Mixed-effects regression modeling

Roland Schäfer

Freie Universität Berlin

August 16, 2017

1 Introduction

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

interpreting the output. Readers are assumed to be familiar with the concepts covered in the previous chapter.

2 Fundamentals

2.1 When are random effects useful?

2.1.1 Introduction to random effects

Generalized Linear 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). 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.[1] 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

[1] 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 the residuals.

maximization. Second, *link functions* are introduced which relate the additive linear term that combines the predictors in a non-linear way to the response variable. Generalized Linear Mixed Models (GLMMs) are not much different. They add what are often called *random effects* 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 2.1.3.

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 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. We know that preferences vary between speakers, and it is therefore reasonable to take care of 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. 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 analysis and understanding of the data set at hand and the differences and similarities in the resulting estimates. Subsection 2.1.2, 2.1.3, and 2.1.4 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, Section 2.2 provide a moderately 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 2.3 then shows how mixed models are specified and interpreted using R.

2.1.2 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

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

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 Daryl and Riley, and Greater London is the unique region value for Dale and Reed. There cannot be exemplars where, for example, the speaker is Daryl 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. 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 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 2.3).

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

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

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 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.[2] Similarly, we might want to describe – in a given study on adjectives – 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.

2.1.3 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 not a technical matter 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. A similar situation would arise in a study of learner corpus data with a learner grouping factor if we also knew

[2] 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.

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

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.[3]

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 ditransitive verb 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

[3]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. See also the last subsection of Section 2.2.4.

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 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 second (linear) model which 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. 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 often 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. Therefore, multilevel modeling will be introduced as the more general framework for so-called mixed effects models in Sections 2.2 and 2.3.

2.1.4 Random slopes as interactions

Before moving on to the more technical discussion of hierarchical model specification in Section 2.2, 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.

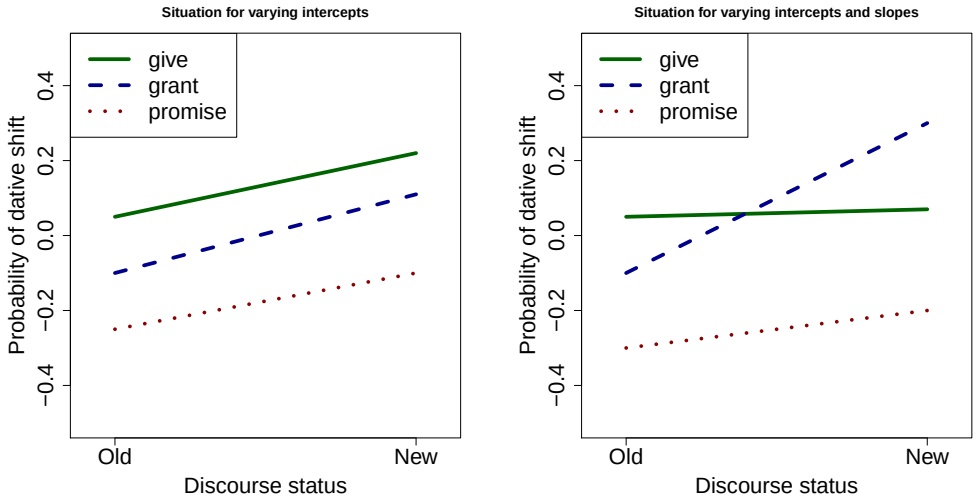

Figure 1: Illustration of data in situations for varying intercepts or varying intercepts and additional varying slopes

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 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 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.

2.2 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, but a few notational conventions are introduced as the absolute minimum required to understand both the output of lme4 and other packages in R as well as the literature on mixed models. To make successful practical use of mixed models, some level of fundamental understanding is required. This section is based mostly on Part 2A of Gelman & Hill (2006) (pp. 235–342).

2.2.1 Simple random intercepts

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 grouping factors discussed in the previous sections) could 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. 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 first specify a minimal model with only the dummies of the lemma grouping factor and one other (binary) 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. In general, α is used for intercepts and β for coefficients. A specification of such a model is given in (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 models the 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 the discourse status for exemplar i (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 in observation i. If in exemplar 64, the lemma is *give* and *give* is encoded as group 12, then

$i = 64$, $j = 12$, and $x^{64}_{l_{12}} = 1$, whereas all $x^{64}_{l_j} = 0$ with $j \neq 12$. 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., $j = 1, \cdots, m-1$.[4] The function $logit^{-1}$ is the *link function*, and its argument is the *linear term* of 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 random intercept, we let the intercept vary by group, and we give the varying intercepts a distribution instead of estimating $m-1$ coefficients. This is the only relevant difference between a fixed effect and a random effect. The model specification then looks like in (2).

