DRAFT of April 13, 2018

Mixed-effects regression modeling

to appear in: St. Gries & M. Paquot (eds.), Practical Handbook of Corpus Linguistics

Roland Schäfer

Freie Universität Berlin

April 13, 2018

1 Introduction

Mixed effects modeling – alternatively called *hierarchical* or *multilevel modeling* is a straightforward extension of (generalised) linear modeling as discussed in the previous chapter. A common characterisation 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 coefficients for each level of such a grouping factor (so-called *fixed effects*), in a mixed model they can alternatively be modeled as a normally distributed random variables (so-called *random effects*) with predictions of group-wise tendencies being made for each group. This chapter introduces readers to the situations where mixed-effects modeling is useful or necessary. The proper specification of models is 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

(Generalised) Linear Mixed Models (GLMMs) are an extension of (Generalised) Linear Models (GLMs). 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 2.1.3. 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* or *unit*) is one atomic exemplar entering into the statistical analysis of a study. In corpus linguistics, single observations can be understood as single lines in a concordance. These concordance lines could contain, for example, clauses or sentences in which one of the alternants of a morpho-syntactic alternation occurs, the goal being to model the influence of diverse properties of the clauses sentences on the choice of the alternants. Along similar lines and under a similar research question, they could contain occurrences of a contracted or a non-contracted form of words (like *am* and *'m* in English). As a another example, the concordance lines could contain NPs where two pre-nominal adjectives are used, the goal being to determine the factors influencing their relative ordering. When such observations are grouped, it is often plausible to assume that there is some variance in the choice of the alternating forms or constructions at the group-level. If this is the case and the grouping factor is not included in the model, the error terms within the groups will be correlated. Put simply, this means that means of the group-wise errors vary. Since the estimators used for estimating the parameters of GLMs work under the assumption of non-correlated errors, standard errors for model coefficients will typically be estimated as smaller than they nominally are, leading to increased Type I error rates in inferences about the coefficients.[1] This gets even worse when there are within-group tendencies regarding the direction and strength of the influence of the other regressors, i. e., when there is an interaction between them and the grouping factor (e. g., Schielzeth & Forstmeier 2009). This is why known variation by group should always be accounted for in the model. Random effects are often a convenient way to do so.

Groups can be defined by any linguistically relevant grouping factor, such as the individual speakers (or authors, writers, etc.), the regions where they were born or live, social groups with which they identify, but also time time periods, genres, styles, etc.[2] Specific lexemes often have idiosyncratic affinities towards alternants in alternating constructions. Therefore, exemplars containing specific lexemes also constitute groups. In cases like the dative alternation individual verbs co-occur with the alternants to different degrees.

The crucial question in specifying models is not whether to include these

[1]Type I errors occur when the null hypothesis is rejected although it is true. The requirement that error should be uncorrelated is often called "independence (of errors)".

[2]Trivially, grouping factors should never be ordinal variables. The are always categorical.

Exemplar	Speaker	Region
1	Daryl	Tyneside
2	Daryl	Tyneside
3	Riley	Tyneside
4	Riley	Tyneside
5	Dale	Greater London
6	Dale	Greater London
7	Reed	Greater London
8	Reed	Greater London

Table 1: Illustration of nested factors

grouping factors at all, but rather whether to include them as fixed effects or as random effects. Random effects structures are very suitable for accounting for group-level variation in regression, but while formulaic recommendations such as "Always include random effects for speaker and genre!" provide useful guidance for beginners, 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 models. The remainder of Section 2.1 introduces three important points to consider about the structure of the data typically used in mixed modeling. Then, Section 2.2 provides a moderately technical introduction to modeling. Section 3.1 shows how mixed models are specified using the `lme4` package in `R`, and Section 3.2 deals with the interpretation of the output.

2.1.2 Crossed and nested effects

This section discusses a distinction that arises when there is more than one grouping factor. 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

Exemplar	**Speaker**	**Mode**
1	Daryl	Spoken
2	Daryl	Written
3	Riley	Spoken
4	Riley	Spoken
5	Dale	Written
6	Dale	Written
7	Reed	Spoken
8	Reed	Written

Table 2: Illustration of crossed factors

nested, as Table 2 illustrates. Since speakers have a unique region of birth, Tyneside is the unique *region* value for the speakers Daryl and Riley, and Greater London is the unique *region* value for Dale and Reed. In this example, the region factor nests the speaker factor. This example was chosen because the nesting is conceptually necessary. However, even when a data set has a nested structure by accident, standard packages in `R` will also treat them as nested (see Section 3.1).

When the grouped entities (themselves groups) do not uniquely belong to levels of the grouping factor, 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 Daryl and Reed. 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 is many-to-one, which is typical of nested factors.

With more than two grouping factors, 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. Similarly, in a study on adjectives we might want to describe 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. However, not all of these structures should be modeled as nested random effects. In the latter case, for example, the low number of levels in one factor (intersectivity with just two levels) predestines it as a second-level predictor rather than a nesting factor; see Section 2.1.3.

2.1.3 Hierarchical or multilevel modeling

This section describes the types of data to be used in true multilevel models. Let us assume that we wanted to account for lexeme-specific variation in a study on an alternation phenomenon such as the dative alternation in English 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. A similar situation would arise in a study of learner corpus data (even of the same alternation phenomenon) 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. In such cases, variables like *frequency* and *number of learning years* are constant for each level of the grouping factor (*lexeme* and *learner*, respectively). In other words, each lexeme has exactly one overall frequency, and each learner has had a fixed number of years of learning the language.[3]

Such variables are thus reasonably interpretable only at the group-level.

