Mixed-effects regression modeling

Roland Schäfer

Freie Universität Berlin

August 9, 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 multi-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 found very often 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 speakers (or authors, writers, etc.), 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 exemplars each written by ten speakers, then the speaker 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.

Exemplar	Speaker	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 exemplars (such as sentences) by the individual speakers who wrote or uttered them, and we can group speakers by their region of birth. Such a data set would intrinsically be *nested*, as Table 2 illustrates.
Since speakers have a unique region of birth, Tyneside is the unique region value for the speakers Bay and Riley, and Greater London is the unique region value for Dale and Hayden. There cannot be exemplars where, for example, the speaker is Bay and the region is Greater London (assuming that speakers are uniquely identified by the labels in the middle column). In this example, the region factor nests the speaker factor.[5] This example was chosen because

[5]The exemplar index should not be called not a grouping factor because its values are unique, and sentences are thus not "nested". They represent the basic level of observations, and at that level, there is nothing to group.

Exemplar	Speaker	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

the nesting is conceptually necessary. However, even when a data set has a nested 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 speaker 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 speakers and modes, which is characteristic of crossed factors. In Table 1, the relation between speakers 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 speaker). 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. When grouping factors 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, speakers might be nested within genres because each of them only contributed material to one genre.[6] Similarly, we might want to describe – in a given study on adjectives

[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 speakers across a wide range of genres.

– adjectives as being either intersective or non-intersective. 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. Table 3 illustrates such a data set (fictional in this case). It might be a small fraction of the data used to predict whether a dative NP is used in the dative shift construction or not. The exemplar indices, again, simply identify single sentences containing one of the constructions of interest. The discourse status status obviously varies at the level observations, and so does the NP length in syllables. To capture verb lemma specific tendencies, a verb lemma grouping factor is added, and the verb lemma frequency necessarily varies at the group level because each lemma has a unique frequency. I such cases, an adequately specified multilevel model uses the group-level variables to partially predict

[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.

Level of observations			Group level	
Exemplar	**Givenness**	**NP length**	**Verb**	**Verb frequency**
1	New	8	give	6.99
2	Old	7	give	6.99
3	Old	5	give	6.99
4	Old	5	grant	5.97
5	New	9	grant	5.97
6	Old	6	grant	5.97
7	New	11	promise	5.86
8	New	10	promise	5.86
9	Old	9	promise	5.86

Table 3: Illustration of a data set which requires multilevel modeling; lemma frequencies are logarithmized frequencies per one million tokens taken from ENCOW14A

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 another model (a linear model) that predicts the group-level random effect itself, and the second-level predictor is a fixed effect in this model. Such second-level models can even contain modeled effects themselves, giving rise to third-level models, and so on. In a way, the data look similar to multilevel nesting, but (1) second-level models can account for continuous numerical predictors at the group-level, which nesting cannot, and (2) there might be situations where specifying even categorical second-level grouping factors as fixed effects in a second-level model is more appropriate than adding nested random effects (see Section 2.5).

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 situations 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 slopes as interactions

Before moving on to the more technical discussion of hierarchical model specification in Section 2.5, one more basic concept will be discussed in this section, namely the data patterns that gives rise to *varying intercepts* and *varying slopes*. Varying intercepts are an adequate modeling tool when the overall tendency

in the outcome variable changes with the levels of the grouping factor. It is shown that random slopes are just another way of modeling an interaction between influencing factors.

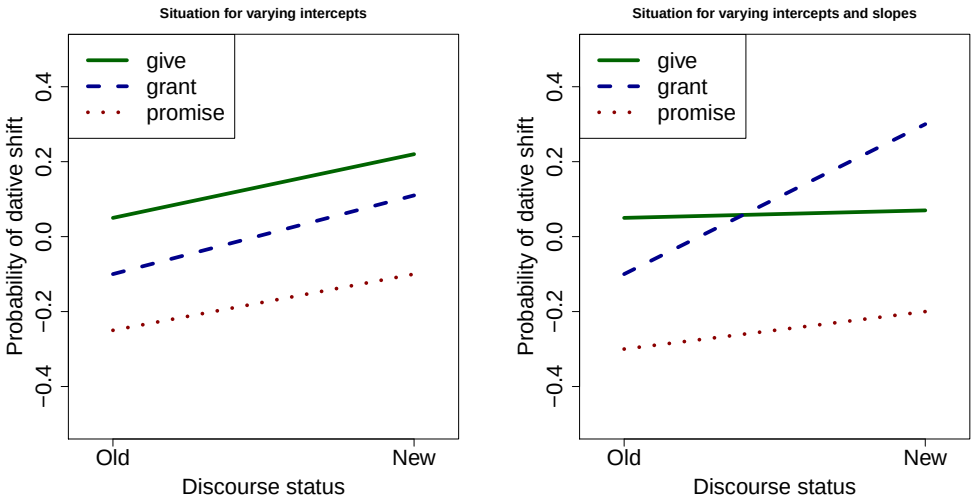

Figure 1: Illustration of data in situations for varying intercepts or varying intercepts and additional varying slopes

For the sake of the argument, we assume that we are looking at an alternation phenomenon like the dative alternation, wherein we are interested in the probability that, under given circumstances, the dative shift construction is chosen. Looking at data set, it turns out that the probability of the dative shift changes for *old* and *new* dative NPs. The verb lemma – a typical candidate to be used as a random effect – also influences the probability of either variant being used. The situation can now be as in the left or the right panel of Figure 1. In the situation illustrated in the left panel, the overall level in probability changes with the verb lemma, but for each verb lemma, the values change roughly accordingly in exemplars with old and new dative NPs. Note that the lines are not perfectly parallel because the figure is supposed to be an illustration of a data set rather than a fitted model, and we always expect some chance variation in data sets. In the situation depicted in the right panel, however, the overall levels are different between lemmas, but the lemma-specific tendencies also vary between exemplars with old and new NPs. This is actually nothing but an interaction between two factors (verb lemma and discourse status). However, if the verb lemma factor is used as a random effect grouping factor, the interaction is modeled as a so-called *random slope*. In the next section, it is shown how all the different types of data sets discussed so far can be modeled using fixed effects models or, alternatively, using mixed effects. Which one is more appropriate will be argued to be better understood as a technical rather than a conceptual question.