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

We now have an intercept $\alpha_l^{j[i]}$ that varies by group (instead of one term with its own coefficient per group). We use the notation $\alpha_l^{j[i]}$ (borrowed in a modified form from Gelman & Hill 2006) to indicate that the correct j-th lemma intercept is chosen for the i-th observation. For example, if in exemplar 64, the lemma is *give*, which is group 12, then $i = 64$ and $j[64] = 12$ (i. e., the group appropriate for exemplar 64 is group 12), and $\alpha_l^{j[64]} = \alpha_l^{12}$. The term $\beta_d \cdot x_d^i$ for the effect of discourse status remains unchanged when going from (1) to (2). Crucially, instead of estimating a batch of coefficients for the lemma effect, α_l is itself modeled, and random terms are predicted for each level of the grouping

[4]Picking 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.

factor. For this, the assumption in (3) is made.

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

This is standard notation to indicate that the values of α_l^j follow a normal distribution with mean μ_l and a variance of σ_l^2. In fact, we can regard (3) as a minimal second-level model already, although one which simply predicts varying intercepts from a normal distribution. All more complex models to be discussed below are extensions of this approach. In the next section, the consequences of going from a fixed effect to a random effect are discussed.

2.2.2 Choosing between random and fixed effects

There are primarily two points to consider which influence the decision to use random effects. First, the variance in the intercepts (and for random intercept-random slope models also the covariance between intercepts and slopes) needs to be estimated. 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). While all conditions which were discussed in the previous chapter (independence of observations, non-collinearity, etc.) must also be met by hierarchical models, these two points add additional conditions.

As was stated above in (3), the random intercepts are assumed to follow a normal distribution, and the variance σ_l^2 needs to be estimated with sufficient precision. From the estimated variance and the data, the estimator then predicts the *conditional modes* in GLMMs (*conditional means* in LMMs) for each group (see Bates 2010: Ch. 1), which is the numerical value which software packages

like lme4 produce, and these values are sometimes sloppily called "random effects" by practitioners. This procedure, however, requires that the number of groups must not be too low to effectively achieve this. As a rule of thumb, fewer than five levels means that a grouping factor should be included as a fixed effect, regardless of its conceptual nature. Even if there is a default recommendation to use a speaker grouping variable as a random effect, it is ill-advised to do so if there are exemplars from less than five speakers in the sample. Very often, the estimator will fail anyway under such conditions. At the same time, the estimation of coefficients for a (dummy-coded) fixed effect becomes less precise and feasible with large numbers of levels, and at a certain point, a random effect might be the best option for technical reasons.

If, however, the number of groups is reasonably large, the next thing to consider is the number of observations per group. As mentioned at the outset of this section, 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 values per group are estimated without taking the between-group variance into account. The conditional modes are *shrunken* towards the overall intercept as a result (*shrinkage*), and they tend to be more leveled compared to fixed effects estimates. This effect becomes considerably weaker with larger per-group sample sizes. With relatively low numbers of observations per group, on the other hand, fixed effect estimates tend to become inexact and will probably be dismissed because of growing uncertainty in the estimate (large confidence intervals, non-significance). In this case, the conditional modes (or means) of random intercepts will just be shrunken more towards the overall intercept. This considered, including a grouping factor as a random effect might be the

only way of using it at all when the estimation as a fixed effect fails.

This section closes with an illustration.[5] For this, 1,000 data sets were simulated which corresponded to the model in (4) and (5). We drop the subscripts on α, β, and μ for convenience since there is only one random intercept.

$$P(y^i = 1) = logit^{-1}(\alpha^{j[i]} + \beta_1 \cdot x_1^i + \beta_2^i \cdot x_2^i) \tag{4}$$

$$\alpha^j \sim N(\mu, \sigma) \tag{5}$$

Again, this could be a model of a binary alternation. x_1 was a binary variable (such as singular/plural) and x_2 a continuous variable (such as NP length). Since the data were simulated, the parameters to be estimated were known: $\beta_1 = 0.8$, $\beta_2 = -1.3$, $\mu = 0$, $\sigma = 1.5$. The number of groups was set to 5 (the recommended minimum), the simulated values of the grouping factor were identical in each simulation, and there were 20 observations per group. Figure 2 shows the distribution of the group levels based on the conditional modes predicted of all but the first group in the $1,000$ simulations. Figure 3 shows the group estimates from a model where the grouping factor was added as a fixed effect.[6]

The per-group predictions lean slightly towards 0 in the GLMM (Figure 2), but the fixed effects estimates in the GLM are prone to massive misestimations (see the large spread and increased 95% intervals). Even with as few as five

[5]The code for these and other simulations is available under a Creative Commons Attribution license: `https://github.com/rsling/Rstuff/tree/master/simulations/glmm`

[6]The plots do not show the distribution of the raw conditional modes and coefficient estimates of the fixed effects. Rather, the overall intercept was taken into account, and the plots thus show the distribution of the per-group prediction of the models, all other things being equal. This is what was pre-specified in the simulations.