[3] In the given example, things would get more complicated if the corpus contained observations of single learners at different points in time. We simplify the scenario for the sake of an easier-to-follow introduction. See also the last subsection of Section 2.2.4.

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

Table 3: Illustration of a fictional data set which requires multilevel modeling; NP length could be measured in words; the lemma frequencies are actual logarithm-transformed frequencies per one million tokens taken from ENCOW14A (Schäfer & Bildhauer 2012); the outcome column encodes whether alternant 1 or 2 was chosen

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 ditransitive verb is used in the dative shift construction or not. The givenness and the NP length status vary at the level observations. To capture verb lemma specific tendencies, a verb lemma grouping factor is added. The verb lemma frequency necessarily varies at the group level because each lemma has a unique frequency. In such cases, an adequately specified multilevel model uses the group-level variables 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 variables. This is achieved by specifying a second (linear) model which 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. 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 2.2).

As in the case of nested vs. crossed factors, standard packages in R usually take care of hierarchical modeling automatically, given that the data are structured and are specified accordingly. 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. See Section 3.1 for details.

2.1.4 Random slopes as interactions

This section introduces 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.

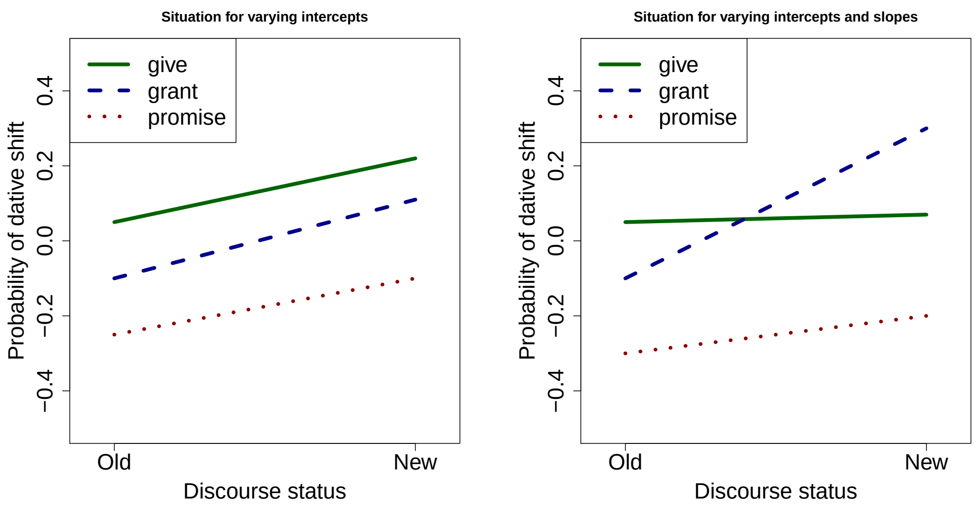

Figure 1: Illustration of fictional data in situations for varying intercepts or varying intercepts and additional varying slopes

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. In the examination of the data, it turns out that the probability of the dative shift changes for *old* and *new* dative NPs. The verb lemma 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 depicted 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 givenness), and we could use a fixed-effect interaction to take it into account. 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 models. Which one is more appropriate will be argued to be better understood as a technical rather than a conceptual question.

2.2 Model specification and modeling assumptions

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 models with 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. Prominently, the specification of models in mathematical notation is not shown here, and model specification is introduced via R notation. For an appropriate understanding of model specification, readers should urgently consult a more in-depth text book, for example Part 2A of Gelman & Hill (2006) (pp. 235–342). Without any knowledge of the mathematical notation conventions, it is impossible to understand many advanced text books and much of the valuable and in-depth advice available online.

2.2.1 Simple random intercepts

Readers with experience in fixed effects modeling should be able to see that a grouping factor encoding the verb lemma and all the other grouping factors discussed in the previous sections could be specified as normal fixed effects in a GLM. This section introduces the main difference between the fixed-effect approach and the random-effect approach. Logistic regression examples are used throughout this section, and we begin with the fictional corpus study of the dative alternation introduced in Sections 2.1.3 and 2.1.4. We focus only on model specification here, and hence the full R commands including the specification of the link function and the distribution family are not shown. They are always assumed to be the logit link (i. e., the inverse logit function) and the binomial distribution in the examples.

First, we specify a minimal model as (1) with only the *Lemma* grouping factor and one other (binary) predictor, namely *Givenness*, both as fixed effects.

$$\texttt{Construction~1+Lemma+Givenness} \tag{1}$$

In the case of logistic regression in alternation modelling, *Construction* is binary (levels 0 or 1, corresponding to the two alternants in the example). Furthermore, *Lemma* has m levels (one for each lemma), and *Givenness* is also binary (levels 0 and 1, corresponding to *not given* and *given*). A line like (1) encodes a theoretical commitment to what the researcher thinks is the mechanism that determines which alternant is chosen. Concretely, it encodes the assumption that the probability of the outcome labelled 1 (often called the "success", which in the example corresponds to one of the alternants) can be predicted from the additive linear term specified as *1+Lemma+Givenness*. Because the influence of the regressors on the outcome is not linear in many cases, the additive linear term is transformed through the link function (here assumed to be the logit function), which is not encoded directly in R-type model formulæ. Also not part of the model formula in R is the specification of the distribution of the residuals (assumed to be binomial), which encodes the assumption that the distribution of the prediction errors follows the binomial distribution.[4] If another distribution (such as the Poisson distribution) and another link function (such as the logarithm, which is the default for Poisson models) is chosen, the specification in (1) remains the same.

