Journal of Memory and Language 94 (2017) 206–234

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The cave of shadows: Addressing the human factor with generalized additive mixed models

Harald Baayen [a,*], Shravan Vasishth [b], Reinhold Kliegl [b], Douglas Bates [c]

[a] University of Tübingen, Germany
[b] University of Potsdam, Germany
[c] University of Wisconsin-Madison, USA

ARTICLE INFO

Article history:
Received 21 June 2015
revision received 8 November 2016
Available online 7 January 2017

Keywords:
Generalized additive mixed models
Within-experiment adaptation
Autocorrelation
Experimental time series
Confirmatory versus exploratory data analysis
Model selection

ABSTRACT

Generalized additive mixed models are introduced as an extension of the generalized linear mixed model which makes it possible to deal with temporal autocorrelational structure in experimental data. This autocorrelational structure is likely to be a consequence of learning, fatigue, or the ebb and flow of attention within an experiment (the 'human factor'). Unlike molecules or plots of barley, subjects in psycholinguistic experiments are intelligent beings that depend for their survival on constant adaptation to their environment, including the environment of an experiment. Three data sets illustrate that the human factor may interact with predictors of interest, both factorial and metric. We also show that, especially within the framework of the generalized additive model, in the nonlinear world, fitting maximally complex models that take every possible contingency into account is ill-advised as a modeling strategy. Alternative modeling strategies are discussed for both confirmatory and exploratory data analysis.

All models are wrong, but some are useful.

[George Box (1979)]

Introduction

Regression models are built on the assumption that the residual errors are identically and independently distributed. Mixed models make it possible to remove one source of non-independence in the errors by means of random-effect parameters. For instance, in an experiment with fast and slow subjects, the inclusion of by-participant random intercepts ensures that the fast subjects will not have residuals that will tend to be too large, and that the slow subjects will not have residuals that are too small (see, e.g. Pinheiro & Bates, 2000, for detailed examples). However, even after including random-effect parameters in a linear model, errors can still show non-independence.

For studies on memory and language, it has been known for nearly half a century that in time series of experimental trials, response variables such as reaction times elicited at time t may be correlated with earlier reaction times at $t-k$, $k \geqslant 1$ (Baayen & Milin, 2010; Broadbent, 1971; Gilden, 2001; Gilden, Thornton, & Mallon, 1995; Sanders, 1998; Taylor & Lupker, 2001; Welford, 1980). One source of temporal dependencies between trials is the presence of an autocorrelational process in the errors, potentially representing fluctuations in attention. Another source may be habituation to the experiment, possibly in interaction with decisions made at preceding trials (Masson & Kliegl, 2013). Alternatively, subjects may slow down in the course of an experiment due to fatigue. A further source of correlational

* Corresponding author at: Seminar für Sprachwissenschaft, Eberhard Karls University Tübingen, Wilhelmstrasse 19, Tübingen, Germany.
E-mail address: harald.baayen@uni-tuebingen.de (H. Baayen).

http://dx.doi.org/10.1016/j.jml.2016.11.006
0749-596X/

structure in sequences of responses is learning. As shown by Marsolek (2008), the association strengths between visual features and object names are subject to continuous updating. Ramscar, Yarlett, Dye, Denny, and Thorpe (2010) and Arnon and Ramscar (2012) documented the consequences of within-experiment learning in the domain of language. Kleinschmidt and Jaeger (2015) report and model continuous updating in auditory processing in the context of speaker-listener adaptation. De Vaan, Schreuder, and Baayen (2007) reported lexical decisions at trial t to be co-determined by the lexicality decision and the reaction time to a prime that occurred previously at $t-40$. Grammaticality judgements that change in the course of an experiment are reported by Dery and Pearson (2015). We refer to the ensemble of learning, familiarization with the task, fatigue, and attentional fluctuations as adaptive processes, or, in short, the 'human factor'. We also refer to data in which the human factor plays no role whatsoever as 'sterile' data, data that are not infected in any way by hidden processes unfolding in time series of experimental trials.