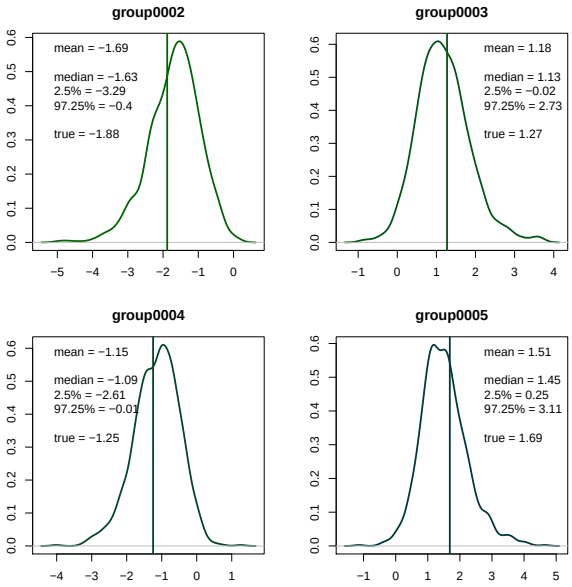

Figure 2: Group levels in sample GLMM based on predicted random effect (conditional mode); 5 groups; 20 observations per group; 1,000 simulations; the horizontal line marks the true value

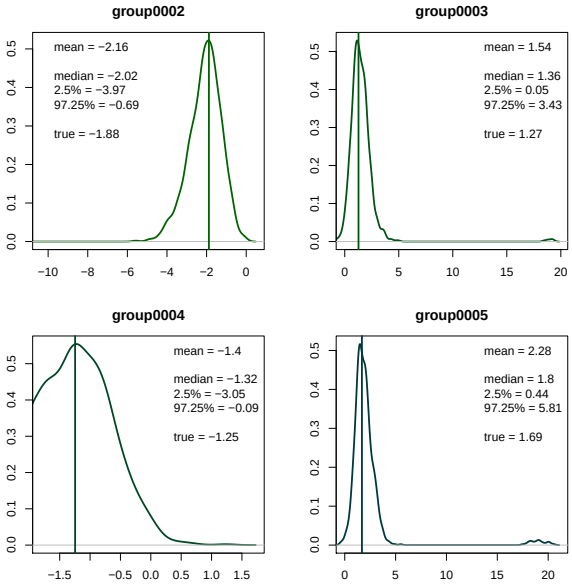

Figure 3: Estimated fixed effects for the grouping factor in sample GLM; 5 groups; 20 observations per level of the grouping factor; 1,000 simulations; the horizontal line marks the true value

levels of a grouping factor, however, random and fixed effects lead to very similar results, albeit with different advantages and disadvantages. A few more differences will be discussed in Sections 2.2.3 and 2.2.4.

2.2.3 Significance testing, model selection and coefficients of determination

One commonly given reason to use a random effect is that "the researchers are not interested in the individual levels of the random effect factor" (or variations thereof). Such recommendations should be taken with a grain of salt. Gelman & Hill (2006: 245–247) summarise the diverging and partially contradicting recommendations for what should be a random effect along with their motivations. They conclude that there is essentially no clean conceptual way of deciding what should be a random effect and what a fixed effect. In this chapter, a more technical approach was therefore suggested; see Section 2.2.2. Clearly, however, it is not adequate to do any kind of significance test on the levels of the random effect factor, as is customary for fixed effects. The conditional modes are not estimates, and only for estimates is significance testing a well-defined activity. There are ways of calculating confidence intervals for conditional modes (see Section 2.3), but they should not be misused for talking about significance. If this is absolutely required, fixed effects are the way to go. 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 a dot plot. Additionally, a random effect allows the researcher to quantify the between-group variance, which is not possible in the same way with fixed effects.