In any type of GL(M)M, the additive linear term consists of a number of sub-terms which are simply added up. Each of these sub-terms (except for the intercepts) multiplies the (estimated) *coefficient* with an observed *value* of one

[4]As the example is still a GLM, this is a recapitulation of the previous chapter. Also, this might be significantly easier to understand in mathematical notation. Readers are encouraged to consult Gelman & Hill (2006).

of the variables. However, R notation for model formulæ simplifies the specification of the actual linear term. First of all, the 1 in *1+Lemma+Givenness* is R's way of encoding the fact that an *intercept* is part of the model. An intercept is a constant sub-term to which all other terms are added, and it can be seen as the reference value when all other sub-terms (corresponding to categorical or numeric regressors) assume 0 as their value. For binary regressors like *Givenness*, the only coefficient that is estimated directly encodes the value added to (in case of a positive coefficient) or subtracted from (in case of a negative coefficient) the linear term when the value of the regressor is 1 (in the example, when the referent is given). When the value of the regressor is 0 (for example, when the referent is not given), 0 is added to the intercept. The intercept thus encodes (among other things) something like a default for a binary regressor. If the default corresponds to, as in the example, non-givenness, phrases like "non-givenness is on the intercept" or "givenness equals zero is on the intercept" are often used. However, a grouping factor such as *Lemma* is usually a categorical variable with more than two levels. In such a case, each of the m levels of the grouping factor are *dummy-coded*, and for all but one of these binary dummy variables, a coefficient is estimated. Dummy coding is a way of encoding a categorical variable as a number of binary variables, see Table 4. Because the

Value of...			
Lemma	$\mathbf{l_1}$	$\mathbf{l_2}$	$\mathbf{l_3}$
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Table 4: Dummy coding of a categorical variable *Lemma* with four levels, resulting in the binary dummy variables l_1, l_2, l_3

first of the m levels of the grouping factor is encoded as the value 0 for all dummy variables, no coefficient is estimated for this level, and only $m-1$ sub-terms are added to the model, which means that only $m-1$ coefficients will be estimated. The first level of the grouping factor is thus "on the intercept" and becomes the reference to which all other levels are compared.[5]
To sum up and clarify, if in a given study *Lemma* has four levels dummy-coded as l_1, l_2, l_3, and *Givenness* is binary and coded as g with $g=1$ if the referent is given, the formula corresponding to (1) looks like (2) in mathematical notation, where the linear additive term is enclosed in [], each sub-term appears in (), and c encodes the choice of the two alternants.

$$Pr(c=1) = logit^{-1}\left[\alpha_0 + (\beta_{l_1} \cdot l_1) + (\beta_{l_2} \cdot l_2) + (\beta_{l_3} \cdot l_3) + (\beta_g \cdot g)\right] \quad (2)$$

In plain English: the probability Pr that the alternant of the construction coded as 1 is chosen $Pr(c=1)$ is calculated as the inverse logit of the linear term. The linear term is just the addition of the intercept α_0 and the measured values, each multiplied by its corresponding coefficient labelled β.
In such a model, the effect of each verb lemma is treated as a fixed population parameter, exactly the same as the effect of givenness. In other words, the algorithm which estimates the coefficients for the $m-1$ dummy variables tries to find a fixed value for each of them without taking the variation between them into account. With many levels, this requires a lot of data, and

[5]Picking one dummy as a reference level is necessary because otherwise infinitely many equivalent estimates of the model coefficients exist as one could simply add any arbitrary constant to the intercept and shift the other coefficients accordingly. However, the estimator works under the assumption that there is a unique maximum likelihood estimate. This extends to any other appropriate coding for categorical grouping variables.

levels for which only a few observations are available in the data set have very imprecise coefficient estimates with large confidence intervals.

This is where random effects come into play as an alternative. If we treat the same grouping factor as a random intercept, we let the intercept vary by group, i. e., each group is allowed to have its own intercept. Furthermore, we give the varying intercepts a (normal) distribution instead of estimating $m-1$ fixed population parameters. This means that the group-wise intercepts are assumed to be normally distributed around 0. This is the relevant difference between a fixed effect and a random effect.

In R, the model specification then looks like (3), where "1|" can be read as "an intercept varying by".

$$\texttt{Construction\~{}1+Givenness+(1|Lemma)} \tag{3}$$

The sub-term *Givenness* remains the same as in (1), and it is still treated as a fixed effect. The sub-term *(1|Lemma)* encodes that an intercept will be added to the linear term depending on which lemma is observed. Notice that the sub-term for the varying intercept (just like the one for the normal intercept) does bot involve multiplication. This is obvious in mathematical notation corresponding to (3) as shown in (4). In addition to the overall intercept α_0, there is another constant term α_{Lemma}, which is chosen appropriately for each level of *Lemma*, which is notated as $\alpha_{[Lemma]}$.

$$Pr(c=1) = logit^{-1}\left[\alpha_0 + \alpha_{[Lemma]} + (\beta_g \cdot g)\right] \tag{4}$$

Crucially, instead of estimating a batch of $m-1$ coefficients for the levels of the grouping variable, one varying intercept (assumed to come from a normal distribution) is predicted for each of its m levels. All more complex models to be discussed below are extensions of this approach. In the next section, it is discussed when fixed and random effects can and should be used.

