Mixed-effects regression modeling

Roland Schäfer

Freie Universität Berlin

August 17, 2017

1 Introduction

Mixed effects modeling – alternatively called *hierarchical* or *multilevel modeling* is a straightforward extension of (generalized) linear modeling as discussed in the previous chapter. A common characterization of mixed-effects modeling is that it accounts for situations where observations are *clustered* or *come in groups.* In corpus linguistics, there could be clusters of observations defined by individual speakers, registers, genres, modes, lemmas, etc. Instead of estimating a coefficients for each level of such a grouping factor (a so-called *fixed effects*), in a mixed model they can be modeled as normally distributed random variable (a so-called *random effect*) with predictions being made for each group. This chapter introduces the situations where mixed-effects modeling is useful, including a discussion of the alternative models without random effects. The proper specification of models without and with R are discussed, as well as some model diagnostics and ways of interpreting the output. Readers are assumed to be familiar with the concepts covered in the previous chapter.

2 Fundamentals

2.1 When are random effects useful?

2.1.1 Introduction to random effects

(Generalized) Linear Mixed Models (GLMMs) are an extension of (Generalized) Linear Models. They add what are often called *random effects* and *mix* them with the normal predictors (*fixed effects*) as used in GLMs. Alternatively, statisticians speak of *multilevel models* or *hierarchical models* (Gelman & Hill 2006), a terminology to be explained in Section

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* or *unit*) is one atomic exemplar entering into the statistical analysis of the study. In corpus linguistics, single observations can be understood as a single line in a concordance, and they typically represent, 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. If this is the case and the grouping factor is not included in the model, the error terms within the groups will be correlated. Since the estimator works under the assumption of non-correlated errors, standard errors for model coefficients can be estimated as smaller than they nominally are, leading to increased Type I error rates in inferences about the coefficients. This gets even worse when there are within-group tendencies regarding the direction and strength of the influence of the