Mixed-effects regression modeling

Roland Schäfer

Freie Universität Berlin

August 8, 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 (a binary or n-valued categorical response, 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. First, 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. Second, *link functions* are introduced which relate the additive linear term that combines the predictors in a non-linear way to the response

[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.

variable. Generalized Linear Mixed Models (GLMMs) are not much different. They add what are often called *random effects* and *mix* them with the normal predictors as used in GLMs. The latter one 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 2.3 and Section 3.

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, 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. Groups can be defined by any linguistically relevant grouping factor (a categorical variable, also called a nominal variable), such as the individual authors or speakers, their sex and gender, the regions where they were born or live, social groups with which they identify, but also time periods, genres, styles, etc. If the concordance in a study contains, say, ten sentences each written by ten authors, then the author grouping factor has ten levels and defines ten groups.[3] We know that preferences vary between speakers, and it is therefore reasonable to model this variance in our statistics in some way. The same goes for the other possible groups just mentioned. Furthermore, it is known that specific lexemes often have idiosyncratic affinities towards alternants in alternating constructions. Therefore, exemplars containing specific lexemes can also be treated as groups with considerable between-group variance. As an example from outside corpus linguistics, variation between participants is standardly modeled by including a random effect for speaker in experimental settings.

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.[4] Random effects structures are very suitable for accounting for group-level variation in regression, but 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

[3] Importantly, it will be advised below that grouping factors with as few as ten levels should *not* standardly be used as random effects, cf. Section 2.5.

[4] 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.

Sentence	Author	Region
1	Bay	Tyneside
2	Bay	Tyneside
3	Riley	Tyneside
4	Riley	Tyneside
5	Dale	Greater London
6	Dale	Greater London
7	Hayden	Greater London
8	Hayden	Greater London

Table 1: Illustration of nested factors

analysis and understanding of the data set at hand and the differences and similarities in the resulting estimates. Subsection 2.2, 2.3, and 2.4 introduce three important points to consider about the structure of the data typically used in mixed modeling. This is intended to show readers that mixed or multilevel/ hierarchical modeling is simply a matter of doing justice to the structure of the data. Then, Sections 2.5 and 2.6 provide a mostly non-technical introduction to the important technicalities in mixed modeling, including a discussion of when a factor should be included as a random effect and when as a fixed effect. Section 3 then shows how mixed models are specified and interpreted using R.

2.2 Crossed an nested effects

It was established in the previous section that random effects are a means of accounting for group-level variance in regression models. This section briefly introduces a distinction that plays a role in modeling when there is more than one grouping factor (to be used either as a fixed or random effect). When this is the case, each pair of grouping factors can be *nested* or *crossed*. By way of example, we can group sentences by the individual authors who wrote them, and we can group authors by their region of birth. Such a data set would intrinsically be *nested*, as Table 2 illustrates.

Since authors have a unique region of birth, Tyneside is the unique region value for the authors Bay and Riley, and Greater London is the unique region value for Dale and Hayden. There cannot be sentences where, for example, the author is Bay and the region is Greater London (assuming that authors are uniquely identified by the labels in the middle column). In this example, the region factor nests the author factor.[5] This example was chosen because the nesting is conceptually necessary. However, even when a data set has a nested

[5]The sentence index is not a grouping factor because its values are unique, and sentences are thus not "nested". They represent the basic level of observations.

Sentence	Author	Mode
1	Bay	Spoken
2	Bay	Written
3	Riley	Spoken
4	Riley	Spoken
5	Dale	Written
6	Dale	Written
7	Hayden	Spoken
8	Hayden	Written

Table 2: Illustration of nested factors

structure by accident, standard packages in R will treat them as nested, and a closer look at data sets should be part of any protocol for using GLMMs in corpus studies (see Section 3.1).