A related question is *model selection*, i. e., whether the inclusion of the random effect improves the model quality. First, the between-group variance should be checked. 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 $\sigma_\epsilon = 0.2$ and the between-group variance is $\sigma_\alpha = 0.8$, then we can say that the between-group variance is four times larger than the residual variance, which would indicate a high importance of the random effect. This comparison is impossible in GLMMs because their (several types of) residuals do not have the same straightforward interpretation as in LMMs.

Traditional model selection for random effects can be achieved through a likelihood ratio (LR) test comparing a model including the random effect and a model not including it. 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). An alternative option are parametric bootstrap replacements for the LR test (see Section 2.3). It is *not* 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 information criteria).

Coefficients of determination (pseudo-R^2) should not be used for model selection, because they usually do not penalize complexity enough (i. e., the pseudo-R^2 measures rise with added complexity). They can still be used to give some idea of the model fit, however. For GLMMs, Nakagawa & Schielzeth (2013) have proposed a method that distinguishes between *marginal* R^2 (fixed-effects-only) and *conditional* R^2 (including random effects). This has become a de facto

standard, and we now show its consistency with Nagelkerke's R^2 for GLMs. Using the simulations described in the last section, Figures 4 and 5 show that the marginal R^2 is roughly the same as Nagelkerke's R^2 in a GLM which ignores the grouping factor, and the conditional R^2 is roughly the same as Nagelkerke's R^2 in a GLM which includes the grouping factor as a fixed effect, given that the estimators converge for all these three types of models.

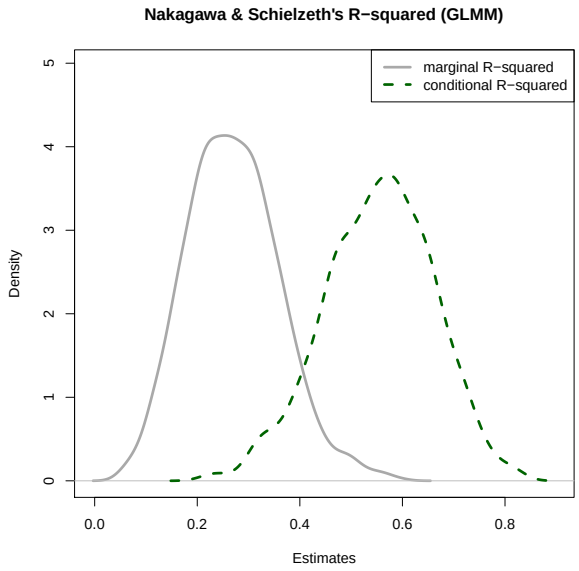

Figure 4: Distribution of Nakagawa & Schielzeth's R^2 in the simulations described in Section 2.2.2

Finally, it should be mentioned that adding a random intercept does not change the estimates nor the inferences (p-values) for fixed effects, at least if the grouping factor and the fixed effect are independent. Figure 6 compares the p-values obtained for the two fixed effects in the $1,000$ simulations described above for the GLMM and the GLM ignoring the grouping factor. Notice that the β_2 effect has the higher effect strength, and p-values are generally lower. If fixed effects and random effects are not independent, adding a varying slope should be considered, which is described in Section 2.2.4.

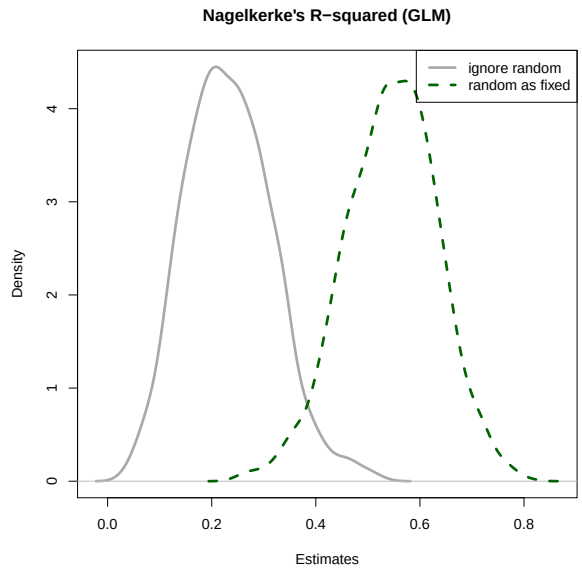

Figure 5: Nagelkerke's R^2 in the simulations described in Section 2.2.2 for a GLM that ignores the grouping factor and a model that includes it as a fixed effect

2.2.4 More complex models

This section shows briefly how more complex models are specified before Section 2.3 demonstrated how models are estimated in R, including interpretations of the output.