2.2.2 Random effect or fixed effect

One commonly given reason to use a random effect instead of a fixed effect is that "the researcher is not interested in the individual levels of the random effect" (or variations thereof). Such recommendations should be taken with a grain of salt. Gelman & Hill (2006: 245–247) summarise this as well as other diverging and partially contradicting recommendations for what should be a random effect as found in the literature. They conclude that there is essentially no universally accepted and tenable conceptual criterion of deciding what should be a random effect and what a fixed effect. The author of this chapter agrees with their conclusion that random effects should be preferred whenever it is technically feasible. Understanding when it is technically feasible requires at least some understanding of two major points. First, the variance in the intercepts needs to be estimated if a random effect is used. Second, the random intercepts can be understood as a compromise between fitting separate models for each group of the grouping factor (*no pooling*) and fitting a model while ignoring the grouping factor altogether (*complete pooling*), see Gelman & Hill (2006: Ch. 12).

As was stated above, the random intercepts are assumed to come from a normal distribution, and therefore the variance between them has to be estimated with sufficient precision. From the estimated variance and the data for a specific group, the estimator predicts the *conditional mode* in a GLMM (or the *conditional mean* in a LMMs) for that group (see Bates 2010: Ch. 1).[6] The conditional mode/mean for a group is the value of the varying intercept for this group. It is the numerical value shown by `R` packages like `lme4` for each level of a random intercept variable. This procedure, however, requires that the number of groups must not be too low. As a rule of thumb, if there are fewer than five levels, a grouping factor should be included as a fixed effect, regardless of its conceptual interpretation. Although one often find default recommendations telling practitioners to use a speaker grouping variable as a random effect, it would be ill-advised to do so if there are exemplars from less than five speakers in the sample. Along the same lines, the mode (typically spoken vs. written) is not a suitable grouping factor for use as a random effect because it has too few levels.

If, however, the number of groups is reasonably large, the next thing to consider is the number of observations per group. Alternatives to using a random effect would be to estimate a separate model for each level of the grouping factor, or to include it as a fixed effect. In both cases the effects are not treated as random variables, and fixed coefficients per group are estimated without taking the between-group variance into account. With a random effect, however, the conditional modes/means are pulled (*shrunken*) towards the overall intercept (*shrinkage*). When there the number of observations in a group is low, the conditional mode/mean is simply shrunken more strongly towards 0, predicting only a small deviation from the overall tendency.[7] On the other hand, fixed effect estimates would become

[6] Notice that *mode* here refers to the statistical usage of the word.

[7] Terminologically, shrinkage is thus *stronger* (and the conditional mode/mean is closer to 0) if there is less evidence that a group deviates from the overall tendency. The lower the number of observations per group, the lesser evidence there is.

inexact and would probably be dismissed because of growing uncertainty in the estimate (large confidence intervals, large p-values) when the number of observations in a level is low. Thus, low numbers of observations in all or some groups are often detrimental for using fixed effects grouping factors. Random effects are much more robust in situations like this because of shrinkage. On the downside, a conditional mode that was strongly shrunken (due to a low number of observations) cannot be distinguished straightforwardly from a conditional mode of a group which simply does not deviate a lot from the average tendency. For fixed effects, we have both a parameter estimate and a possible significance test, but for random effects, we only have the prediction of the conditional mode/mean. However, so-called *prediction intervals* can be calculated for individual per-group intercepts, and we return to them in the following section.

2.2.3 Model quality and model selection

Significance It is not adequate to do any kind of significance testing on the levels of the random effect because they are not estimates in the conceptual and technical sense.[8] There are ways of calculating *prediction intervals* (which are not the same as confidence intervals) for conditional modes in order to specify the quality of the fit (see Section 3.1), but they should not be misused for talking about significance. Not doing significance tests for single levels of the grouping factor does, however, not mean that the researcher is not interested in the individual conditional modes, which is proven by the fact that they are often reproduced in research papers, for example in the form of

[8] Again, we do not assume them to be fixed population parameters, which would be the case for estimates such as fixed effects coefficients.

a dot plot. Also, the simulation in Section 2.2.2 shows that we can use a random effect and still get a good idea of the per-group tendencies. Additionally, a random effect allows the researcher to quantify the between-group variance, which is not possible for fixed effects.

