Mixed-effects regression modeling

Roland Schäfer

Freie Universität Berlin

August 7, 2017

1 Introduction

2 Fundamentals

2.1 Introduction to random effects

Generalized Linear Models (GLMs; as discussed in the previous chapter) allow us to estimate the effects which various *predictors* or *regressors* (i. e., corpus linguistic variables) have on an *outcome* or *response* (i. e., another corpus linguistic variable).[1] Surely, the most typical application (in corpus linguistics) is the modeling of *alternations*, i. e., phenomena where the response variable of interest encodes a choice of forms or constructions, for example a case alternation (binary or n-valued categorical, depending on the richness of the language's case system), alternations of graphemic form such as contracted vs. non-contracted, ordering preferences such as the order of prenominal adjectives, or syntactic constructional alternations such as the dative alternation.[2] The approach is called *generalized* in contrast to normal linear models because the response need not be numerical, and the *errors* or *residuals* do not have to be (approximately) normally distributed. This is achieved by allowing for different types of exponential distributions for the residuals – which requires the use of a more general estimator than least-squares, typically likelihood maximization – and *link functions* that relate the additive linear term that combines

[1] It should be noted that I use the term *variable* here in its neutral sense which is common in most strains of empirical and statistical research. The definition of a variable in Labovian variationist linguistics is more restricted and theoretically burdened.

[2] In this article, I restrict the discussion to GL(M)Ms with categorical responses, simply because the continuous responses in Linear (Mixed) Models – or LM(M)s – are not often found in corpus linguistics. Also, an L(M)M can be understood as a GL(M)M with an identity link function and a Gaussian distribution for residuals.

the predictors in a non-linear way to the response variable. Generalized Linear Mixed Models (GLMMs) are not much different. They add what are often called *random effects*, *mixing* them with the normal predictors which are used in GLMs, and which are called *fixed effects* in this terminology. Alternatively, statisticians speak of *multilevel models* or *hierarchical models* (Gelman & Hill 2006), a terminology to be explained in Section 2.

The purpose of including random effects is usually said to be the modeling of variance between groups of observations. A single observation (or *data point* or *measurement*) is one atomic exemplar entering into the statistical analysis of the study. In corpus linguistics, these are usually represented by a single line of a concordance containing for example (and with reference to the above examples) a clause or sentence in which one of the alternants of a case alternation occurs, an NP where two pre-nominal adjectives are used, a single occurrence of a contracted or non-contracted form, etc. When such observations are grouped, it is often plausible to assume that some variance in the choice of the alternating forms or constructions occurs at the group-level and not at the level of observations. Groups could be defined any linguistically relevant grouping factor, such as authors, authors' sex and gender, regions where authors were born or live, social groups with which authors identify themselves, time periods, genres, styles, etc. If the concordance in a study contains, say, ten sentences each written by ten authors, the grouping factor *Author* has ten levels and defines ten groups. We know that preferences vary between speakers, and it is therefore reasonable to include this variance in our modeling in some way. The same goes for the other possible groups just mentioned. In experimental settings, variation between participants is standardly modeled by including a random effect for speaker.

While random effects are often presented like this using conceptual arguments, the crucial question in specifying models is not whether to include these grouping factors at all, but rather whether to include them as fixed effects or as random effects.[3] Contrary to formulaic recommendations such as "Always include random effects for speaker and genre.", the choice between fixed and random effects can and should be made based on an analysis and understanding of the data set at hand and the differences and similarities in the resulting estimates. After some further clarifications, this is the subject of Section 2.4.

[3] I assume that no corpus linguist would make the decision not to include relevant factors in their models when the corpus contains the corresponding meta data or these meta data can be annotated reliably with acceptable effort.

2.2 Crossed an nested effects

2.3 Hierarchical models

2.4 The details

3 Estimation for hierarchical models in R

3.1 The standard lme4 package

3.2 Improving lme4 inferences and model selection using the bootstrap

3.3 Using a more robust “Bayesian” estimator

References

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.