Varying intercepts and slopes A varying slope is an interaction between a fixed effect and a random effect, see Section 2.1.4. 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. We extend the simple model from (2), and the fixed effect coefficients for which a random slope is specified simply receive group indices; see (6). Put simply, instead of a fixed coefficient, coefficients are predicted and assumed to come from a random (normal) distribution. We use $\beta_{d:l}$ to denote the coefficient for discourse status varying by lemma.

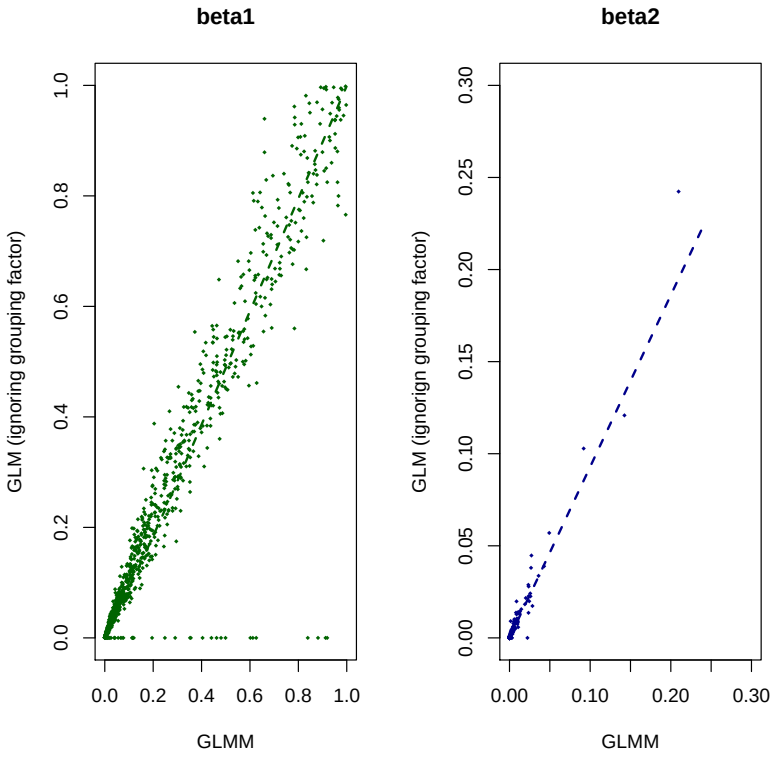

Figure 6: Comparison between the p-values of the fixed effects in the simulations described in Section 2.2.2 for the GLMM and the GLM ignoring the grouping factor; the line was drawn using LOWESS (locally weighted scatterplot smoothing); the axes between the left and the right panel are on a different scale

$$P(y^i = 1) = logit^{-1}(\alpha_l^{j[i]} + \beta_{d:l}^{j[i]} \cdot x_d^i) \tag{6}$$

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. Therefore, condition (7) is added, where the indices l and $d : l$ are omitted for readability.

$$\begin{pmatrix} \alpha^j \\ \beta^j \end{pmatrix} \sim \left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right) \tag{7}$$

(7) says that the joint distribution of the intercepts α^j and the slopes β^j follow a bivariate normal distribution with means μ_α and μ_β. The variance in the intercepts is σ_α, the variance in the slopes is σ_β, and the covariance between them is ρ. Figure 7 shows the bivariate density distributions for two (1) negatively correlated, (2) non-correlated, and (3) positively correlated normally distributed variables.

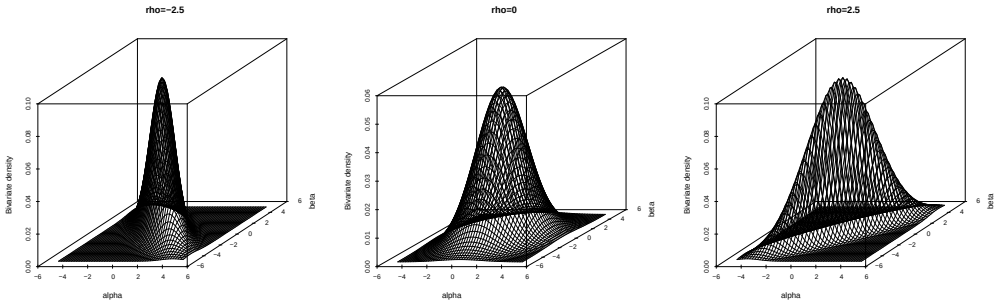

Figure 7: Bivariate normal density distribution with different correlation strengths (ρ); $\sigma_\alpha = \sigma_\beta = 3$; $\mu_\alpha = \mu_\beta = 0$

Importantly, the number of variance parameters to be estimated obviously in-

creases 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-parametrized* and needs to be simplified. In Section 2.4, it will be shown how to interpret the lme4 output appropriately.