Model selection A related question is *model selection*, i. e., whether the inclusion of the random effect improves the model quality. It is recommended here to include all conceptually necessary random effects and only remove them if they have no effect. To check whether they have an effect, the estimated between-group variance is the first thing to look at. If it is close to 0, there is most likely not much going on between groups, or there simply was not enough data to estimate the variance. In LMMs, it is possible to compare the residual (observation-level) variance with the between-group variance to see which one is larger, and to which degree. If, for example, the residual variance is 0.2 and the between-group variance is 0.8, then we can say that the between-group variance is four times larger than the residual variance, which would indicate that the random effect has a huge impact on the response. This comparison is impossible in GLMMs because their (several types of) residuals do not have the same straightforward interpretation as in LMMs. Furthermore, models can be compared using likelihood ratio (LR) tests. In such tests, a model including the random effect and a model not including it are compared, similar to LR tests for the comparison of fixed effects. Such pairs of models, where one is strictly a simplification of the other, are called *nested models* (not to be confused with *nested effects* discussed in Section 2.1.2). A sometimes more robust alternative to the ordinary LR test are parametric bootstrap tests (see also Section 3.1). With all this, it should be kept in mind that it is *never* appropriate to compare a GLMM with a random effect and a GLM with the same factor as a fixed effect using any test or metric (including so-called information criteria such as Akaike's or Bayes').

Quality of the fit To measure how well a GLMM fits the data, any metric that is based on prediction accuracy can be used in the same way as with GLMs. For example, the prediction accuracy on the data used for model estimation or cross-validation methods can be used.

Coefficients of determination (pseudo-R^2) can be used to give some idea of the overall model fit. For GLMMs, Nakagawa & Schielzeth (2013) have proposed a method that distinguishes between *marginal* R^2 (only fixed effects) and *conditional* R^2 (fixed and random effects). This has become a de facto standard. In cases where an effect works well as a fixed or a random effect (for example, if it has between five and ten levels with enough data points for each level), the marginal and conditional R^2 measures for the GLMM converge in an expected way with Nagelkerke's R^2 for corresponding GLMs. The marginal R^2 for a GLMM estimate is roughly the same as Nagelkerke's R^2 for a GLM estimate where the grouping factor is ignored. Also, the conditional R^2 for a GLMM estimate is roughly the same as Nagelkerke's R^2 for a GLM estimate which includes the grouping factor as a fixed effect.

2.2.4 More complex models

Varying intercepts and slopes In Section 2.1.4, it was shown under which conditions a varying-intercept and varying-slope (VIVS) model might be useful. Readers might want to review the example before continuing on. While it is possible to have just a varying slope, this is rarely useful, and we discuss only varying-intercept and varying-slope (VIVS) models.

A random slope is useful when the strength or direction of some fixed effect varies by group. We extend the simple model from (1) to include random slopes for *Givenness* varying by *Lemma* using `R` notation in (5). Each variable from the fixed effects part of the formula which we expect to vary by *Lemma* is simply repeated before the | symbol.

```
Construction~1+Givenness+(1+Givenness|Lemma)
```
(5)

With this model specification, a fixed coefficient for *Givenness* will still be estimated. However, an additional value will be predicted for each lemma, and this value has to be added to the fixed coefficient. In mathematical notation, this is very transparent, as shown in (6). The varying slope for *Givenness* to be chosen appropriately for each *Lemma* is specified as $\beta_{g[Lemma]}$.

$$Pr(c = 1) = logit^{-1}\left[\alpha_0 + \alpha_{[Lemma]} + ((\beta_g + \beta_{g[Lemma]}) \cdot g)\right] \tag{6}$$

A source of problems in VIVS models is the fact that in addition to the variance in the intercepts and slopes, the covariance between them has to be estimated. If in groups with a higher-than-average intercept, the slope is also higher than average, they are positively correlated, and vice versa. These relations are captured in the covariance. Technically speaking, the joint distribution of the random intercepts and the random slopes is assumed to follow a bivariate normal distribution with means, variances, and covariances to be estimated. The number of variance parameters to be estimated thus obviously increases with more complex model specifications, and the estimation of the parameters in the presence of complex variance-covariance matrices requires considerably more data than estimating a single variance parameter. The estimator might converge, but typically covariance estimates of −1 or 1 indicate that the data was too sparse for a successful estimation of the parameter. In this case, the model is *over-parametrised* and needs to be simplified (see Bates, Kliegl, et al. 2015, Matuschek et al. 2017).

Nested and crossed random effects As it was explained in Section 2.1.2, nested random effects are adequate when grouping factors are nested within other grouping factors. Technically, while varying slopes can be understood as interactions between a fixed and a random effect, nested random intercepts can be understood as interactions between two or more random effects. Crossed random effects are just several unrelated random effects. (7) shows the model specification, extending (**??**) with a varying intercept α_s. This could be for example semantic classes which nest individual lemmas. It could also be another grouping factor for speaker, completely unrelated to the lemmas.

$$P(y^i = 1) = logit^{-1}(\alpha_s^{k[i]} + \alpha_l^{j[i]} + \beta_d \cdot x_d^i) \tag{7}$$

The difference is that in the nested case, $k[i] = k[j]$, i. e., the level of the nesting factor can be selected based on the nested factor as well as based on the single observation. As was mentioned in Section 2.1.2, the question is rather one of how the way the data are organised.

Second-level predictors In Section 2.1.3, situations were introduced where the random effects themselves can be partially predicted from fixed-effects. In this case, an additional linear model is specified for the random effect instead of the simple normal distribution predictor. We extend (**??**) by a predictor γ_f for the lemma frequency. The lemma frequencies themselves we denote by u_f, and we index them with j, just like the verb lemmas. This is reasonable because for each verb lemma, there is exactly one frequency. The first-level model specification remains the same, namely (8).

$$P(y^i = 1) = logit^{-1}(\alpha_l^{j[i]} + \beta_d \cdot x_d^i) \tag{8}$$

However, instead of (??), the varying intercept is predicted from (9).

$$\alpha_l^j \sim N(\gamma_0 + \gamma_f \cdot u_f^j, \sigma_l^2) \tag{9}$$

Instead of just the mean of the α_j values, the model in (9) specifies a second-level intercept γ_0 and a second-level fixed coefficient γ_f.

