Linear Mixed Effects Models

Giovanni Cassani

July 22, 2019

CAVEAT: In this explanation of Linear Mixed-effects Models it is assumed that the reader knows about basic statistical concepts (distributions, mean, variance, standard deviation, standard error), linear regression, hypothesis testing (t-test, ANOVA, ANCOVA). If you are not, first learn about these concepts and then get back here: using mixed models without a proper understanding of the aforementioned concepts is strongly discouraged, because you need to know about all that to decide which is the right model given your data and hypothesis, and to correctly interpret the results.

Mixed-effects models, also known by the name of multilevel models, are an extension of traditional linear regression models that have been introduced to better take into account hierarchically organized samples, where observations belong to nested groups. The hierarchy can consist of temporally ordered measurements taken on the same subjects or of measurements taken on subjects that, due to other factors, cannot be said to be independent, e.g. children in a same classroom, trees in a same garden, and the like (see Figure 1).

Mixed-effects models extend more common and traditional statistical tests like the t-test, ANOVA (continuous outcome and categorical predictors), ANCOVA (continuous outcome and mixed - continuous and categorical - predictors) by dispensing with the assumption that all measurements are independent of each other. The paired t-test already goes in this direction but is not very flexible given that it can only be applied when two measurements per subject have been taken. Unlike these traditional methods, mixed models can take the association between observations into account, and model hierarchical (or clustered) data better.

The general equation of a linear mixed model looks like the following one:

$$Y \sim X_1 + X_2 + ... + X_N + (1|RE_1 + ... + RE_K) \tag{1}$$

where Y is a continuous dependent variable and X_{1-N} are continuous or categorical predictors whose effect on Y we want to assess. The Xs constitute the fixed part of the model. The last part of the equation is the most interesting, since it contains the random effects, which account for the correlation between observations. From an experimental point of view, random effects account for the variable part of an experiment: if we were to run an experiment again, certain manipulations would stay constant, since we are interested in them. Other factors, however, would change, for example the subjects. Whatever would stay fixed in a replication should be included as a fixed effect; everything that, on the contrary, would change in a replication is modeled as a random

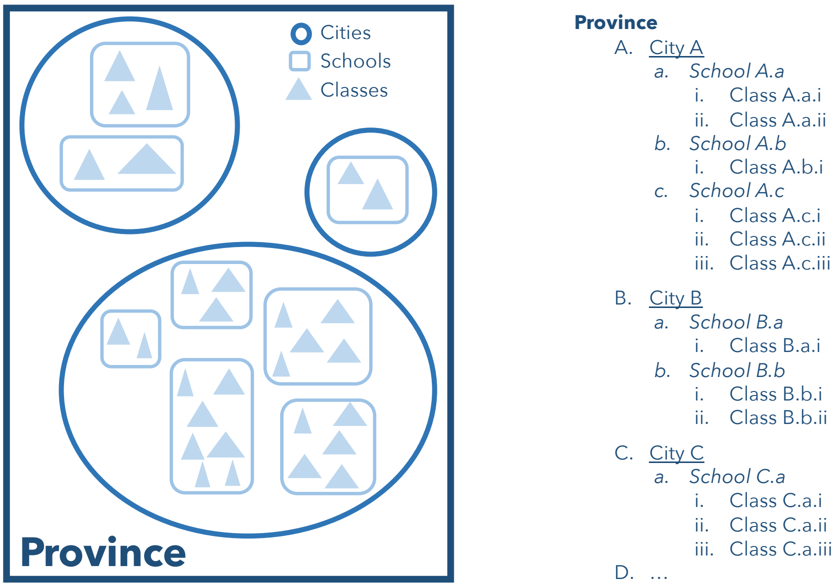

Figure 1: Graphical representation of hierarchical (or clustered) data. in the left panel, the clustered nature of the data is made explicit: classes in a same school in a same city are 'closer together', i.e. more similar than classes from different schools or even from different cities. The right panel show the hierarchy better: classes are nested within schools, that are in turn nested within cities, that belong to a province of interest.

effect. In general, the fixed effects model the expected outcome of Y, the mean of our dependent variable, while the random effects model the variance structure of the data.