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 ca be understood interactions between two or more random effects. Crossed random effects are just several unrelated random effects. In the case of crossed and nested random intercepts, there are just more than one random intercept. (8) shows the model specification, extending (2) with a varying intercept α_c. 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.

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

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. In the crossed case, this is not the case. As was mentioned in Section 2.1.2, the questions is rather one of how the way the data are organised. In the model specification (including model specification in R), there is not

much to consider.

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 (2) 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 (9).

$$P(y^i = 1) = logit^{-1}(\alpha_l^{j[i]} + \beta_d \cdot x_d^i) \tag{9}$$

However, instead of (3), the varying intercept is predicted from (10).

$$\alpha_l^j \sim N(\gamma_0 + \gamma_f \cdot u_f^j, \sigma_l^2) \tag{10}$$

Instead of just the mean of the α_j values, the model in (10) specifies a second-level intercept γ_0 and a second-level fixed coefficient γ_f. If the lemma frequency plays a role, this approach can lead to a more precise prediction of the varying first-level intercept compared to the less elaborate approach in (3). True multilevel models increase the complexity of GLMMs, especially if third-level models, fourth-level models, etc. are used. Situations for multilevel modeling are quite frequently encountered, however. Especially when it comes to speakers as random effects, the age, the gender, the region of birth (if this grouping factor has too few levels to be used as a random effect nesting speakers), etc. are ideal second-level predictors. The lme4 syntax treats second-level predictors like interactions between a fixed and a random effect, see also

Section 2.3.

Readers finding it difficult to see why (10) is a plain linear model could consider the equivalent formulation in (11) and (12).

$$\alpha_l^j = \gamma_0 + \gamma_f \cdot u_j^j + \epsilon_l^j \tag{11}$$

$$\epsilon_l^j \sim N(0, \sigma^2) \tag{12}$$

Models for longitudinal studies A longitudinal study is one wherein single subjects (usually speakers) are observed at different points in time, for example second-language learners after different years of learning a second language. This section briefly discusses the main points to consider when such models – which can get quite complex – are specified. First, the observations are obviously grouped by the individual speakers. It is thus recommended to include the speaker grouping factor, either as a fixed effect (up to five speakers) or random effect (more speakers). Second, there might be time parameters such as years of learning. Third, the errors might be correlated.

We now assume we have a sample from a learner corpus of German as a second language. In a logistic regression, we examine whether Swedish and English learners use the weak or the strong forms of attributive adjectives in NPs with a determiner. Grammatically, the crucial variable is whether the determiner has itself a strong ending or not, and we include an appropriate first-level term $\beta_d \cdot x_d^i$ in the model. A random effect for individual learners is also included as $\alpha_l^{j[i]}$. Additionally, we add a term for the number of years the learners have learned German (potentially logarithmized or otherwise transformed). However, since learners progress with different speed, a random slope is indicated, and the

term becomes $\beta_{y:l}^{j[i]} \cdot x_y^i$. The first-level model is thus (13).

$$P(y^i = 1) = logit^{-1}(\alpha_l^{j[i]} + \beta_d \cdot x_d^i + \beta_{y:l}^{j[i]} \cdot x_y^i) \tag{13}$$

The main purpose of our study might be to find out whether Swedish and English learners differ with respect to the phenomenon at hand. Therefore, the first language is added as a second-level predictor with the term $\gamma_f \cdot u_f^j$, where u_f^j is 1 if the language of learner j is Swedish, and 0 if it is English. Since we have a random intercept and a random slope, the second-level model becomes more complex, and we need to distinguish between $\gamma_f^\alpha \cdot u_f^j$ for the random intercept and $\gamma_f^\beta \cdot u_f^j$ for the random slope. The second level model to go with (13) is (14).

$$\begin{pmatrix} \alpha^j \\ \beta^j \end{pmatrix} \sim \left(\begin{pmatrix} \gamma_0^\alpha + \gamma_f^\alpha \cdot u_f^j \\ \gamma_0^\beta + \gamma_f^\beta \cdot u_f^j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right) \tag{14}$$

This is already a quite complex model, but most likely not yet adequate. Since the performance of learners after $n+1$ years of learning is usually correlated to a high degree with their performance after n years. This will lead to *autocorrelation* in the errors, violating basic assumptions. To take care of this, model which allow for explicitly specified error structures must be used. While these are well-understood and implemented in R packages for linear models (see Fox 2016 and Zuur et al. 2009), there exists no simple solution for GLM(M)s. Anyone wishing to use such models will have to consult original literature.