2.5 Model specification

In this section, it is discussed how the specification of mixed models differs from that of fixed effects models, and that for each model with random effects there is an alternative using only fixed effects. A major focus is on the question of when to use fixed and random effects. The amount of technicality and notation is kept at the absolute minimum, but a few notational conventions are introduced as the absolute minimum required to understand the literature on mixed models. To make successful *practical* use of mixed models, some level of fundamental understanding is required.

Readers with experience in fixed effects modeling (see the previous chapter in this handbook) should see that a grouping factor encoding the verb lemma, the speaker, the mode, the genre corresponding to a corpus exemplar (and all the other potential random effects grouping factors discussed in the previous sections) can be specified as a normal fixed effect in a GLM. In such a case, each of the m levels of the speaker factor is dummy-coded, and for all but one of these binary dummy variables, a coefficient is estimated. A logistic regression example is used throughout this section. We begin with a minimal model with only the dummies of the lemma grouping factor and one other predictor, namely discourse status. There are m verb lemmas (i. e., groups) and n observations. As index variables, we use j for groups and i for observations. A specification of such a model is given in equation (1).

$$Pr(y^i = 1) = logit^{-1}(\alpha_0 + \beta_d \cdot x_d^i + \beta_{l_1} \cdot x_{l_1}^i + \beta_{l_2} \cdot x_{l_2}^i + \cdots + \beta_{l_{m-1}} \cdot x_{l_{m-1}}^i) \quad (1)$$

This model provides an estimate of the probability (Pr) that in observation i, the outcome variable y^i is 1, i. e., that dative shift occurs. α_0 is the intercept, β_d is the coefficient for the effect of discourse status. x_d^i is the value of the variable that encodes discourse status for exemplar i with the value 0 for discourse-old NPs and 1 for discourse-new NPs. Furthermore β_{l_j} are the coefficients for the lemma dummy variables. Finally, $x_{l_j}^i$ is the value (0 or 1) for lemma j and observation i. If in exemplar 64, the lemma is *give* and *give* is encoded as group 12, then $i = 64$, $j = 12$, and $x_{l_{12}}^{64} = 1$, whereas all $x_{l_j}^{64} = 0$ with $j \neq 12$.[8]

Because one verb lemma dummy variable is on the intercept α_0 and thus used as a reference, we only estimate $m-1$ instead of m coefficients, i. e., $j = 1, \cdots, m-1$.[9]

The function $logit^{-1}$ is the *link function*, and its argument is the *linear term* of

[8] It is unfortunate that multiple indexation is required to such an extent. However, any alternative notation has at least the same potential to confuse readers.

[9] It is often not explained in text books for practitioners that using one dummy as a reference level is necessary because otherwise infinitely many equivalent estimates of the model coefficients exist because one could simply add a constant to the intercept and subtract it from the dummies. However, the estimator works under the assumption that there is a unique maximum likelihood estimate for the coefficient matrix.

the model. It is obvious that in such a model, the effect of each verb lemma is treated as a fixed population parameter, exactly the same as the effect of discourse status. The coefficient β_d is estimated in exactly the same way as each β_{l_j}.
If we treat the same grouping factor as a so-called random intercept, we add an atomic term to the linear term and give it a distribution instead of estimating $m - 1$ coefficients. The model looks like in equation (2).

$$P(y^i = 1) = logit^{-1}(\alpha_0 + \alpha_l^{j[i]} + \beta_d \cdot x_d^i) \tag{2}$$

We now have an overall intercept α_0, an atomic term α_l^j that varies between groups (instead of a single additive term for each of the lemmas), where in $\alpha_l^{j[i]}$ the correct lemma intercept j is chosen for observation i, which we write $j[i]$, a notation borrowed from Gelman & Hill (2006). If, again, in exemplar 64, the lemma is *give*, which is group 12, then $i = 64$ and $\alpha_l^{j[12]} = \alpha_l^{12}$ because $j[64] = 12$. The term $\beta_d \cdot x_d^i$ for the effect of discourse status remains unchanged. Crucially, instead of estimating coefficients for the lemma effect – as in the model in (1) –, α_l is itself modeled, and random terms are *predicted* for each level of the grouping factor. For this, the assumption in (3) is made.

$$\alpha_l^j \sim \mathbf{N}(0, \sigma_l^2) \tag{3}$$

This is to be read as "the values of α_l^j follow a normal distribution with mean 0 and a variance of σ_l^2". Equation 3 already specifies a very simple second-level linear model. We can see this if we reformulate it as 4.

$$\begin{aligned} \alpha_l^j &= z_j + \epsilon_j \\ \epsilon_j &\sim \mathbf{N}(0, \sigma_l^2) \end{aligned} \tag{4}$$

References

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