1 Random intercepts models

Consider a study on the educational system in a province, like the one pictured in figure 1: cities are randomly sampled in the province; within a city, schools are randomly sampled; within a school, classes are randomly sampled; within a class, children are randomly sampled. The aim of the study is to explore the impact of class size, native language of children, SES, and teacher experience on the literacy of pupils. The same experiment could be run in a different province, with different cities, schools, classes, and children, and yet investigate the same predictors on the same outcome aiming to test the same experimental hypothesis (a typical example of a factor that is always a fixed effect is gender: in every study that investigates both men and women, gender will be constant and thus it doesn't make sense to include it as a random effect). Thus, the nested level of sampling should all be included in a model as random effects, as they would change in a replication study, they are not variables of interest (we don't want to know here whether being in school A improves your literacy, but what factors make children in school A have a higher literacy). Children from the same class are likely to be more correlated than children in the same school but different class and even more than children in the same city but different school. Cities are sampled independently and there is no further level on top of it: thus, by using linear mixed-effects models, we are assuming here that schools are independent. Even though this might not be true (schools in the same country are certainly more correlated than schools from different countries), we are investigating the effects of several predictors on literacy ***in the province***: this means that our conclusions will only apply to that context, but doesn't affect the validity of the model per se, if cities have indeed been randomly sampled from all schools in the province. A random sample is always a crucial assumption to derive robust conclusions from a statistical model.

The model we would fit to analyze this dataset would look like this:

$$Literacy \sim size + lang + ses + exp + (1|city + school + class) \tag{2}$$

This is a simple additive model, but interactions between fixed effects can be included. There is no random effect for child since only one measurement is taken for every child. If our study would have followed children for a certain period of time with multiple measurements per child, our model would have been different, similar to the one in (3):

$$Literacy \sim \mathbf{time} + size + lang + ses + exp + (1|city + school + class + \mathbf{child}) \tag{3}$$

where time is added as fixed effect and child as a random effect.

Mathematically, we can formulate the model in (2) as in equation (4)[1]:

$$Y_{ijkl} \sim \mu + a_i + b_{j(i)} + c_{k(j)(i)} + \beta_1 x1_{ijkl} + \beta_2 x2_{ijkl} + \beta_3 x3_{ijkl} + \beta_4 x4_{ijkl} + \epsilon_{ijkl} \tag{4}$$

where:

- Y_{ijkl} is the literacy value for the lth child, in the kth class from the jth school in the ith city of our sample
- μ is the global intercept estimated from the data, i.e. the average literacy value across our whole sample
- a_i is the random effect for cities, i.e. the estimated change in literacy estimated for the ith city. as are assumed to be drawn from a population of cities and distributed $\sim N(0, sigma^2_{city})$: a_i represent variability between cities, i.e. how much a city average deviates from the overall average
- $b_{j(i)}$ is the random effect for schools, i.e. the estimated change in literacy estimated for the jth school in the ith city. bs are assumed to be drawn from a population of schools and distributed $\sim N(0, sigma^2_{school})$: b_j reflect the variability between schools in a same city, i.e. how much a a school deviates from the average of the city the school is in

[1] that there is no alignment between (2) and (4): when writing models like in (2), random effects are usually places at the end; on the contrary, when writing mathematical formulations, random effects are usually placed right after the global intercept μ

- $c_{k(j)(i)}$ is the random effect for classes, i.e. the estimated change in literacy estimated for the kth classroom in the jth school in the ith city. bs are assumed to be drawn from a population of classrooms and distributed $\sim N(0, sigma^2_{class})$: c_k represent the variability between classes within a same school, i.e. how much a class deviates from the average of the school it belongs to