2.3 Specifying model using lme4 in R

This section focusses on lme4, the standard package to do multilevel modeling in R with maximum likelihood methods. First, specifying models is discussed.

Second, it is shown how the output can be interpreted.

Varying intercepts Most important are the functions `lmer` and `glmer`, which extend the syntax of `lm` and `glm`. The simple varying intercept model in 2 looks as follows in R. Informative variable names are used instead of Greek letters.

```
glmer(construction ~ discourse + (1 | lemma),
      family = binomial(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 (6) would be specified as follows. Only the formula is given from now on since the rest of the function call remains the same.

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

In other words, 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, the lemma is automatically also modeled as varying. The last formula can therefore be abbreviated to the following equivalent one.

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

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

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

Multiple random effects When there is more than one random effect, several bracketed terms can simply be added. The following is the recommended specification for models like (8), regardless of whether the effects are nested or crossed.

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

Sometimes the recommendation is found to use the following notation for nested random effects, where `semantics` nests `lemma`.

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

In other words, 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 an illustration of what this means in terms of the data structure), lme4 automatically treats it as nested, and the results are exactly the same with

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 4: Illustration of nested factors, organized suboptimally

all the three aforementioned notations. However, the following specification is *not* equivalent and leads to problematic results.

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

This would instruct lme4 to try to estimate the variance of `lemma` not just restricted to specific 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 to 0, while the variance estimate for `semantics : lemma` will be distorted.

There is one situation where the explicit notation for nested factors is necessary. This is when the data are stored in a suboptimal way. The suboptimal version of Table 1 would look something like Table 4. 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 recognizing that the

data structure is nested, and the user has to explicitly provide that information. It would, of course, be better *not* to organize data that way.

Second-level predictors (9) and (10) have the following lme4 syntax.

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

If the data is organized 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. This should be kept in mind when interpreting the results of the estimation.

Finally, and without going into the details for space reasons, we mention that second-level predictors for random slopes are more tricky to specify (see Gelman & Hill 2006: 280-282). Assuming that the effect of discourse status varies with the lemma, which itself comes with a second-level model including frequency as a regressor, the specification looks as follows.

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

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

2.4 Interpreting the output of lme4

Basics and varying intercepts The output for GLMMs in lme4 can be understood straightforwardly after what was said in Sections 2. Here is a possible

output of the summary function for a fit of model (2) and (3), repeated here as (15) and (16). Artificial data was used for the estimation.

$$P(y^i = 1) = logit^{-1}(\alpha_l^{j[i]} + \beta_d \cdot x_d^i) \tag{15}$$

$$\alpha_l^j \sim N(\mu_l, \sigma_l^2) \tag{16}$$

```
Generalized linear mixed model fit by maximum likelihood
 Family: binomial  ( logit )
Formula: construction ~ discourse + (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
discourse1    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 $\hat{\sigma}_l^2$.[7] The estimate for `discourse1` (1.5064) corre-

[7] In general, the notation $\hat{v}$ denotes an estimate of the variable v.

sponds 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 above one, 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 … 2.6433 and the bootstrap with 250 simulations gives $3.9665 \cdot 10^{-6}$ … 1.8023 as the 95% confidence interval. Since the bootstrap (especially with smaller original sample sizes) tends to run into replications where the estimation of the variance fails and is thus output 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 GLMM and a GLMM with an added single random effect can be performed with the `anova` function.

```
alternation.null <- glm(construction ~ discourse,
```

```
                        data = observations,
                        family = binomial(link=logit))
anova(alternation, alternation.null)
```

The GLMM must be the first argument to `anova`. In this case, the output looks like this (AIC and BIC removed for space reasons), indicating a significant effect, although the p-values should not be considered highly reliable.

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

The coefficient of determination (pseudo-R^2) can be computed using the function `r.squaredGLMM` from the `MuMIn` package.

```
library(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)
```

To plot the predictions for the levels of a random effect, packages like `sjPlot` offer customizable functions, as in the following example, which plots Figure 8. Notice that calls the conditional modes BLUPs (best linear unbiased predictors), which is an alternative term for conditional means in linear models.

```
sjp.glmer(alternation, sort.est = "sort.all",
          facet.grid = F, show.ci = T)
```