When the grouped entities do not uniquely belong to grouping factors, the factors are *crossed*. Continuing the example, crossed factors for author and mode are illustrated in Table 2. While there are only spoken sentences by Riley and only written sentences by Dale in the sample, there is one spoken and one written sentence each by Bay and Hayden. There is a many-to-many relation between authors and modes, which is characteristic of crossed factors. In Table 1, the relation between authors and regions was many-to-one, which is typical of nested factors. In experimental settings, the design often makes sure that the combinations of nested or crossed factors are represented by equal numbers of observations (such that, for example, there is an equal number of written and spoken sentences from each author). Contrarily, the situation in Table 2 is typical of corpus studies where pseudo-random sampling from a pre-compiled corpus such as the BNC was used. This does not affect the practical modeling procedures much, especially when random factors are used, as will be shown below. However, practitioners must be aware of it when interpreting the data.

Finally, it should be noted that grouping factors can form hierarchical structures. Especially when they are nested, there can be more than one level of nesting. Mode could nest genre if genres are defined such that each genre is either exclusively spoken or written, and in a given corpus, authors/speakers might be nested within genres because each author/speaker only contributed material to one genre.[6] Similarly, we might want to describe – in a given study on adjectives – adjectives as being either intersective or non-intersective.

[6]In this example, the second level of nesting is not a conceptual necessity. In fact, it would be quite surprising if the real world were shaped like this. However, standard corpus compilation techniques might easily lead to a situation where exactly this is the case, simply because it is often difficult to sample texts and utterances from single authors/speakers across a wide range of genres.

Within the two groups, a finer-grained semantic classification might be nested, which itself nests single adjective lexemes. This gives rise to potentially complex hierarchical structures, which can often be modeled more effectively using random effect structures compared to fixed effect structures.

2.3 Hierarchical or multilevel modeling

This section introduces the idea – often ignored in introductory texts written by practitioners and handbook articles – that so-called random effects actually introduce new levels of modeling, or *secondary models*. It is argued that this is, again, not a technical thing but required by the structure of certain data sets. Assume that we wanted to account for lexeme-specific variation in a study on an alternation phenomenon by specifying the lexeme as a random effect in the model. Additionally, we suspect or know that a lexeme's overall frequency influences its preferences for occurring in the construction alternants. Now, we could simply quantize the frequency variable and turn it into an ordinal variable (for example in the form of frequency bands) and interpret it as a grouping factor which nests the lexeme grouping factor. However, frequency obviously is a numerical and not a categorical variable, and by using it as a grouping factor we would destroy valuable information that is encoded in the data. A similar situation would arise in a study of learner corpus data with a learner grouping factor if we also knew that the number of years learners have learned a language influences their performance with regard to a specific phenomenon. It should have become clear that in such cases, variables like *frequency* and *number of learning years* are constant for each level of the grouping factor (*lexeme* and *learner*, respectively), but we cannot treat them as nesting grouping factors themselves. In other words, each lexeme has exactly one overall frequency, and each learner has had a fixed number of years of learning the language.[7]

Such variables are thus reasonably interpretable only at the group-level. I such cases, an adequate multilevel model uses them to partially predict the tendency of the grouping factor. Put differently, the idiosyncratic effect associated with a lexeme, speaker, genre, etc. is split up into a truly idiosyncratic preference and a preference predictable from group-level factors. This is achieved, in fact, by specifying a linear model that predicts the group-level random effect itself. Such second-level models can even contain modeled effects themselves, giving rise to third-level models, and so on.

As in the case of nested vs. crossed factors, standard packages in R often take care of hierarchical modeling automatically, given that the data are structured and are specified accordingly (see Section 3). This might, however, lead to sit-

[7] In the given example, things would get more complicated if the corpus contained data by single learners from different points in time. We simplify the scenario for the sake of an easier-to-follow introduction.

uations where practitioners specify multilevel models without even knowing it, which in turn can lead to misinterpretations of the results. Therefore, multilevel modeling will be introduced as the more general framework for so-called mixed effects models in Sections 2.5 and 3.

2.4 Random intercepts and slopes

2.5 Model specification

2.6 Pooling and shrinkage

3 Estimation of hierarchical models in R

3.1 Using lme4

3.2 Bootstrap methods for lme4

3.3 Estimation using Markov-Chain Monte Carlo

References

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