Mixed effect models for interactions and cross-cultural variation: a conceptual primer

Luke Maurits <luke_maurits@eva.mpg.de>

07/04/2021

Introduction

This document attempts to clear up what I perceive to be common points of confusion regarding, and to emphasise underappreciated aspects of, *mixed effect models*, i.e. models which contain both *fixed effects* and *random effects*. It is fairly wide-ranging in scope, but in particular tries to address these points:

- Thinking about "fixed effects" and "random effects" as two separate kinds of things is not the best way to understand these models and what they are really doing.
- Random slopes are a kind of interaction between variables - yes, your model can have interactions even if there's no `*` in the formula!
- Almost nobody talks about "random effects" in a way that makes senses, and in particular saying something like "we included a random effect of culture" is almost always woefully ambiguous.
- It's not really all that important whether the different grouping variables in a mixed effect model are nested or crossed.

The document is more aimed at conceptual understanding of what mixed effects models *actually do* rather than a practical tutorial on how to run and report such analyses (which I hope to address in separate documents).

A quick note on terminology

A lot of people are starting to push back against the "fixed effects" and "random effects" terms (and hence, as a necessary consequence, "mixed effects" too), pointing out that they are often used inconsistently. Instead, these people promote terms like "multi-level" or "hierarchical" modelling. I'm very sympathetic to the idea that the standard terminology is less than ideal - as this document will make clear - but at the same time I think the logical interpretation of "multi-level modelling" refers to a much broader class of models than I'm going to discuss here. Traditional mixed effect models are just one special, simple case of multi-level modelling. Since this document isn't going to talk about anything more general, I'm going to pinch my nose and stick with the traditional terms.

What actually are random effects, anyway?

Let's start from a friendly and familiar model with only fixed effects in it, say a logistic regression which tries to predict the probability that a survey respondent answers "yes" to a yes/no question, using age as an explanatory variable. Don't panic if you aren't familiar with logistic regression - random effects work the same way no matter what kind of response distribution you're using, as they are all about adding extra terms to the linear predictor. So, you might write this model using base R's `glm` function something like this:

```
glm(y ~ 1 + age, family=binomial)
```

which corresponds, under the hood, to this:

$$y_i \sim \text{Bernoulli}(p_i) \tag{1}$$
$$p_i = \text{logit}^{-1}(\eta_i) \tag{2}$$
$$\eta_i = \beta_0 + \beta_1 x_i \tag{3}$$

Where β_0 is the intercept, β_1 is the slope of age, and x_i is the i-th participant's age.

Suppose you've distributed carefully translated versions of this survey to participants in different countries, and you suspect that participants with the same cultural background might tend to give more similar answers to one another, which you'd like to account for in your model (rather than treating all participants with the same age as totally independent from one another, which is what the simple `glm` model above does). You know that this is what random effects are for, so you update your model by switching to the `lme4` package and using:

```
glmer(y ~ 1 + age +  (1 + age|culture), family=binomial)
```

What does this change under the hood? The most important change to grasp in the beginning is that the linear estimator, which changes from equation 3 to:

$$\eta_i = (\beta_0 + u_{0c_i}) + (\beta_1 + u_{1c_i})x_i \tag{4}$$

Here c_i is the culture of the i-th participant. Each participant's data is modelled using an intercept which is equal to β_0 plus some extra term u_{0c_i} which represents the deviation from β_0 for their particular culture. And a similar story holds for the slopes. To make this a little more concrete, suppose you have data for 9 participants, three from Germany, three from the UK and three from Vanuatu. The full set of linear predictor terms is:

$$\eta_1 = (\beta_0 + u_{0\text{DE}}) + (\beta_1 + u_{1\text{DE}})x_1 \tag{5}$$
$$\eta_2 = (\beta_0 + u_{0\text{DE}}) + (\beta_1 + u_{1\text{DE}})x_2 \tag{6}$$
$$\eta_3 = (\beta_0 + u_{0\text{DE}}) + (\beta_1 + u_{1\text{DE}})x_3 \tag{7}$$
$$\eta_4 = (\beta_0 + u_{0\text{UK}}) + (\beta_1 + u_{1\text{UK}})x_4 \tag{8}$$
$$\eta_5 = (\beta_0 + u_{0\text{UK}}) + (\beta_1 + u_{1\text{UK}})x_5 \tag{9}$$
$$\eta_6 = (\beta_0 + u_{0\text{UK}}) + (\beta_1 + u_{1\text{UK}})x_6 \tag{10}$$
$$\eta_7 = (\beta_0 + u_{0\text{VA}}) + (\beta_1 + u_{1\text{VA}})x_7 \tag{11}$$
$$\eta_8 = (\beta_0 + u_{0\text{VA}}) + (\beta_1 + u_{1\text{VA}})x_8 \tag{12}$$
$$\eta_9 = (\beta_0 + u_{0\text{VA}}) + (\beta_1 + u_{1\text{VA}})x_9 \tag{13}$$