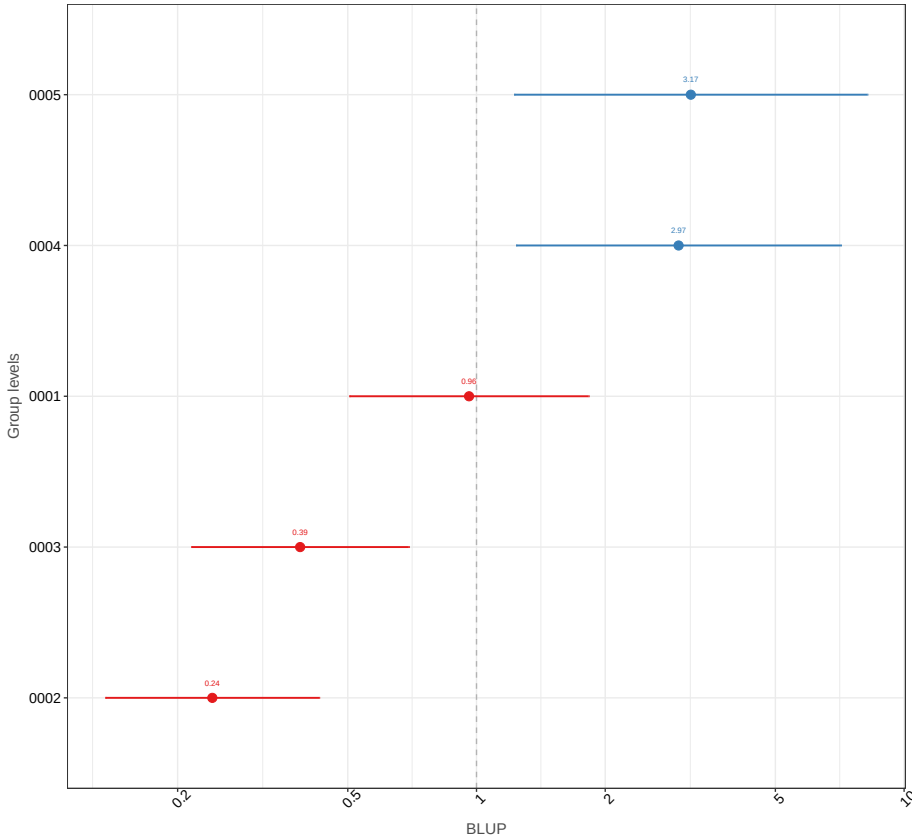

Figure 8: Estimated conditional modes with confidence intervals

Varying intercepts and slopes If we add a varying slope to the model as in (6), the most important thing to look for is the report of the variances and

covariances.

```
Random effects:
 Groups Name        Variance Std.Dev. Corr
 lemma  (Intercept) 0.6103   0.7812
        discourse   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 ***
discourse    -1.0428     0.1492  -6.990 2.75e-12 ***
```

This output tells us that the estimated variance in the intercepts is $\hat{\sigma}^2_\alpha = 0.6103$, the estimated variance in the slopes is $\hat{\sigma}^2_\beta = 0.8944$, and the covariance coefficient estimate is $\hat{\rho} = -0.39$. The means are estimated as $\hat{\mu}_\alpha = 0.6332$ and $\hat{\mu}_\beta = -1.0428$. Compare this to Section 2.2.4), especially (7). It is possible to reconstruct group-wise models from this output and a lookup of the group-specific predictions. For the first group, for example, the following can be done.

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

The output is as follows.

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

This means that for the first lemma, actual predictions for the outcome of the alternation can be made using (17), where values are rounded to two decimal

digits. Compare this to (6).

$$Pr(y^i = 1) = logit^{-1}([0.63 + 0.44] + [-1.04 - 1.23] \cdot x_d^i) \tag{17}$$

2.5 Summary and recommendations for a protocol

In this chapter, readers should have learned that the output of R is transparent when practitioners are able to relate it to (1) the structure of their data set and (2) the specification of the model. As a general protocol, it is recommended that for each study, researchers first examine their data set, then specify a model in mathematical notation *including the variance-covariance matrix* based on their knowledge of the data and the theory behind their study. After the re-specification in R and the estimation, the R output and the model specification can be easily related, and informed inferences can be made.

3 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, 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. They do not give any reason except that otherwise "difficulties" would arise, but it is the advantage of random effects modeling that it can deal with a situation where categories have low numbers of observations (see *shrinkage* in 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 ...

Gries (2015)

Research questions

Data

Method

4 Further reading

This article was written in a way such that readers should be able to move on to text books written by statisticians rather than practitioners. That said, 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) is an obligatory read for users of lme4.

References

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

Fox, John. 2016. *Applied regression analysis & generalized linear models*. 3rd edn. London: Sage Publications.

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.

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.

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.