Why might we expect that experimental data are not sterile? Because, unlike molecules or plots of barley, human beings adapt quickly and continuously to their environment, and as the work mentioned above has shown, this includes the environment of psycholinguistic experiments.

When temporal autocorrelations are actually present in the data, but not brought into the statistical model, the residuals of this model will be autocorrelated in experimental time. The proper evaluation of model components by means of t or F tests presupposes that residual errors are identically and independently distributed. By bringing random intercepts and random slopes into the model specification, clustering in the residuals by item or subject is avoided. However, such random slopes and random intercepts do not take care of potential trial-to-trial autocorrelative structure. The presence of autocorrelation in the residuals leads to imprecision in model evaluations and uncertainty about the validity of any significances reported. When strong autocorrelation characterizes the residuals, this uncertainty will make it impossible to draw well-founded conclusions about statistical significance.

It might be argued that adaptive processes, if present, will have effects that are so minute that they are effectively undetectable. If so, the experimental design, and only the experimental design, could serve as a guide for determining the statistical model to be fitted to the data. Alternatively, one might acknowledge the presence of adaptive processes but claim that their presence gives rise to random and temporally uncorrelated noise. Any such adaptive processes would therefore be expected not to interact with predictors of theoretical interest.

However, it is conceivable that adaptive processes are present in a way that is actually not harmless. We distinguish two cases. First, adaptive processes may be present, without interacting with critical predictors of theoretical interest. In this case, measures for dealing with the autocorrelation in the errors will be required, without however affecting the interpretation of the predictors. In this case, elimination of autocorrelation from the errors will result in p-values that are more trustworthy. Second, it is in principle possible that adaptive processes actually do interact with predictors of theoretical interest in non-trivial ways. If so, it is not only a potential autocorrelational process in the residual error that needs to be addressed, but also and specifically the adaptive processes. These processes, which themselves may constitute a considerable source of autocorrelation in the errors, will need to be examined carefully in order to provide a proper assessment of how they modulate the effects of the critical predictors.

In this study, we discuss three examples of non-sterile data demonstrably infected by adaptive processes unfolding in the experimental time series constituted by the successive experimental trials. First, we re-analyze a data set with multiple subjects, and a $2 \times 2 \times 4$ factorial design with true treatments (Kliegl, Kuschela, & Laubrock, 2015) and a single stimulus 'item'. We then consider a mega-study with auditory lexical decision (Ernestus & Cutler, 2015) using a regression design with crossed random effects of subject and item. The third analysis concerns a self-paced reading study in which subjects were reading Dutch poems, following up on earlier analyses presented in Baayen and Milin (2010).

The analyses of these three data sets make use of the generalized additive mixed model (GAMM). Before presenting these analyses, we first provide an introduction to GAMMs. Following the analyses of the three data sets, we discuss regression modeling strategies for dealing with the human factor when conducting confirmatory or exploratory data analysis, and the final discussion section, after summarizing the main results, closes with some reflections on the importance of parsimony in regression modeling.

The generalized additive mixed model

In linear regression, a univariate response y_i (where i indexes the individual data points) is modeled as the sum of a linear predictor η_i and a random error term with zero mean. This linear predictor is assumed to depend on a set of predictor variables. Often, the response variable is assumed to have a normal distribution. If so, a regression model such as

$$y_i = \eta_i + \epsilon_i \text{ where } \epsilon_i \underset{\text{ind}}{\sim} N(0, \sigma^2) \text{ and } \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

describes a response variable y that is modeled as a weighted sum of two predictors, x_1 and x_2, together with an intercept (β_0) and Gaussian error with standard deviation σ.

Generalized linear models let the response depend on a smooth monotonic function of the linear predictor. This family of models allows the response to follow not only the normal distribution, but other distributions from the exponential family, such as Poisson, gamma, or binomial. An example of a binomial GLM with the same linear predictor η is

$$y_i \underset{\text{ind}}{\sim} \text{binom}(\exp(\eta_i)/\{1+\exp(\eta_i)\}, 1) \quad \text{where } \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

This equation specifies that y_i follows a binomial distribution with 'number of trials' = 1, and a probability of success