: `Item`, `Subj` and the observation noise (`Residual`). The second column specifies whether the random e ect concerns the intercept or a slope. The third column reports the variances, and the fourth column the square roots of these variances, i.e., the corresponding standard deviations. The sample standard deviations calculated above on the basis of Table 1 compare well with the model estimates, as shown in Table 2.

The high correlation of the intercept and slope for the subject random e ects (

lists only random intercepts for subject and item, as desired.

The reader may have noted that summaries for model objects fitted with `lmer` list standard errors and t-statistics for the fixed e ects, but no p-values. This is not without reason.

With many statistical modeling techniques we can derive exact distributions for certain statistics calculated from the data and use these distributions to perform hypothesis tests on the parameters, or to create confidence intervals or confidence regions for the values of these parameters. The general class of linear models fit

els with fixed-e ects only. Crucially, the MCMC technique applies to more general models and to data sets with arbitrary structure.

Informally, we can conceive of Markov chain Monte Carlo (

tribution, which is generally the case for such parameters. After we have checked this we can evaluate p-values from the sample with an ancillary function defined in the languageR package, which takes a fitted model as input and generates by default 10,000 samples from the posterior distribution:

We obtain p-values for only the first two parameters (the fixed e ects). The first two columns show that the model estimates and the mean estimate across MCMC

ple evaluated at zero works because the value 0 cannot occur in the MCMC

However, this counterintuitive inhibitory priming e ect is no longer significant when the decision latency at the preceding trial (RTmin1) is brought into the model,

```
> print(lmer(RT
log(RTmin1) + Condition + (1 Word) + (1 Subject)),
corr = FALSE)Woreluntecept())-9763.90.0034623n
Subjecluntecept())-9763.90.0-53773n
```

2005). The code for the simulations is available in the languageR package in the CRAN archives (http://cran.r-project.org, see

Most psycholinguistic experiments yield much larger numbers of data points than in the present example. Table 5 summarizes a second series of simulations in which we increased the number of subjects to 20 and the number of items to 40. As expected, the Type I error rate for the mixed-e ects models evaluated with tests based on p-values using the t-test are now in accordance with the nominal levels, and power is perhaps slightly larger than the power of the quasi-F test. Evaluation using MCMC sampling is conservative for this specific fully balanced example. Depending on the costs of a Type I error, the greater power of the t-test may o set its slight anti-conservatism. In our experience, the di erence between the two p-values becomes very small for data sets with thousands instead of hundreds of observations. In analyses where MCMC-based evaluation and t-based evaluation yield a very similar verdict across coefficients, exceptional disagreement, with MCMC sampling suggesting clear non-significance and the t-test suggesting significance, is a diagnostic of an unstable and suspect parameter. This is often confirmed by inspection of the parameter's posterior density.

It should be kept in mind that real life experiments are characterized by missing data. Whereas the quasi-F test is known to be vulnerable to missing data, mixed-e ects t

compared to the F1 analysis proposed by Raaijmakers et al. (1999), as illustrated in Table 7, which lists Type I error rate and power for 1000 simulation runs without and with an e ect of SOA. Simulated datasets were constructed using the parameters given by `latinsquare.lmer4`. The upper half of Table 7 shows power and Type I error rate for the situation in which the F1 analysis includes the interaction of SOA by List,

The estimates are close to the parameters that generated the simulated data: $\sigma_i = 20$, $\sigma_s = 50$, $\sigma = 80$, $\beta_{int} = 400$, $\beta_{priming} = 30$, $\beta_{list} = 18.5$, $\beta_{list:priming} = 0$.

Table 8 lists power and Type I error rate with respect to

BLUPs (the subject and item specific adjustments to intercepts and slopes), which allow enhanced prediction for these items and subjects (see, e.g., Baayen, 2008, for further discussion). Another important advantage is the possibility to include simultaneously predictors that are tied to the items (e.g., frequency, length) and predictors that are tied to participants (e.g., handedness, age, gender). Mixed-e ects models have also been extended to generalized linear models and can hence be used e - ciently to model binary response data such as accuracy in lexical decision (see Jaeger, this volume).

To conclude, we briefly address the question of the extent to which an e ect observed to be significant in a

mtly to4e[gende.7(bys521(t0and)ralized)-3e[gend)-cros21(t0alsobr)-1(t0the)ects)-2780-1.2203TD[gend-318.10and)sF-34t-334.4gen1

in an RT study, and then analyze this data in (what in the neuroimaging community is called) a random e ects analysis.

The estimation methods used to calculate the statistical parameters of these models include Maximum Likelihood or Restricted Maximum Likelihood, just as in the application of the multilevel models used in education research described earlier. One reason that these techniques are used is to account for correlation between successive measurements in the imaging time series. These corrections are similar to corrections familiar to psychologists for non-sphericity (Greenhouse & Geisser, 1958).

Similar analysis concerns are present within electrophysiology. In the past, journal policy in psychophysiological research has dealt with the problems posed by

ing their behavior as the experiment proceeds to optimize performance. Procedures requiring prior averaging across subjects or items, or procedures that are limited to strictly factorial designs, cannot provide the researcher with the analytical depth typically provided by a mixed-e ects analysis.

For data with not too small numbers of observations, mixed-e ects models may providear

mum likelihood estimates. The + symbols in each panel denote the values of the deviance components at the maximum likelihood estimates.

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modeling of data on teaching styles. Journal of the Royal Statistical Society, A, 144, 148–161.

Aitkin, M., & Longford, N. (1986). Statistical modeling in school e ectiveness studies.