- β_{1-4} are the estimated coefficients for each of the four predictors included in the model ($x_1 = size, x_2 = lang, x_3 = ses, x_4 = exp$. Each beta is multiplied by the value of the matching predictor for the lth child from the kth classroom in the jth school in the ith city, and gives the estimated difference in literacy for each 1-unit change in a predictor (for continuous predictors) or for a different level of the predictor (for categorical ones) controlling for all other predictors

- ϵ_{ijkl} is the residual variance that is not explained by the fitted statistical model, i.e. the difference between each observed literacy score and the corresponding predicted literacy score after all other parameters have been estimated. Errors are assumed to be normally distributed $\sim N(0, \sigma^2_{res})$

In (4), $a_i + b_{j(i)} + c_{k(j)(i)}$ are referred to as **random intercepts**, as they represent the deviation of the intercept of a specific city, school in a city, and class within a school in a city, from the average literacy value (global intercept, or Y value when all other parameters are 0) estimated from the sample. By allowing variation around the estimated intercept, we acknowledge that certain classes might do better and other worse, and that this has an effect on the scores of individual children, that cannot be assumed to be independently sampled. The same holds true for all other cluster levels.

Random intercept models estimate one regression line for each cluster, each with its own intercept but all with the same slope. In a way, random intercept mixed effect models can be seen as cluster-specific regression models, i.e. a regression model with cluster-specific regression parameters.

A bunch of important assumptions are hidden in this model: first of all, the coefficients of the random effects (a,b,c) are assumed to be normal for mathematical convenience, but this might not be the case in the data. Dispensing with this assumption however makes the estimation process problematic, and is not considered here also because violations of it don't affect the estimation of fixed effects. A second important assumption is the independence of ϵ_{ijkl} and all random coefficients: if the prediction error is consistently larger for certain observations with specific random effects values, the model is not reliable. A third, crucial assumption is that the variance is constant for all values of the predictors. This assumption is the easiest to violate, since it might well be the case that the variance is larger for larger values of a predictor. Luckily enough, there are ways to increase the model flexibility to model situations of unequal variance for different values of a fixed effect. Moreover, since clusters at the highest level are assumed to be independent, the covariance between observations from different clusters is assumed to be 0: in our example, two schools from different cities are independent, and so are two classes from two schools in different cities, all the way down to children. On the contrary, two schools from the same city have covariance equal to σ^2_{city}. If we go down the hierarchy to classes from two schools in the same city, their covariance is equal to $\sigma^2_{city} + \sigma^2_{school}$, and so on.

Within a same cluster, variance only comes from error terms and the variance at the lower levels. Thus, at the class level, variance that is not explained by the fixed effects only results from errors between predicted and observed values; at the school level, however, variance comes from the error terms and from the variance at the class level, and so on, up to the highest-level cluster. Given the assumption of independence between random effects at all levels and error terms, we can simply add variances from the lower level to obtain the variance at the higher level. Formally:

$$\begin{aligned} Var(Y_{ijkl}) &= Var(\mu + a_i + b_{j(i)} + c_{k(j)(i)} + \beta_1 x1_{ijkl} + ... + \beta 4 x4_{ijkl} + \epsilon_{ijkl}) \\ &= Var(a_i + b_{j(i)} + c_{k(j)(i)} + \epsilon_{ijkl}) \\ &= Var(a_i) + Var(b_{j(i)}) + Var(c_{k(j)(i)}) + Var(\epsilon_{ijkl}) \\ &= sigma^2_{city} + sigma^2_{school} + sigma^2_{class} + sigma^2_{res} \end{aligned} \tag{5}$$

μ and all the βs don't affect variance because they are constant, they don't depend on the cluster. Thus, the estimate of the variance is fully additive and consists of the sum of the estimated variance at all cluster levels. This total variance is decomposed in two main parts, **within-cluster** and *between-cluster variability*: the latter corresponds to the residual variance $sigma^2_{res}$, while the former includes all variance components relative to the random factors, here $sigma^2_{city} + sigma^2_{school} + sigma^2_{class}$. Thus, we can compute the proportion of the total variance that is explained by every cluster: the higher this proportion, the more important the random effect. Typically, the closest random effect to the observations, *classes* in our example, is the random effect that explains the highest proportion of variance. This proportion decreases when considering higher-level clusters. This pattern makes intuitive sense: two observations from the same class are more correlated than two observations from a same school, which are even more correlated than two observations from the same city. Including a child random effect in a longitudinal study would make it even clearer: measuring a child multiple times would result in measurements being highly correlated, much more than two measurements taken for two different children.