Things are written out here the way the model views things, with separate *fixed effects* (β_0 and β_1) and *random effects* (the various u_{0c_i} and u_{1c_1} terms). But it's conceptually simpler to collapse this distinction. If we define, say, $\beta_{0\text{DE}} = \beta_0 + u_{0\text{DE}}$, and $\beta_{1\text{DE}} = \beta_1 + u_{1\text{DE}}$, and so on for the other cultures, then everything above simplifies to:

$$\eta_1 = \beta_{0\text{DE}} + \beta_{1\text{DE}} x_1 \tag{14}$$
$$\eta_2 = \beta_{0\text{DE}} + \beta_{1\text{DE}} x_2 \tag{15}$$
$$\eta_3 = \beta_{0\text{DE}} + \beta_{1\text{DE}} x_3 \tag{16}$$
$$\eta_4 = \beta_{0\text{UK}} + \beta_{1\text{UK}} x_4 \tag{17}$$
$$\eta_5 = \beta_{0\text{UK}} + \beta_{1\text{UK}} x_5 \tag{18}$$
$$\eta_6 = \beta_{0\text{UK}} + \beta_{1\text{UK}} x_6 \tag{19}$$
$$\eta_7 = \beta_{0\text{VA}} + \beta_{1\text{VA}} x_7 \tag{20}$$
$$\eta_8 = \beta_{0\text{VA}} + \beta_{1\text{VA}} x_8 \tag{21}$$
$$\eta_9 = \beta_{0\text{VA}} + \beta_{1\text{VA}} x_9 \tag{22}$$

In other words, all we're doing is giving each culture its own intercept and its own slope.

A strange familiarity...

At this point, you might be wondering how this mixed effects model actually differs in any meaningful way from a simpler analysis which just adds a fixed effect of culture. After all, if you run a fixed effects model like this:

```
glm(y ~ 1 + age*culture, family=binomial)
```

you probably have a clearer understanding of what will happen. You'll be fitting a model like this:

$$\eta_i = \beta_0 + \Delta\beta_{0\text{UK}} x_{i\text{UK}} + \Delta\beta_{0\text{VA}} x_{i\text{VA}} + (\beta_1 + \Delta\beta_{1\text{UK}} x_{i\text{UK}} + \Delta\beta_{1\text{VA}} x_{i\text{VA}}) x_i \tag{23}$$

Here β_0 and β_1 are the intercept and slope for the "reference culture", in this case Germany. We've added dummy variables $x_{i\text{UK}}$ and $x_{i\text{VA}}$ which take values of 0 or 1 depending on which culture participant i is from, and each of the two non-reference cultures gets its own intercept and slope which is added to the corresponding values for the reference cultures. I've named these terms to start with Δ, which is a common notation for referring to things which are a difference in value from starting point, since that's exactly what these are. Writing it all out in full, we have:

$$\eta_1 = \beta_0 + \beta_1 x_1 \tag{24}$$
$$\eta_2 = \beta_0 + \beta_1 x_2 \tag{25}$$
$$\eta_3 = \beta_0 + \beta_1 x_3 \tag{26}$$
$$\eta_4 = (\beta_0 + \Delta\beta_{0\text{UK}}) + (\beta_1 + \Delta\beta_{1\text{UK}}) x_4 \tag{27}$$
$$\eta_5 = (\beta_0 + \Delta\beta_{0\text{UK}}) + (\beta_1 + \Delta\beta_{1\text{UK}}) x_5 \tag{28}$$
$$\eta_6 = (\beta_0 + \Delta\beta_{0\text{UK}}) + (\beta_1 + \Delta\beta_{1\text{UK}}) x_6 \tag{29}$$
$$\eta_7 = (\beta_0 + \Delta\beta_{0\text{VA}}) + (\beta_1 + \Delta\beta_{1\text{VA}}) x_7 \tag{30}$$
$$\eta_8 = (\beta_0 + \Delta\beta_{0\text{VA}}) + (\beta_1 + \Delta\beta_{1\text{VA}}) x_8 \tag{31}$$
$$\eta_9 = (\beta_0 + \Delta\beta_{0\text{VA}}) + (\beta_1 + \Delta\beta_{1\text{VA}}) x_9 \tag{32}$$

But we can simplify this a bit too. Since Germany is the reference culture here, let's rename β_0 to $\beta_{0\text{DE}}$ and β_1 to $\beta_{1\text{DE}}$. Then, we can define $\beta_{0\text{UK}} = \beta_{0\text{DE}} + \Delta\beta_{0\text{UK}}$ and so on - this is a lot like how we simplified the mixed-effects model earlier, if we think of β_0 and β_1 as being like the fixed effects and the terms like $\Delta\beta_{0\text{UK}}$ as being like the random effects. Then we end up, once again with:

$$\eta_1 = \beta_{0\text{DE}} + \beta_{1\text{DE}} x_1 \tag{33}$$
$$\eta_2 = \beta_{0\text{DE}} + \beta_{1\text{DE}} x_2 \tag{34}$$
$$\eta_3 = \beta_{0\text{DE}} + \beta_{1\text{DE}} x_3 \tag{35}$$
$$\eta_4 = \beta_{0\text{UK}} + \beta_{1\text{UK}} x_4 \tag{36}$$
$$\eta_5 = \beta_{0\text{UK}} + \beta_{1\text{UK}} x_5 \tag{37}$$
$$\eta_6 = \beta_{0\text{UK}} + \beta_{1\text{UK}} x_6 \tag{38}$$
$$\eta_7 = \beta_{0\text{VA}} + \beta_{1\text{VA}} x_7 \tag{39}$$
$$\eta_8 = \beta_{0\text{VA}} + \beta_{1\text{VA}} x_8 \tag{40}$$
$$\eta_9 = \beta_{0\text{VA}} + \beta_{1\text{VA}} x_9 \tag{41}$$

i.e. we've given each culture its own intercept and its own slope. This is, in fact, exactly the same model structure we saw earlier for the mixed effects model - compare Equations 14 - 22 to Equations 33 - 41. It's just that the various per-culture terms are built up in different ways from other parts of the model. Here it is all laid out in a table:

Table 1: Corresponding fixed and mixed effects model terms

Per-culture term	**Fixed effects form**	**Mixed effects form**
$\beta_{0\text{DE}}$	β_0	$\beta_0 + u_{0\text{DE}}$
$\beta_{0\text{UK}}$	$\beta_0 + \Delta\beta_{0\text{UK}}$	$\beta_0 + u_{0\text{UK}}$
$\beta_{0\text{VA}}$	$\beta_0 + \Delta\beta_{0\text{VA}}$	$\beta_0 + u_{0\text{VA}}$
$\beta_{1\text{DE}}$	β_1	$\beta_1 + u_{1\text{DE}}$
$\beta_{1\text{UK}}$	$\beta_1 + \Delta\beta_{1\text{UK}}$	$\beta_1 + u_{1\text{UK}}$
$\beta_{1\text{VA}}$	$\beta_1 + \Delta\beta_{1\text{UK}}$	$\beta_1 + u_{1\text{VA}}$

So, we've seemingly taken two different routes to end up at more or less the same place. What's going on here?

Same same, but different

Well, there *are* differences between these two models, and they're important, and we'll take a look at them shortly. But before moving on, it's important to understand that despite the differences, there really *is* also a sense in which these two models:

```
glmer(correct ~ 1 + age + (1 + age|culture))
glm(correct ~ 1 + age*culture)
```

are doing broadly the same thing. They are both models where each individual culture will get its own intercept and its own age slope. In particular, you should appreciate that random slopes are just another way to formulate an interaction between two explanatory variables. Even though the first model above has no `*` in its formula, there's absolutely still an interaction between age and culture going on. This highlights an important principle to keep in mind: the syntax you use to specify a model to a package like `lme4` or `brms` is abstracted quite far away from the actual statistical structure of the model. Small differences in syntax can reflect substantial model changes, and large differences in syntax can obscure deep underlying similarities. This is the great pedagogical value of the syntax used in Richard's `rethinking` R package: there's no abstraction at all and you specify the actual model directly.

Once you appreciate that random slopes are a kind of interaction, you can also see how the way a lot of people talk about random effects makes no sense. Nobody would ever say "we included an interaction of sex" or "we included an interaction of culture", because interactions fundamentally include *two* (or more!) variables. It's obvious that you have to say something like "we included an interaction *between* sex and culture" for your statement to mean anything. And yet, people will often say things like "we included a random effect of

culture" and nothing more, as if that meant something by itself. However, this kind of language is in fact completely ambiguous without specifying what it is that varies by culture! If you're talking about random slopes, then you're talking about an interaction, and you need to specify *both* parts of the interaction for it to be clear what your model is actually doing. If you're talking about random intercepts, then there is only one variable involved, but in that case you should explicitly say "we included a random *intercept* of culture", so your readers know what you're saying. The only time that ambiguous phrasing like "we included a random effect of culture" can be correctly understood from context is when you have a very simple model which *only* includes an intercept and nothing else. That's rare, so this kind of imprecise language should be rare also, but unfortunately it's actually the norm!

Identifiability and constraint

So, what is the actual difference between these two models? It comes down to how the random effect terms, like u_{0c_i} and u_{1c_i} are fitted. There's an important additional aspect of a mixed effects model beyond simply changing the linear predictor, which we haven't talked about yet. We'll get to it very shortly, but first I want to take a quick detour to explain *why* there necessarily has to be something more involved. This will introduce you to a very useful concept which applies in many other contexts.

Remember that when simplifying the mixed effects model above, we redefined things like this:

$$\beta_{0\text{DE}} = \beta_0 + u_{0\text{DE}} \tag{42}$$
$$\beta_{0\text{UK}} = \beta_0 + u_{0\text{UK}} \tag{43}$$
$$\beta_{0\text{VA}} = \beta_0 + u_{0\text{VA}} \tag{44}$$

(and similarly for the slopes). The problem with this is that there is no unique way to decompose these per-culture intercepts into a common fixed effect β_0 and per-culture random effects. Suppose β_0 was 0.20 and $u_{0\text{DE}}$ was -0.05, so that $\beta_{0\text{DE}} = 0.20 - 0.05 = 0.15$. We could just as easily add 0.50 to β_0, making it $0.20 + 0.50 = 0.70$, and subtract the same amount from $u_{0\text{DE}}$, making it $-0.05 - 0.50 = -0.55$. This pairing would result in exactly the same value of $\beta_{0\text{DE}} = 0.70 - 0.55 = 0.15$. In fact, we can add *any* amount we like to β_0, and so long as we subtract the same amount from all the u_{0c_i} terms, then nothing changes overall. We can think of arbitrarily many different values of these parameters which will all yield the same likelihood or posterior probability for a given dataset. If this is true, how can we fit the model at all?!

This situation is referred to as a problem with *model identifiability*: as it stands, the model outlined above is said to be *unidentifiable*.

Note that this isn't a problem for the simple fixed effects model. The value of β_0 is constrained by the fact that it, and it alone, has to explain the intercept of all the German participants. Once β_0 is fixed by that requirement, the $\Delta\beta_0$ terms can't vary freely either.

The true nature of random effects revealed!

The reason random effects models *can* be fit, at least some of the time, is that there is an important constraint placed upon the value of the random effect terms. The *full* form of the model is this:

$$\eta_i = (\beta_0 + u_{0c_i}) + (\beta_1 + u_{1c_i})x_i \tag{45}$$
$$u_{0c_i} \sim \mathcal{N}(0, \sigma_0) \tag{46}$$
$$u_{1c_i} \sim \mathcal{N}(0, \sigma_1) \tag{47}$$

That is, the random effects are assumed to be drawn from a normal distribution with a mean of 0 (this is all that the fancy $\sim \mathcal{N}$ notation in Equations 46 and 47 above means). The random intercepts and the random

slopes come from separate distributions, each with their own standard deviations, which are estimated from the data (and in a Bayesian setting from prior distributions, too). This solves the identifiability problem - we can't just move β_0 around freely and shift the u_{0c_i}s in the opposite direction to compensate, because shifting the random effects all in the same direction at once would shift their mean away from zero, which this model will fight back against. Note that fixing the mean at *any* value would have removed this freedom and allowed us to fit unique estimates to the parameters, but fixing the mean at zero in particular is in some sense natural, and here's why.

Setting the random effects mean to zero is basically saying that, on average, we do not expect there to be any difference between any particular culture's intercept and slope values and the fixed effects of β_0 and β_1. In other words, β_0 and β_1 are the mean intercept and slope values over *all* cultures. It's easier to see this by, once again, ignoring the distinction between fixed and random effects and thinking about the distribution of the per-culture terms. Since

$$\beta_{0\mathrm{DE}} = \beta_0 + u_{0\mathrm{DE}} \tag{48}$$

and

$$u_{0\mathrm{DE}} \sim \mathcal{N}(0, \sigma_0), \tag{49}$$

it follows straightforwardly that:

$$\beta_{0\mathrm{DE}} \sim \mathcal{N}(\beta_0, \sigma_0). \tag{50}$$

And the same is true for the other cultures, too:

$$\beta_{0\mathrm{UK}} \sim \mathcal{N}(\beta_0, \sigma_0) \tag{51}$$
$$\beta_{0\mathrm{VA}} \sim \mathcal{N}(\beta_0, \sigma_0) \tag{52}$$

In other words, we're just drawing each per-culture intercept from a single, shared distribution, and the model parameters that we estimate and can see in the model summary output are exactly the mean (β_0) and standard deviation (σ_0) of that distribution (and similarly for the per-culture slopes). I contend that this is the most *natural* and enlightening way to think about mixed-effects models. Thinking about them in terms of two different components, fixed effects and random effects, is a bit of a distraction from what the model is actually doing, estimating per-culture terms *and* fitting a single distribution to them.

Having an explicit model of what the variation across different cultures looks like is a great thing because it lets us straightforwardly extend the results of our analysis to *new* cultures not in our sample. Suppose we then want to predict survey responses for participants from, say, the Republic of the Congo. We have no data on what the values of $\beta_{0\mathrm{RC}}$ or $\beta_{1\mathrm{RC}}$ are, but we know that their *expected* values are simply β_0 and β_1, respectively, and we know that it's very unlikely that $\beta_{0\mathrm{RC}}$ is going to be less than $\beta_0 - 2\sigma_0$ or that it's going to be more than $\beta_0 + 2\sigma_0$, because 95% of a normal distribution's probability mass is between 2 standard deviations either side of the mean.

When using the fixed-effects only model, if we want to extend the results of our study to new cultures outside our sample, things are not so straightforward. None of the parameter estimates that we get in the output of our model correspond directly to the overall mean intercept or slope across all cultures. Instead, we get the mean values for one culture (Germany, or whatever is chosen as the reference level), and then the differences between that culture and all the others. And there's no explicit estimate at all of how much variation we expect over cultures. If we want to make predictions for a new culture not in our sample, there is no easy, obvious way forward.

If you've ever been taught a rule like "use fixed effects when your data includes all possible levels of your grouping variable, and use random effects when it doesn't", this is probably what the person teaching you that had in mind. If your data had *all* the world's cultures in it, then the fact that without a mixed-effects model there's no easy way to extend your predictions outside of your sample is not really a disadvantage. That said, I consider that rule bad advice, or at least overly simplistic advice. There are still reasons to prefer mixed-effects modelling even when you *do* have all the levels of a grouping factor in your data - which is not necessarily as rare a situation as it might seem based on our running example. You'll certainly never be able get study participants from every single country, or native speakers of every single language, but it's entirely feasible to e.g. get participants from all 16 states of Germany, or to test the performance of all 8 species of Great Ape on some cognitive task.

You might be thinking that we're making a mountain out of a molehill here. After all, it's not *that* hard to make an overall intercept estimate from the output of the simpler model like this:

$$\hat{\mu} = \frac{\beta_{0\text{DE}} + \beta_{0\text{UK}} + \beta_{0\text{VA}}}{3} \tag{53}$$

$$= \frac{\beta_0 + (\beta_0 + \Delta\beta_{0\text{UK}}) + (\beta_0 + \Delta\beta_{0\text{VA}})}{3} \tag{54}$$

$$= \frac{3\beta_0 + \Delta\beta_{0\text{UK}} + \Delta\beta_{0\text{VA}}}{3} \tag{55}$$

$$= \beta_0 + \frac{\Delta\beta_{0\text{UK}} + \Delta\beta_{0\text{VA}}}{3} \tag{56}$$

And then once you've got that you could estimate the variance in the usual way, by calculating the sample standard deviation from the formula:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}, \tag{57}$$

which, plugging in Equation 56 from above and referring to Table 1, becomes:

$$\hat{\sigma} = \sqrt{\frac{(\beta_0 - \hat{\mu})^2 + (\beta_0 + \Delta\beta_{0\text{UK}} - \hat{\mu})^2 + (\beta_0 + \Delta\beta_{0\text{VA}} - \hat{\mu})^2}{2}} \tag{58}$$

And you're right, while this is tedious and unpleasant to look at, it only needs to be programmed once and then your computer can do it for you quickly every time. This brings us to probably the most important reason to use mixed effects modelling. The estimates it provides of the population means and variances aren't just easier to get at, they will actually tend to be *better estimates* than what you'd get from the above, because of something called *partial pooling*.

Partial pooling

When using fixed effects to model cultural variation, each culture's coefficients are estimated independently of one another, using only the data for that culture. On the face of it, this probably sounds like the most natural thing in the world, and doing anything else might sounds like a very strange proposition. However, there is actually a lot to be said for allowing information to flow between cultures as well, which is exactly what happens in a mixed effects model. One easy way to motivate this is to consider a situation regarding unequal sample sizes.

Suppose we have a lot of data from some of our cultures but not so much from some others. With a fixed effects model, the coefficients for the under-sampled cultures will be less reliably estimated, especially if the few data points we have happen to include some outliers. In a mixed effects model, the good estimates of the

coefficients for the well-sampled cultures can provide the model with a lot of information about the typical extent of variation between cultures (by informing the estimate of the random effects variance terms σ_0 and σ_1). This in turn can help to better inform the estimates of the random effects for the under-sampled cultures: if the few data points we have for one of these cultures are suggestive of random effect values which would make that culture a lot more different from the others than we would expect based on how different the well-sampled cultures are from one another, then the model can correct for this by interpreting that culture's data as an unusual outcome of sampling a typical culture, rather than as a typical outcome of sampling an unusual culture.

This is called "partial pooling" because its somewhere between "complete pooling", where *all* the datapoints across cultures are combined (or "pooled together"), and "no pooling", where none of them are. The very first model we considered in this document, where culture was not involved at all (`glm(y ~ 1 + age, family=binomial)`) is an example of complete pooling; all the data are thrown together regardless of culture to estimate parameters which model all participants identically. Our fixed-effects model (`glm(y ~ 1 + age*culture, family=binomial)`) is an example of no pooling; the data points from different cultures are kept strictly separate.

The way I've described partial pooling above makes the whole thing sound a bit ad-hoc and arbitrary, but the way it actually happens in a real inference is actually perfectly principled. The model doesn't first somehow divide cultures into "well-sampled" and "under-sampled" categories, and then analyse one group first to get an idea of how to best analyse the others; Everything is fit simultaneously. And the strength of the correction which is applied to the under-sampled culture(s) depends in a perfectly principled way on just how many data points you have for each culture, how many well-sampled cultures there are, and just how unusual the outlier points are.

In fact, in a mixed effects model, this "correction for undersampling" can happen even if you have an equal number of samples for every culture. Thinking about this in terms of correcting for sample size variation was just an easy perspective for me to take in a pedagogical context motivate the idea that it's not crazy to estimate coefficients for one culture using data from another. What's actually going on is something a bit more general. The differences in your data across cultures come from two sources: the actual variation across cultures, and variation due to random sampling (unless you take an impractically large number of samples, you'd get different data from two cultures even if the people in them acted identically). By explicitly modelling the cross-cultural variation, mixed effect models can separate the two kinds of variation out to some extent, resulting in better estimates. A fixed effects model is forced to explain more of the variation as being due to differences across cultures, and because of this, fixed effects models will typically *overestimate* the amount of cross-cultural variation. Or, looking at it from the other perspective, mixed effects models are said to exhibit *shrinkage*, because they will typically "shrink" each culture's intercept and/or slope estimates back toward the population mean. Shrinkage is a kind of protection against over-fitting, and it allows more reliable estimates from smaller sample sizes. These statistical advantages are a compelling reason to use mixed effects instead of fixed effects for modelling variation across groups even when you have data for all possible groups in your dataset.

Note that shrinkage is not something which *always* happens with random effects models. If you have a large number of datapoints from each culture, and the genuine differences across cultures are large, then the degree of shrinkage will be very limited, and a fixed effects model and a mixed effects model will yield very similar results. But with random effects in the picture, shrinkage will happen when the situation warrants it, to the degree that the situation warrants it, while with fixed effects only it will never happen, even when it should. Mixed effects models basically offer a kind of statistical safety net which kicks in automatically when needed but which gets out of the way when it isn't. The price of this safety net is increased computational complexity of fitting the model, but computing power is cheap nowadays, so it's a good deal.

The Bayesian big picture

From a Bayesian perspective, there really is nothing special about mixed effects models, and they are treated just like any other model. The constraints on the random effects terms (i.e. that they're normally distributed

with a mean of zero) are nothing but prior distributions over those terms. We also put priors on the standard deviations of the random effects too - typically regularising priors, so that our models don't get excited too easily and over-estimate the extent of variation across cultures (this means that Bayesian mixed effect models can induce more shrinkage than their frequentist counterparts). This means that the full Bayesian form of our model might look something like this:

$$\eta_i = (\beta_0 + u_{0c_i}) + (\beta_1 + u_{1c_i})x_i \tag{59}$$
$$\beta_0 \sim \mathcal{N}(a, b) \tag{60}$$
$$\beta_1 \sim \mathcal{N}(c, e) \tag{61}$$
$$u_{0c_i} \sim \mathcal{N}(0, \sigma_{0c}) \tag{62}$$
$$u_{1c_i} \sim \mathcal{N}(0, \sigma_{1c}) \tag{63}$$
$$\sigma_{0c} \sim \text{Exponential}(1) \tag{64}$$
$$\sigma_{1c} \sim \text{Exponential}(1) \tag{65}$$

I've included non-specific normal priors on the β_0 and β_1 terms, which aren't really important to our present discussion. I don't want to bog us down here talking about priors, but if you're curious, the exponential priors on the two random effect variance terms are an example of regularising priors, which discourage the model from over-estimating the amount of variation across cultures.

You'd fit this model using MCMC, and β_0, β_1, all the individual u_{0c_i} and u_{1c_i} terms as well as σ_{0c} and σ_{1c} would all get estimated simultaneously, in the same way as each other. None of the parts of the model are really different from any other part. Once your chain is finished, you'd get posterior samples for all of these terms, and that means that simply by adding together pairs of values from those samples you could compute posterior distributions for the per-culture values like $\beta_{0\text{DE}}$ which we defined earlier. Everything is simple and consistent and lovely. In the frequentist world, things are... different. But we'll get to that later.

The point I want to emphasise here is that when thinking strictly in a Bayesian sense, the only real difference between the fixed effects and mixed effects models that we've been comparing this whole time is in the priors. We've seen that you can massage the formulation of both models in such a way that they have the exact same overall structure, with separate intercepts and slopes for each culture. All that's different is how we assign priors to those per-culture intercepts and per-culture slopes.

In the mixed effects version, all the per-culture intercepts share a single common prior, and all the per-culture slopes share a separate common prior. Before we see any data, we have exactly the same pre-existing beliefs about all cultures, which seems like a pretty sensible default stance to take. Both of those priors are normals, so they have means and standard deviations, and we put priors on those parameters too (sometimes called "hyperpriors", which is just a fancy name for "priors on parameters of other priors") and estimate them from the data along with everything else. When setting those hyperpriors, the fact that they are priors for parameters which have straightforward interpretations in terms of the population of all cultures makes it comparatively easy to use existing knowledge to choose prior distributions.

In the fixed effects only version, every culture gets its own *separate* prior for its intercept and another one for its slope. In the case of the reference culture, those priors are direct, or explicit - the prior for β_0 really is just a prior for the intercept for German participants, $\beta_{0\text{DE}}$. But for all the other cultures, the priors are indirect, or implicit. You don't get to set a prior for $\beta_{0\text{UK}}$, instead you get to set a prior for $\Delta\beta_{0\text{UK}}$ - the difference between the intercept for the UK and the intercept for Germany. If you give this different a prior with a mean of zero then you can at least tell the model you have the same expected intercept value for both cultures, but it's actually *impossible* to tell the model you have equal (un)certainty about the intercept for the reference culture and any non-reference culture. The model structure forces you to use asymmetric priors, which is kind of nuts[1]. Things are even worse when it comes to the implicit prior on the variance across

[1] Okay, this particular wart is more of a problem with the use of dummy variables for categorical variables than fixed effects *per se*, and with sufficient effort you can work around it, but the issue is solved for free with the mixed effects approach.

cultures, which is not a simple sum of two priors, but something much uglier that I don't even want to try to figure out the exact form of[2].

So basically, we have the same model with two different approaches to setting priors - one approach is simple and direct, reflects the natural structure of the inference problem, gives you the estimation advantages of partial pooling and lets you use regularising priors to further guard against over-fitting the amount of variation across cultures. The other is opaque and implicit and hard to think about, doesn't give you enough freedom to set what are arguably the sensible default priors, and doesn't do partial pooling or regularisation. Unsurprisingly, my advice is that when working in the Bayesian paradigm you should prefer the mixed effects approach and use it as your default approach - literally all you are doing compared to the fixed effects approach is choosing better priors, which is always a good idea.

It's also worth noting that in a Bayesian setting, there's no real reason that the variation across cultures *has* to be a normal distribution. If it reflected your beliefs about the problem at hand, you could replace that distribution with anything else you liked. MCMC doesn't care, and it'll estimate things all the same. This is the reason I decided to talk about "mixed effects" models in this document - "multi-level" or "hierarchical" modelling to me means *any* model at all involving "distributions over distributions", while mixed/random effects means specifically the case of normal distributions centred on zero.

Some frequentist considerations

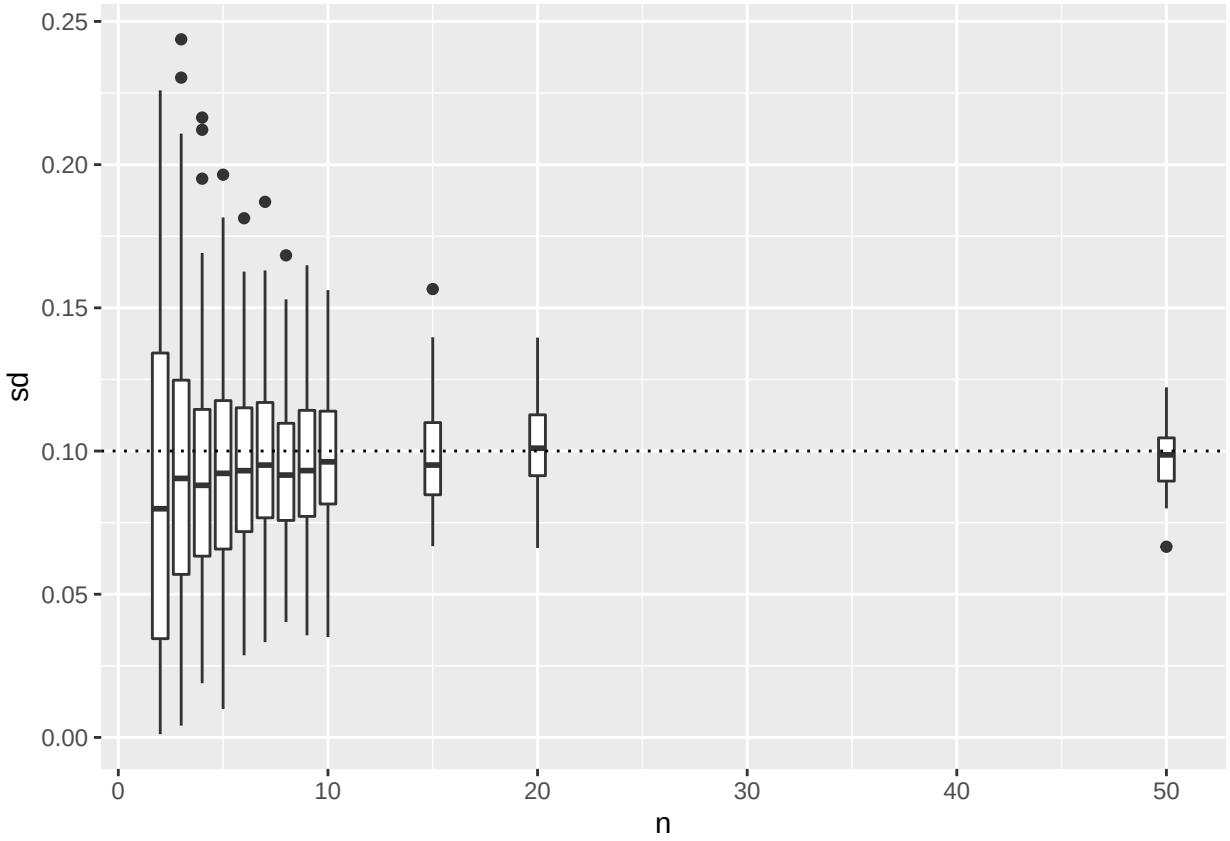
The frequentist implementation of these models is not quite so simple and elegant, even though the "big picture" is identical. There are two things worth knowing in particular.

Minimum number of groups

The first point arises from the fact that in frequentist modelling, there are no prior distributions. In particular there is no prior on the standard deviations of the random effects, σ_0 and σ_1. When estimating these terms, the model has nothing to go on but the data itself, and this has important consequences for when using random effects is feasible. In our ongoing example here, we have three different cultures in our analysis, so the model would be trying to estimate the variance of a distribution like $\mathcal{N}(0, \sigma)$ on the basis of just three samples. That's a *really big* ask. If you only get three samples from a normal distribution, and all the values are pretty close together, you really have no way of knowing whether that's because the standard deviation really is small, or whether it's large and you just got an "unlucky" sample. Similarly if they're spread far apart. This is easy to see by just sampling small numbers of values from a normal distribution and computing the standard deviation of the sample:

```
d <- tibble(n=c(), sd=c())
for(n in c(2,3,4,5,6,7,8,9,10,15,20,50)) {
  for(i in 1:100) {
    d <- add_row(d, n=n, sd=sd(rnorm(n, 0, 0.1)))
  }
}
ggplot(d, aes(group=n, x=n, y=sd)) +
  geom_boxplot() +
  geom_hline(yintercept = 0.10, linetype="dotted")
```

[2] This is still ugly even if you switch from dummy variables to indexed variables, so it really truly is a limitation of fixed effects.



As you can see, with small numbers of samples the estimates can be pretty far from the mark. If you've ever been taught a rule that you should only use random effects if you have more than n different levels of a group (often n is 5 or 6), this is the reason why. The more levels you have the better, and if you only have 2 or 3 and you're committed to a frequentist approach, you might want to consider using the fixed effect approach (and aiming for a lot of samples from each culture). In a Bayesian setting, this problem is not quite so severe. More levels are still better, of course, for estimating variance or anything else. But a strong regularising prior can at least prevent severe overestimates of variability across cultures, making mixed effects less risky with fewer groups.

Per-group estimate quality

The second point to be aware of is regarding the actual values of the random effect terms (the u_{0c_i} and u_{1c_i} values). Sometimes you might not be interested in these. For example, if you are including random intercepts or slopes that vary by participant ID because you have repeated measurements of individual participants and you want to properly account for the non-independence of their responses, it's likely that you don't actually care which particular participants had slightly higher or lower than average values of some parameter. But other times you might care, e.g. if you have random effects of culture like in our example, you might actually want to know how Vanuatan participants compare with German participants.

Bayesian and frequentist mixed-effects methods will both give you access to these values (just pass a fitted `glmer` or `brm` model object to `ranef()`), but the frequentist version will only give you a point estimate, not a full posterior distribution like in the Bayesian version (which could be used to e.g. compute the posterior probability that Vanuatan participants of average age are more likely to give a positive response than German participants of average age). Not only that, but whereas MCMC will sample all the model parameters simultaneously, fully accounting for how uncertainty in one induces uncertainty in others, the frequentist approach estimates parameters in turn; first the variance parameters (the most computationally demanding part of the process), then the fixed effects are estimated based on the point estimates of variance parameters,

then finally the random effects values themselves based on the point estimates of the fixed effects and variance parameters (strictly speaking, the random effects are "predicted", not "estimated", due to weird philosophical distinctions that frequentists make and Bayesians ignore). The point is that the frequentist random effect values lie at the end of a chain of point estimation procedures, and uncertainty is not propagated across the consecutive steps. I wouldn't go so far as to call them "bad estimates"; Each step in the chain is well-justified on theoretical grounds. But if I really cared about these values, and in particular wanted to compare them across societies, I'd feel a lot more comfortable with the Bayesian version, because the idea of throwing out uncertainty at every step freaks me out.

Multiple random effects

Same procedure as every variable

Our example mixed effects analysis so far has included only a single random intercept and a single random slope. Often times you'll want multiple variables to have random effects on a given slope and/or on the intercept. The good news is that essentially nothing changes model-wise when going from one set of random effects to two or more! Each one is handled in exactly in exactly the same way, so you've already learned everything you need to know.

For example, suppose we want to allow our intercept and age slope to vary not just across cultures but also across religion. All we do is change our linear estimator from this:

$$\eta_i = (\beta_0 + u_{0c_i}) + (\beta_1 + u_{1c_i})x_i \tag{66}$$

to this:

$$\eta_i = (\beta_0 + u_{0c_i} + u_{0r_i}) + (\beta_1 + u_{1c_i} + u_{1r_i})x_i \tag{67}$$

where r_i is the religion of the i-th participant, just like c_i is the culture of the i-th participant. As you might expect, these new u_{0r_i} and u_{1r_i} terms are drawn from Normal distributions centred on 0, just like the u_{0c_i} and u_{1c_i} terms:

$$u_{0c_i} \sim \mathcal{N}(0, \sigma_{0c}) \tag{68}$$
$$u_{1c_i} \sim \mathcal{N}(0, \sigma_{1c}) \tag{69}$$
$$u_{0r_i} \sim \mathcal{N}(0, \sigma_{0r}) \tag{70}$$
$$u_{1r_i} \sim \mathcal{N}(0, \sigma_{1r}) \tag{71}$$

They get their own variance terms, so the model can predict more or less variation across religions compared to across cultures as the data allows.

The end result of this addition is that the model assigns separate intercepts and separate age slopes to each unique combination of culture and religion. If we had five religions in addition to our three cultures, there would be $3 \times 5 = 15$ intercepts and 15 slopes. However, the model does not have complete freedom to fit all 15 intercepts and 15 slopes separately, because each intercept or slope is made by summing up the population mean (the fixed effect - a single parameter), the random effect of a particular culture (3 parameters) and the random effect of a particular religion (9 parameters). Thus, $1 + 3 + 5 = 9$ parameters are estimated, and summed up in 15 different ways.

Nested vs crossed structure

You'll often find textbooks or tutorials making a big deal about particular structures in mixed effect models with random effects of multiple variables. The classic example is analysing data about school students, where

each student belongs to a particular classroom, each classroom belongs to a particular school, and each school belongs to a particular "school district" or some other administrative grouping, and there are random effects (either intercepts and/or slopes) for each of these grouping variables. These are known as *nested* random effects[3]. In contrast, sometimes random effects are not nested, e.g. if you have random effects of nationality, language, religion and education level, all possible combinations of values for these variables might appear in your dataset - it's certainly not the case that everybody who speaks a particular language then necessarily practices a certain religion, in the same way that, say, everybody who lives in Leipzig also lives in Germany. Non-nested random effects are known as *crossed* random effects.

The fact that there are names for these structures and that they are often given a lot of attention in educational material is, IMHO, entirely out of proportion to how important they actually are nowadays. From a modelling perspective, it just doesn't matter. Nothing is different at all between nested and crossed structures in terms of the variables that go into the linear predictor and what they mean or how they are distributed.

Why do these things even have names, then? They matter for *implementation*, specifically when doing maximum likelihood fitting of mixed effects models. If your mixed effects are nested, you can take advantage of this structure to fit the model in a way which is faster and more reliable. Back in ye olde days, this was of practical importance. Models with nested random effects were practical to fit on larger datasets than those with crossed random effects without needing expensive computing equipment. Some statistics software didn't even support crossed random effects, because it required more complicated programming. Nowadays, this is not anywhere near as important as it used to be. Computing power is cheap, and the methods in `lme4` can handle crossed random effects without problems. And in a Bayesian context, this has never been important because there's no way to take advantage of nested structure. MCMC, once again, just doesn't care[4] and will work the same way in either case.

Further reading

- Here's a nice and quick visualisation of what adding random intercepts, random slopes, or both to a linear model looks like. However, this document taken alone could easily lead somebody to miss the point of random effects - it says nothing at all about the assumption of a normal distribution with a mean of zero. The *exact* same visualisations of lines against data points could correspond to a strictly fixed-effects model.
- The paper Understanding Mixed-Effects Models Through Data Simulation adocates for the use of mixed-effects models for experimental data instead of more familiar ANOVA and F-test methods. Simulating data is a fantastic way to get a deeper understanding of *any* statistical model.
- The paper Random effects structure for confirmatory hypothesis testing: Keep it maximal argues for including random slopes as well as random intercepts when doing analyses which aim to test a pre-existing hypothesis.

[3]The word "nested" unfortunately has quite a few meanings in statistics, which can be confusing. For example, you might have also heard about "nested models" in the context of model comparison. Whether or not two models are nested in that context has nothing to do with whether or not they have nested random effects.

[4]MCMC is the honey badger of statistical inference.