3 Practical guide with R

3.1 Specifying models using `lme4` in `R`

This section and the next focus on `lme4`, an often used package to do multilevel modeling in `R` with maximum likelihood methods (Bates, Mächler, et al. 2015).

Varying intercepts The functions `lmer` and `glmer` extend the syntax of `lm` and `glm`. The varying intercept model in (??) is specified as follows in `R` (using informative variable names instead of Greek letters).

```
glmer(formula = construction ~ given + (1 | lemma),
      family = binominal(link=logit), data = my.data)
```

The pipe operator `x1|x2` can be read as *x1 varies by x2*. The intercept is denoted by `1`, and hence `(1|lemma)` simply says that the intercept varies by lemma.

Varying intercepts and slopes The VIVS model in (??) is specified as follows (only the formula).

```
construction ~ given + (1 + given | lemma)
```

Before the pipe, the part of the model is repeated that should be modeled as varying by the grouping factor after the pipe. If a varying slope is specified, a varying intercept is silently assumed. The last formula can therefore be abbreviated to the following equivalent one.

```
construction ~ given + (given | lemma)
```

In order to let *only* the slope vary, the intercept has to be removed explicitly from the random part of the formula.

```
construction ~ given + (given - 1 | lemma)
```

Multiple random effects When there is more than one random effect, several bracketed terms are added. The following is the recommended specification for models like (7), regardless of whether the effects are nested or crossed.

```
construction ~ given + (1 | lemma) +
                       (1 | semantics)
```

Sometimes the following notation is used for nested random effects, where `semantics` nests `lemma`.

```
construction ~ given + (1 | semantics / lemma)
```

`lme4` expands this to the following underlying syntax, which shows more clearly that nesting is handled as a kind of interaction.

```
construction ~ given + (1 | semantics) +
               (1 | semantics : lemma)
```

There is a random intercept for `semantics` and one for each combination of `semantics` and `lemma`. While these notations are seemingly very explicit about the nesting structure, they are not necessary under normal circumstances. If the grouping factor `lemma` is nested within `semantics` (see Table 1 for a similar situation), `lme4` automatically treats it as nested, and the results are exactly the same with all the three aforementioned notations. However, the following specification is *not* equivalent and leads to problematic results.

```
construction ~ given + (1 | semantics) +
               (1 | lemma) +
               (1 | semantics : lemma)
```

This instructs `lme4` to estimate the variance of `lemma` not just restricted to the permutations of the levels of `lemma` and `semantics` (i. e., `semantics:lemma`), but also outside of these specific permutations. In the nested case, there are no occurrences outside of these permutations, however, and the variance for `lemma` alone will be estimated close (but not exactly equal) to `0`. To compensate for the spurious estimate for `lemma`, the variance estimate for `semantics:lemma` will be shifted unpredictably.

Exemplar	Speaker	Region
1	D	Tyneside
2	D	Tyneside
3	R	Tyneside
4	R	Tyneside
5	D	Greater London
6	D	Greater London
7	R	Greater London
8	R	Greater London

Table 5: Illustration of nested factors, organised suboptimally

There is one situation where the explicit notation for nested factors is necessary. This is when the data are stored in a suboptimal way. Such a suboptimal version of Table 1 would look something like Table 5. Here, the speaker factor is encoded as the initial letter of the name only. Hence, Daryl and Dale (coming from two different regions) cannot be distinguisehd from each other, and Riley and Reed cannot, either. This leaves `lme4` no way of recognising that the data structure is nested, and the user has to explicitly provide that information. It would, of course, be better *not* to organise data that way.

Second-level predictors (8) and (9) have the following `lme4` syntax.

```
construction ~ given + frequency + (1 | lemma)
```

If the data is organised as shown in Table 3 – i. e., , with the second-level regressor not having any variance within the levels of the grouping factor –, `lme4` will detect this and treat `frequency` as a second-level effect. However, second-level predictors for random slopes are more tricky to specify (see Gelman & Hill 2006: 280-282). Assuming that the effect of givenness varies with the lemma, which itself comes with a second-level model including frequency as a regressor, the specification looks as follows.

```
construction ~ given + frequency +
               given : frequency +
               (1 + given | lemma)
```

A second-level regressor on a varying slope is thus an interaction between a first-level and a second-level fixed effect.