Variances are also used to determine the intraclass correlation, that determines the correlation between observations belonging to the same cluster. At the highest cluster level, intraclass correlation is computed as in (6):

$$\rho_{I(city)} = \frac{\sigma^2_{city}}{\sigma^2_{city} + \sigma^2_{school} + \sigma^2_{class} + \sigma^2_{res}} \tag{6}$$

The intraclass correlation between two observations from the same city is thus proportional to the variance across cities, and inversely proportional to the variance across schools within the same city, the variance across classes within the same school, and the residual error.

If we want to see how correlated are observations from our lowest level of clustering, then, straightforwardly, we use the formula in (7):

$$\rho_{I(city)} = \frac{\sigma^2_{city} + \sigma^2_{school} + \sigma^2_{class}}{\sigma^2_{city} + \sigma^2_{school} + \sigma^2_{class} + \sigma^2_{res}} \tag{7}$$

Intuitively, two observations from the lowest cluster level are much more correlated than two observations generally belonging to the highest cluster level. Moreover, when we have fitted our model, the only variance component that is left out between two observations from the same class (lowest level of clustering) in the residual error. Essentially, the equation in (7) tells that the correlation between two children in the same class is proportional to the variance explained by the random component of our model and inversely proportional to the residual error that cannot be accounted for by the model itself. Lastly, this formula makes it clear that if the variance of a random effect is 0, it is useless to include that random effect in the model, since it doesn't contribute to explain any further variability in the data.

Importantly, intraclass correlation and Pearson's r are not the same, since Pearson's r (unlike intraclass correlation) doesn't assume variances to be constant within a same cluster. This assumption needs to be checked, and if it is not the case that, e.g., classes in a same school have constant variance, the model should be specified differently or changed altogether.

After describing the variance structure of the data and detailing how it can be modeled using random effects, we turn to the estimate of the outcome variable of interest, i.e. the literacy of pupils. The expected value of our outcome Y_{ijkl} can be derived as shown in (8) where Y_{ijkl} is again the literacy value for the lth child, in the kth class from the jth school in the ith city of our sample:

$$\begin{aligned} E(Y_{ijkl}) &= \mu + a_i + b_{j(i)} + c_{k(j)(i)} + \beta_1 x1_{ijkl} + ... + \beta 4 x4_{ijkl} + \epsilon_{ijkl} \\ &= \mu + \beta_1 x1_{ijkl} + ... + \beta 4 x4_{ijkl} \end{aligned} \tag{8}$$

The random component $a_i + b_{j(i)} + c_{k(j)(i)}$ and the error ϵ_{ijkl} go away because they have expected value, or μ, equal to 0 by assumption. Thus, they don't affect the expected value of our dependent variable Y. From (5) and (8) it becomes clear that the fixed effects model the estimated value of our dependent variable, while the random effects model the variance structure of the dependent variable itself. If we fitted a simple linear model, that only included the fixed part of our model, we would have assumed a much more fixed structure for the variance, due to the assumption that all observations were independent. When they are not, including random effect makes the model much more flexible.

2 Random slopes models

We have seen in the previous section that random intercepts models allow each cluster to have its own intercept, but variance is assumed to be constant across clusters: thus, one regression line is estimated for the each cluster in the dataset, each crossing the Y axis at a different point but all having the same slope. This follows the assumptions of constant variance and constant within-cluster correlation. These assumptions, however, might be too strict. Consider a longitudinal analyses as the one sketched in equation (3): in one class, the literacy score is very high already at start, thus there is not much room for improvement over time. In another class, however, the literacy score was very low at start, and a big leap forward happens with time. The two classes, two clusters in our mixed-effects model, have two different intercepts, i.e. literacy values at start, but similar literacy scores at the end, but also have two different slopes. Importantly, class is not a factor in our analysis: if we replicated the study, we would have different classes, so we include it as a random effect: we are not interested in which classes evolve faster/slower within our sample.

Another problem for random intercepts models could come from situations where the variance within a cluster is not constant, but varies with time. This can often happen when there is a large average numeric increase in the outcome variable. A straightforward example is human height measured at birth and progressively with age. At birth, variability in height is low, while it increases with time because certain children might turn out very tall and other stay relatively short. Differences of 5cm at birth might increase to 20 or more later in development, and these differences might be correlated with some factor of interest (a fixed effect in the model). Thus, the variability in height at start is not the same as the variability at the end. A simple random intercepts model wouldn't be able to account for these situations, thus we need to extend it to also allow clusters to have cluster-specific slopes.

Consider now a simplified version of the literacy score analysis we have used so far: now, instead of having 4 nested clusters (city, school, class, child), we have a longitudinal study involving children sampled randomly across the schools of a province. This time, there is no hierarchical structure in our data: children are truly randomly sampled from the population of interest. However, we take multiple measurements per child, and measurements from the same child are certainly more correlated than measurements from different children, violating the assumption of traditional linear models that all observations are independent. For this reason we still use a mixed-effects model.

More specifically, we want to assess the effect of SES, which is categorical and can take three values, *low, medium, high*, on literacy score. We are interested in testing a) whether SES affect the literacy of children when they arrive at school and b) whether children from different SES show the same improvement in literacy, i.e. we want to test whether there is an interaction between time and SES on literacy score. Our model will look like the one in equation (9):

$$Y_{ij} = \begin{cases} (\beta_1 + b_{1i}) + (\beta_4 + b_{2i})t_j + \epsilon_{ij} \\ (\beta_2 + b_{1i}) + (\beta_5 + b_{2i})t_j + \epsilon_{ij} \\ (\beta_3 + b_{1i}) + (\beta_6 + b_{2i})t_j + \epsilon_{ij} \end{cases} \tag{9}$$

where

- Y_{ij} is the jth literacy score taken at time t_j for the ith cluster, i.e. child
- $\beta_1, \beta_2, \beta_3$ are the SES-level intercepts, i.e. the estimated average literacy score for each SES level at time 0
- b_{1i} models the cluster-specific variability in the literacy score, i.e. the deviation from the overall intercept for each cluster within each SES level
- $\beta_4, \beta_5, \beta_6$ are the SES-level slopes, i.e. the slopes of the regression lines estimated separately for each SES level

- b_{2i} models the cluster-specific variability in the change in literacy score over time, i.e. the different evolution of literacy scores for each cluster within each SES level
- t_j is the jth observation
- ϵ_{ij} is the usual residual error resulting from the difference between predicted and observed value for the literacy score of the ith child at the jth time point

Once we have fitted this model to our data, we can compute the literacy score of a child by filling in the relevant values in the model. Suppose that the first line from (9) models low SES children, the second line from (9) models medium SES children, and the third line from (9) high SES children. Thus, to obtain the literacy score at time 7 for child 4 from the high SES group, we take the estimate for β_3, the estimated difference in intercept for child 4 wrt to the high SES average, i.e. b_{14}, the estimated group slope for high SES children, $\beta 6$, the estimated difference in slope for child 4 wrt the high SES overall slope, i.e. b_{24}, at time t_7. This gives the predicted literacy score. To get the observed one we finally add the residual error ϵ_{47}. This is summarized in equation (10):

$$Y_{47} = (\beta_3 + b_{14}) + (\beta_6 + b_{24})t_7 + \epsilon_{47} \tag{10}$$

Going back to the hypotheses of interest we sketched before, we want to see whether children of different SES have different literacy score at the beginning of the period of interest and we want to see whether they improve in a similar way. Looking at our model, we want to test whether $\beta_1 = \beta_2 = \beta_3$ to answer the first question (is there a difference in starting literacy score depending on the child SES?) and $\beta_4 = \beta_5 = \beta_6$ to answer the second question (does SES level affect the time evolution of literacy score?). Importantly, the second hypothesis only looks at potential differences in the evolution, but is not concerned with testing whether an evolution was there at all. This second test would take the form $\beta_4 = \beta_5 = \beta_6 = 0$.

As is the case with random intercepts models, random effects are assumed to be normally distributed with mean 0 and cluster specific variance, and the residuals are again assumed to be normally distributed with $\mu = 0$ and $\sigma^2 = \sigma^2_{res}$, independent of the random effects. Unlike random-intercepts models, however, our random effects come from a multivariate normal distribution

$$b_i = (b_1, b_2)' \sim N(0, D) \tag{11}$$

where D is a symmetric 2x2 co-variance matrix, since we have 2 interdependent random effects in this example (and a bi-variate distribution). The main diagonal of the matrix represents the variance of the random effects, with D_{11} being the variance of b_{1i}, i.e. the group specific random intercepts, and D_{22} being the variance of b_{2i}, i.e. the group specific random slopes. Since the matrix is symmetric, $D_{12} = D_{21}$: the number in these cells represent the co-variance between b_{1i} and b_{2i}, i.e. between the intercepts and the slopes. From this, we can derive the correlation between cluster-specific intercepts and cluster-specific slopes, as specified in equation (12):

$$Corr(b_{1i}, b_{2i}) = \frac{d_{12}}{\sqrt[2]{d_{11}} \cdot \sqrt[2]{d_{22}}} \tag{12}$$

The random intercepts model can be now seen as a special case of the random-slopes model, where random intercepts and random slopes don't co-vary. If the off-diagonal cell in the co-variance matrix D is 0, our bi-variate distribution governing (b_{1i}, b_{2i}) amounts to two separate uni-variate distributions for b_{1i} and b_{2i}, both with $\mu = 0$ and their own cluster specific variance. In general, when the variance of a random effect is 0, the random effect is not needed: we have shown that the random effects are required to model the variance component of the data, thus, if a random effect doesn't account for any variance in the data, it is not required in the model. A good way of checking if the random structure of a model is actually accounting for the variance of the data, is to plot the observed variance versus the estimated one. If they look radically different, then the model is not good for the data. For the sake of explanation, let's assume that our children have very similar literacy values at the beginning of the study and also at the end. They start around the same score and end up around the same score. However, some improve a lot at the beginning, some other improve a lot towards the end, and some other improve more gradually. In this situation, we have low variance at the extremes, and high variance in between. If we plot it, we see a sort of bell-shape. We know that to model the variability in the data in mixed-effects models we need to modify the random part. We have already included a random intercept for children (our clusters) and a random slope for time. What can we include further? The idea is to extend the model by including a quadratic factor of time that affects our random slopes. Quadratic functions have the aspect of a parabola, which makes sense to model a variance that starts low, increases and then goes down again. Other options are exponential and logarithmic transformations, for situations where the increase is monotonic but not linear.

When we change the random part of a model, we see that parameter estimates stay constant - which we would indeed expect since those depend on the fixed effects. What changes, beyond the variance components, are the standard errors of those estimates, and this affects the p-values (and possibly the conclusions we derive from the data). Thus, correctly modeling the variance structure of the data is crucial for reliable and robust hypothesis testing.

3 Estimating Random Effects

By now it should be clear that random effects, both intercepts and slopes, reflect cluster specific variation with respect to the average value of the dependent variable in the case of intercepts and with respect to the change in the dependent variable in the case of slopes. We know that random effects are assumed to be normally distributed with mean 0 and cluster specific variance. From equation (10) we also know that to get the expected value of the dependent variable for a specific cluster at a specific time point, we need to estimate both fixed and random effects, and sum them together according to the model formula.

Since random effects are samples from a normal distribution, Bayesian methods are applied: for this reason, estimates for random effects are called **Empirical Bayes estimates** and are the expected variation of a cluster wrt the general mean given the observed data for that cluster. Importantly, we have said that random effects don't model experimentally relevant aspects of the data. Then why should we care about how much a cluster deviates from the average expected value of our dependent variable? After all, knowing that a child has higher literacy scores than another child doesn't really mean anything to us. We want to know whether SES has a systematic effect on literacy, and we include random effects because we expect observation from the same child to be correlated, not because we want to know the identity of the best/worst child. Still, estimating random effects can be a useful diagnostic tool to *detect outliers.* If we see anomalies in estimated random effects such that in order to model certain observations we need a disproportionately large random effect, we can infer that observations from the corresponding cluster might be problematic. Alternatively, we might realize that our random effects are not normally distributed. However, estimates for the fixed effects are very robust wrt the violation of this assumption provided that there is a sufficient number of independent clusters, i.e. the highest order hierarchical level (it doesn't really matter how many observations per cluster). Still, it makes sense to know which assumptions are violated, and even more to know whether there are outliers in the data.

4 Wrapping up

A linear mixed-effects model is essentially a linear regression with two sets of parameters:

- βs modeling the fixed effects, i.e. the experimentally relevant predictors and their effect on the average value of continuous dependent variable of interest

- $b_i \sim N(0, D)$ modeling the random effect, i.e. the subject-specific variability in the data and the co-variance structure. D is the (symmetric) co-variance matrix and its dimensionality is the same as the number of random effects included in the model. When the matrix is diagonal, with off-diagonal cells all being 0, the model is a simple random intercepts model and the multivariate distribution boils down to a combination of independent, uni-variate distributions with $\mu = 0$ and cluster-specific variance. When the matrix is not diagonal, with off-diagonal cells being different from 0, we have a random slopes model, where each cluster not only has a specific intercept value, but also a different change than the average change. In random-slopes models, the random effects interact and co-vary. Going back to the example on literacy (equation (9)), we have a situation where children's improvement in literacy depends on the SES level, modeled as a fixed effect, but is also child-specific, modeled by the random slopes, i.e. a deviation in the slope of a child from the slope of its SES group. Moreover, there is systematic co-variance between a child's intercept value and the slope of his/her improvement in literacy.

5 Missing values

Mixed models are very robust to missing values, provided that these are not systematic and don't depend on the study design. An example of problematic missing values might be a study on cancer treatment where the worst patients die before the end of the study, or a study about depression where the best patients drop out of the program because they feel good. In these cases, missing values can severely bias the estimated model. If, however, values are missing randomly, mixed models are much better than other statistical methods. They counter missing values by using the estimated weights to predict missing values, and hypothesis testing is carried out on a full dataset, where what was missing have been estimated using the model itself. This makes it even clearer that systematic missing values can cause huge biases. Go back to the depression example: at start, we have a bunch of patients with different levels of depression. We start our treatment, randomizing patients in the different conditions. Suppose that our treatment works very well: after a while, patients from the treatment group will start feeling well and will stop coming to the lab for tests. This is a first bias: missing values will come from the treatment group much more than from the control group. But there's another bias lurking: it is likely that people dropping out would be those that came in with lower levels of depression. Thus, we would start with a nice, randomized sample, randomly divided in control and treatment group, each with comparable average values of depression and comparable variances. But we would end up with a control group that stayed pretty much the same, and a treatment group where people feeling better were progressively dropping out. This would leave us with observations at later time points being worse than expected if the whole sample was measured. And thus we would be using biased depression scores to predict the depression scores of healthier people for which we stopped having measurements at some point in time. This could even result in not detecting an effect of our treatment even though there was one.