3.2 After fitting models with `lme4`

Basics and varying intercepts The output for GLMMs in `lme4` can be understood straightforwardly after what was said in Section 2. Here is a possible output of the `summary` function for a fit of model (**??**) and (**??**), repeated as (10) and (11). Artificial data were used.

$$P(y^i = 1) = logit^{-1}(\alpha_l^{j[i]} + \beta_d \cdot x_d^i) \tag{10}$$

$$\alpha_l^j \sim N(\mu_l, \sigma_l^2) \tag{11}$$

```
Generalized linear mixed model fit by maximum likelihood
 Family: binomial  ( logit )
Formula: construction ~ given + (1 | lemma)
   Data: observations
Random effects:
 Groups Name        Variance Std.Dev.
 lemma  (Intercept) 1.29     1.136
Number of obs: 250, groups:  lemma, 5
```

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7638     0.5513   1.385    0.166
given1        1.5064     0.3626   4.154 3.26e-05 ***
```

Some clutter as well as information which we do not interpret here (AIC, BIC, and information about the residuals) have been removed. In this output, the `(Intercept)` estimate (0.7638) is $\hat{\mu}_l$, and the `Variance` estimate for the `lemma` random intercept (1.29) is $\hat{\sigma}_l^2$.[9] The estimate for `given1` (1.5064) corresponds to $\hat{\beta}_d$. Finally, we learn that there were five different lemmas and 250 observations in total.

To see whether the random intercept has a considerable influence, we should first look at the variance estimate. Here, it is larger than 1, which would be surprising if there were nothing going on in terms of between-lemma variation. It is possible to compute confidence intervals for the variance estimate using the `confint` function. Assuming the original model was stored in `alternation`, the following two alternatives work.

```
confint(alternation, parm="theta_", method = "profile")
confint(alternation, parm="theta_", method = "boot",
        nsim = 250)
```

The profile method uses LR tests and the bootstrap method uses a parametric bootstrap. For this model (where the variance estimate was 1.29 and the true value used to generate the data was 1.5), the profile method gives $0.5808\ldots2.6433$ and the bootstrap with 250 simulations gives

[9] The notation $\hat{v}$ is used to denote and estimate of or a prediction for the variable v.

$3.9665 \cdot 10^{-6} \ldots 1.8023$ as the 95% confidence interval. Since the bootstrap (especially with smaller original sample sizes as in this case) typically tends to run into replications where the estimation of the variance fails and is thus returned as 0, the bootstrap interval is skewed to the left, while the profile confidence interval frames the true value symmetrically. The bootstrap is thus not always more robust or intrinsically better.

Although the authors of the `lme4` package advise against it, a significance test on the deviances of a simple GLM and a GLMM with an added single random effect can be performed with the `anova` function.

```
alternation.0 <- glm(construction ~ given,
                     data = observations,
                     family = binomial(link=logit))
anova(alternation, alternation.0)
```

The GLMM object must be the first argument to `anova`. In this case, the output looks like this, indicating a significant effect, although the p-values should not be considered highly reliable.

```
Data: observations
Models:
alternation.0: construction ~ given
alternation:   construction ~ given + (1 | lemma)
             Df  logLik deviance  Chisq Df Pr(>Chisq)
alternation.0 2 -134.52   269.05
alternation   3 -119.12   238.24 30.801  1  2.859e-08 ***
```

If the nested (simpler) model still contains a random effect, the `update` function can be used to build the simpler model. Also, in addition to the anova command, there is a drop-in replacement for LR tests using bootstrap methods in the pbkrtest package (Halekoh & Højsgaard 2014).

```
alternation.1 <- glmer(construction ~ given +
                       (1 | lemma) + (1 | semantics),
                       family = binominal(link=logit),
                       data = my.data)
alternation.2 <- update(object = alternation,
                        formula = construction ~
                                  given +
                                  (1 + | lemma))


require(pbkrtest)
PBmodcomp(alternation.1, alternation.2, nsim = 250)
```

The coefficient of determination (pseudo-R^2) according to Nakagawa & Schielzeth (2013) can be computed using the function r.squaredGLMM from the MuMIn package (Bartoń 2016).

```
require(MuMIn)
r.squaredGLMM(alternation)
```

In this case, it gives us `R2m = 0.1101` and `R2c = 0.3608`, so there is a considerable difference between the marginal R^2 (without random effects) and conditional R^2 (with random effects).

To inspect the conditional modes, the ranef function can be used, and it can also output standard errors for them.

```
ranef(alternation, drop = T, condVar = T)
```

Varying intercepts and slopes In the `summary` output for a VIVS model such as (??), the columns `Variance` and `Corr` can be regarded as specifying the lower triangle of the variance-covariance matrix.[10]

```
Random effects:
 Groups Name        Variance Std.Dev. Corr
 lemma  (Intercept) 0.6103   0.7812
        given       0.8944   0.9457   -0.39
Number of obs: 2500, groups:  lemma, 50

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6332     0.1226   5.163 2.43e-07 ***
given        -1.0428     0.1492  -6.990 2.75e-12 ***
```

This output tells us that the estimated variance in the intercepts is $\hat{\sigma}_{\alpha}^{2} = 0.6103$, the estimated variance in the slopes is $\hat{\sigma}_{\beta}^{2} = 0.8944$, and the covariance coefficient estimate is $\hat{\rho} = -0.39$ (a healthy value inasmuch as it is not 1 or -1). The means are estimated as $\hat{\mu}_{\alpha} = 0.6332$ and $\hat{\mu}_{\beta} = -1.0428$. Compare this to Section 2.2.4, especially (??). It is possible to reconstruct group-wise models from this output and a lookup of the group-specific predictions using the `ranef` function. For the first group, for example, the following can be done.

[10] Alternatively, the `VarCorr` function could be used to extract this information.

```
ranef(alternation)$lemma[1,]
```

The output is as follows.

```
     (Intercept) given
0001   0.4351156 -1.227842
```

This means that for the first lemma, actual predictions for the outcome of the alternation can be made using (12), where values are rounded to two decimal digits. Compare this to (??).

$$Pr(y^i = 1) = logit^{-1}([0.63 + 0.44] + [-1.04 - 1.23] \cdot x_d^i) \tag{12}$$

4 Representative studies

Wolk et al. (2013)

Research questions The authors aim to achieve two things. First, they want to compare changes in two word order-related alternations in the history of English between 1650 and 1999: the dative alternation and the genitive alternation. They look for influencing features shared in both cases as well as construction-specific features. Second, they aim to show that historical data fits well into a probabilistic, cognitively oriented view of language.

Data The authors use the ARCHER corpus (Biber, Finegan & Atkinson 1994), which contains texts from various registers from 1650 to 1999. For both constructions, carefully designed sampling protocols were used (see their Section 4). For the annotation of the data, both available corpus meta data were used (text ID, register, time in fractions of centuries, centered at 1800) as well as a large number of manually coded variables (constituent length, animacy, definiteness, etc.). Furthermore, the possessor head lemma (genitive alternation) and the verb lemma (dative alternation) were coded.

Method Two mixed effects logistic regression models are estimated. For the genitive alternation, the text ID and the possessor head lemma are used as crossed random effects. The authors state on p. 399 that they collapsed all head noun lemmas with less than four occurrences into one category because otherwise "difficulties" would arise. However, it is the advantage of random effects modeling that it can deal with a situation where categories have low numbers of observations (see *shrinkage*, Section 2.2.2). For the dative alternation, the model includes the text ID, the register (which nests the text ID) as well as the lemma of the theme argument and the verb.

Results It is found that many factors have a shared importance in both alternations, e. g., definiteness, animacy, construction length. It is also argued that the observed tendencies – such as *short-before-long* and *animate referents first* – are in line with synchronic corpus based and experimental findings about general cognitive principles underlying the framework of probabilistic grammar. These principles remain in effect, but the strength of their influence changes over time.

Gries (2015)

Research questions The paper is programmatic in nature. The author re-analyses data from a previously published study on verb particle placement in English. He uses a GLMM instead of a fixed-effects logistic regression to show that including random effects in order to account for variation related to mode, register, and subregister increases the quality and predictive power of the model. He also argues that by not doing so, corpus linguists risk violating fundamental assumptions about the independence of the error terms in models.

Data The data are 2,321 instances of particle verbs showing either verb–direct object–particle or verb–particle–direct object order, taken from the ICE-GB. The grouping factors derived from the structure of the corpus are mode (only two levels), register (five levels), and subregister (13 levels). They are nested: mode nests register, which nests subregister. Additionally, verb and particle lemma grouping factors are annotated. Finally, two fixed effects candidates are annotated (the type of the head of direct object and the logarithmised length of the direct object in words).

Method The author uses the model selection protocol described in Zuur et al. (2009) to first find the optimal random effects structure using ANOVAs and AIC comparisons as well as analyses of the estimated variance for single random effects. He then goes on to find the optimal fixed effects structure. Additionally, he compares the pseudo-R^2 measure of the resulting mixed models.

Results Gries finds that the verb and particle lemma play and the subregister play a significant role. Notably, the variance estimate for mode is close to 0 from the beginning of the model selection procedure. This is not surprising, as two levels are not nearly enough in order for the variance to be reliably estimated, and it could maybe be used as a second-level predictor instead (see Section 2.2.2). The R^2 values of the final model are very high, with a considerable difference between marginal $R^2_m = 0.57$ and conditional $R^2_c = 0.748$, which indicates that the random effects do in fact improve the model fit. It is also shown that the classification accuracy is considerably improved over that of a GLM without random effects, but differently for different lexical groups and subregisters. The paper thus shows that it is not appropriate to ignore lexical grouping factors and grouping factors derived from the corpus structure, especially as both are easy to annotate automatically.

5 Further reading

Chapters 1–15 and Chapters 20–24 of Gelman & Hill (2006) are a highly recommended read, especially for `R` and `lme4` users. Similarly, Zuur et al. (2009) has a reputation among `R` users of mixed effects models in many fields. The companion to `lme4`, Bates (2010) and the overview in Bates, Mächler, et al. (2015) are obligatory reads for users of `lme4`.

References

Bartoń, Kamil. 2016. *MuMIn: Multi-Model Inference.* R package version 1.15.6. `https://CRAN.R-project.org/package=MuMIn`.

Bates, Douglas M. 2010. Lme4: mixed-effects modeling with R. http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth & R. Harald Baayen. 2015. *Parsimonious Mixed Models*. https://arxiv.org/abs/1506.04967.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Biber, Douglas, Edward Finegan & Dwight Atkinson. 1994. Archer and its challenges: compiling and exploring a representative corpus of historical english registers. In Udo Fries, Peter Schneider & Gunnel Tottie (eds.), *Creating and using english language corpora*, 1–13. Amsterdam: Rodopi.

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

Gries, Stefan Th. 2015. The most underused statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10(1). 95–126.

Halekoh, Ulrich & Søren Højsgaard. 2014. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software* 59(9). 1–30.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Douglas Bates. 2017. Balancing type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.

Nakagawa, Shinichi & Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2). 133–142.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul, Turkey: European Language Resources Association (ELRA).

Schielzeth, Holger & Wolfgang Forstmeier. 2009. Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology* 20(2). 416–420.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.

Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer.