Mixed effects models

Oct 17th 2019

Richardson Lab Meeting

Roseanna Gamlen-Greene

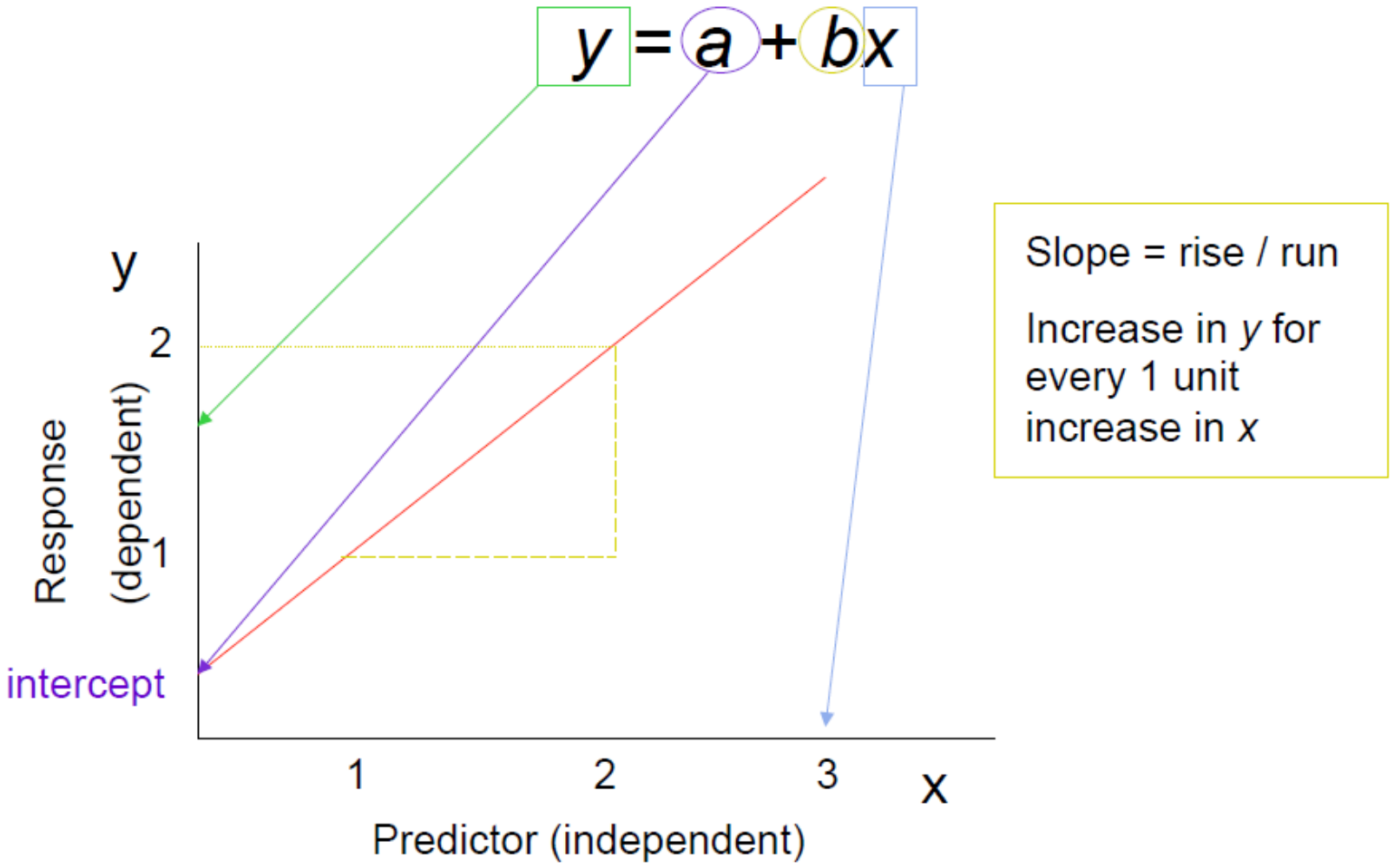
Mixed effects models are an extension of a general linear model

- The general linear model has this basic form:
- $\mathbf{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i}$
- $\mathbf{\varepsilon_i \sim iid\ N(0, \sigma^2)}$ **(The error term)**

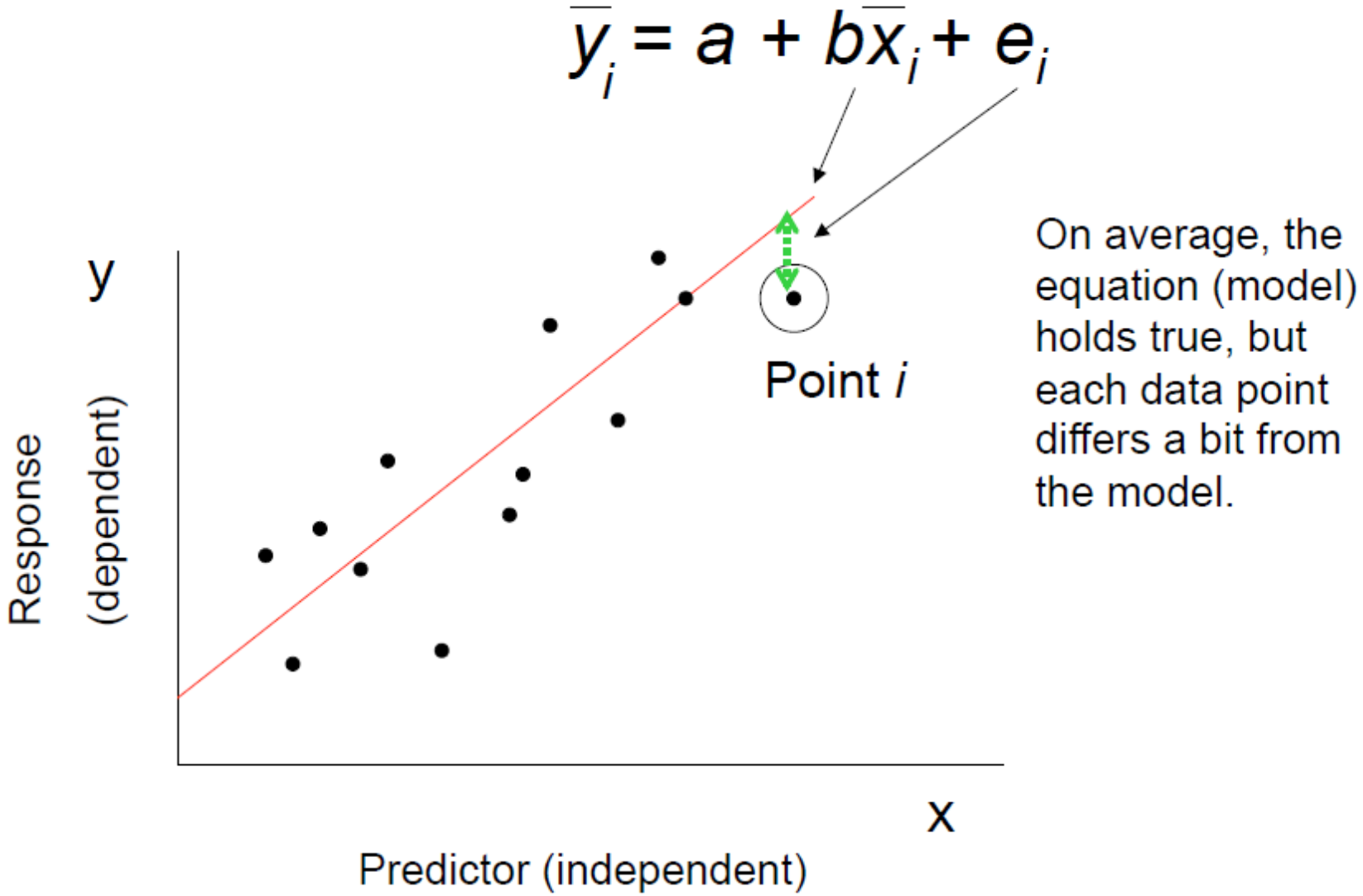
- And has these assumptions (among others)

1. the residuals are independent of each other
2. the residuals are normally distributed
3. the relationship between Y and the model parameters is linear

General linear model (GLM)



General linear model (GLM)



On average, the equation (model) holds true, but each data point differs a bit from the model.

Why use mixed effects models rather than just a GLM?

- When you have a hypothesis that you're testing and have data you've measured but aren't testing the effect of – you just want to control for these extra factors in your model – so that it helps explain the variance around the means, but we don't actually test these factors – they just "explain away" some of the variance
- Uses less degrees of freedom than treating the data as fixed effects = leaves model more power to test the things you really care about

How to know when you have random effects in your study

Generally, you have random effects:

- whenever your sampling design is nested: quadrats within transects; transects within woodlots; woodlots within districts, individuals in families.
- whenever your replicates are grouped spatially or temporally (i.e., in blocks, which are typically analyzed as random effects).
- whenever you divide up natural biological groups (plots, families, clones, ponds, etc.), and apply separate treatments to subgroups (subplots, siblings, individuals from clones, pond-halves, etc.).
- whenever you take measurements on related individuals.
- whenever you measure subjects or other sampling units repeatedly.

Fixed effects

- Predetermined categories of a variable, of **direct interest, repeatable**.
- For example:
  - medical treatments in a clinical trial
  - predetermined doses of a toxin
  - diet or fertilization treatments
  - age groups in a population
  - habitat, season
- Any conclusions reached in the study about differences among groups can be applied only to the groups included in the study. The results cannot be generalized to other treatments, habitats, etc. not included in the study.

Example of factorial experiment with fixed effects

Data: Field transplant experiment to investigate how herbivores affect the abundance of the red alga, *Mazzaella parksii* in the intertidal habitat of coastal Washington State using (Harley 2004). We examine a subset of treatments:

	herbivory treatment	
intertidal zone	**present**	**absent**
low	*n* = 16 plots	*n* = 16 plots
mid	*n* = 16 plots	*n* = 16 plots

N = 64 plots in a completely randomized design

ANOVA table for factorial experiment with fixed effects

The denominator of the *F* statistic for the treatment effect is $MS_{residual}$

Source	SS	df	MS	*F*	*P*
herbivory	1512	1	1512	1512/237.8= 6.36	0.014
zone	89	1	89	89/237.8= 0.37	0.543
herbivory*zone	2617	1	2617	2617/237.8= 11.0	0.002
residual (error)	14271	60	237.8		

64 plots, and as many as 60 degrees of freedom in the error term to test main treatment effect (herbivory). Excellent!

What are random effects? cont

Randomly sampled categories of a variable, representing groups of measurements or units.

For example:

- families made up of siblings
- subjects measured repeatedly (“repeated measures”)
- transects of quadrats in a sampling survey
- field plots of plants
- environment chambers containing aquaria

Groups are assumed to be randomly sampled from a population of groups. Therefore, conclusions reached about groups can be generalized to the population of groups.

What are random effects?

In some cases the random effects are a nuisance, of no interest themselves.

- field plots of plants
- environment chambers containing aquaria

In other cases, measuring the variance associated with different levels of random groupings is the main point of the study.

- families made up of siblings
- subjects measured repeatedly (“repeated measures”)
- transects of quadrats in a sampling survey

In either case, random effects must be incorporated into the model, because units within groups are not independent (e.g, repeated measures). Modeling random effects explicitly avoids pseudoreplication.

Example: a factorial experiment with 1 fixed and 1 random effect

Futuyma and Philippi (1987)

Fall cankerworm, *Alsophila pometaria*

http://cfs.nrcan.gc.ca/subsite/glfc--sugarbush/alsophila--pometaria--images

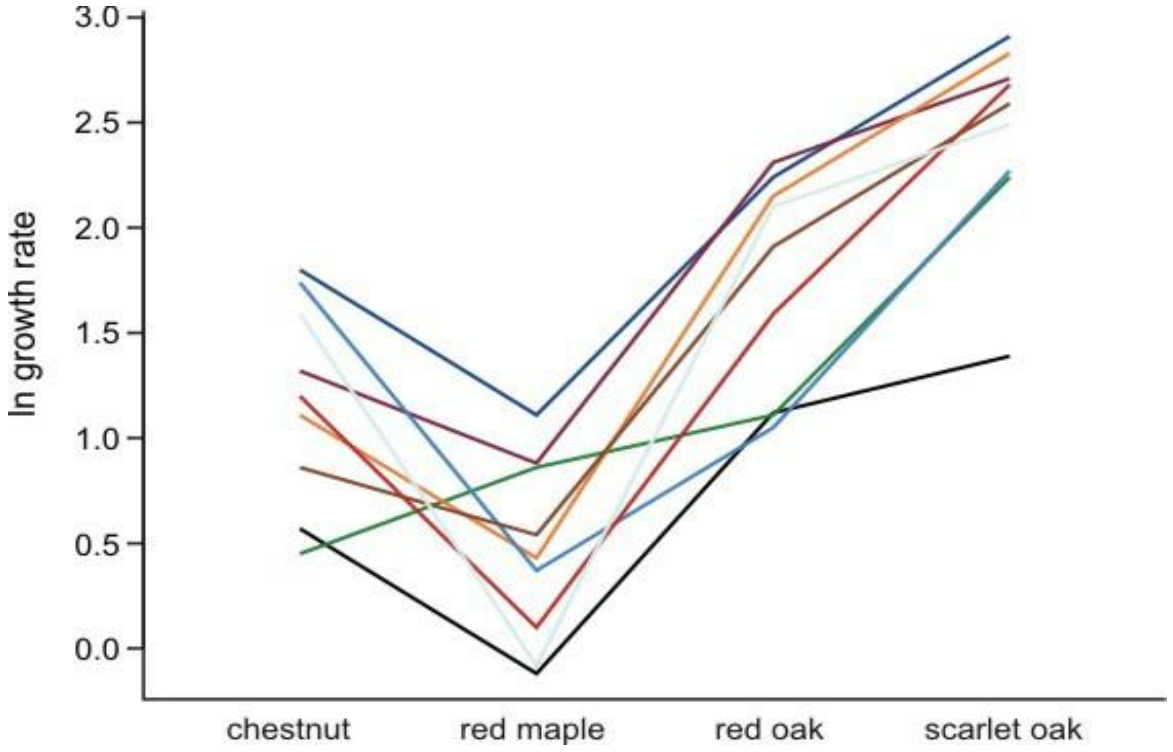
Caterpillars of the fall cankerworm feed on the leaves of hardwood trees. Adult female moths are wingless. Many reproduce clonally, producing only daughters genetically identical to themselves.

Research question: how much genetic variation is present in the moth population for performance on different tree species?

Interaction plot of responses

Design: Sample 9 female moths from a population in NY. Raise larvae from 9 clones on leaves of 4 tree species. Measure growth after 15 days.

Two factors: **Tree species** (Fixed), ***Clone*** (Random)



Mean growth rates of caterpillars from nine families (clones) raised on four tree species.

N = 326 caterpillars total.

Why random effects must be analyzed differently

F--statistics are calculated differently than with fixed effects.

E.g., the denominator of *F* for the test of the fixed effect is not $MS_{residual}$

Source	SS	df	MS	*F*	*P*
tree	167.7	3	55.9	55.9/0.9 = 64.3	<0.001
clone	35.3	8	4.4		
clone*tree	21.0	24	0.9		
residual	122.4	290	0.4		

290 df for error and they can't be used to test treatment effects!

This is because caterpillars from the same clone are not independent.

There are only 9 clones, and it would be pseudoreplication to base test of treatment effect on the number of caterpillars.

What the *F*--statistic for treatment is made of

F is a ratio of Mean Squares. The Mean Squares are chosen so that any variance due to treatment stands out from "extra variance"

$$F = \frac{\text{(variance due to treatment) + (extra variance)}}{\text{(extra variance)}}$$

In a *Fixed Effects* model:

$MS_{treatment}$ = variance due to treatment + variance due to sampling error

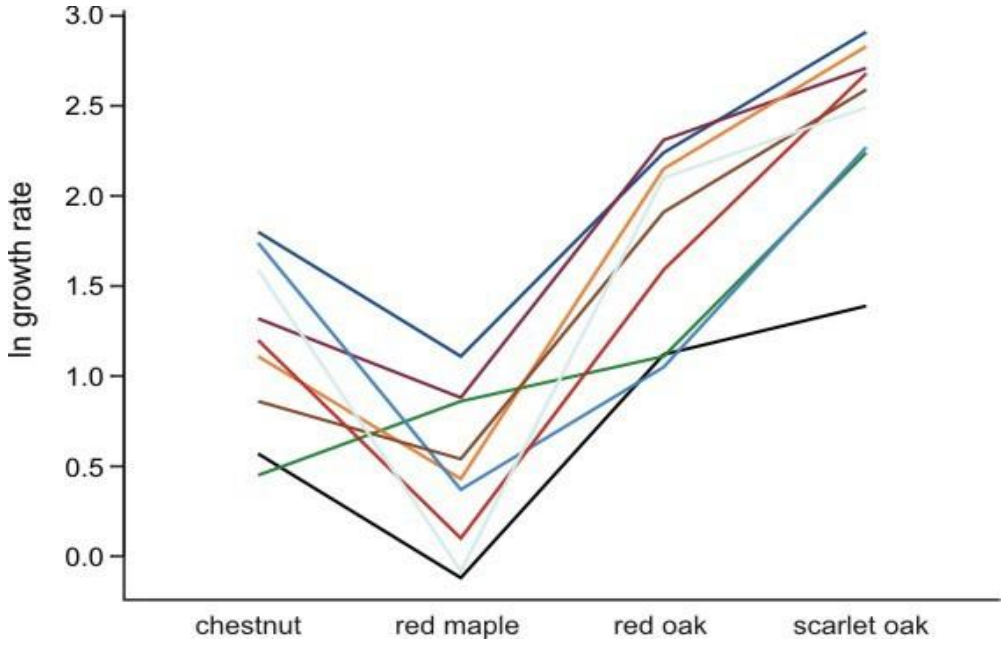
$MS_{residual}$ = variance due to sampling error

In a fixed effects model, the residual error variance (representing sampling error) makes up all the "extra variance". It is the only source of randomness in the design. All else is fixed.

Why the *F*--statistic is calculated differently with random effects

The presence of a random factor adds another source of random variation when estimating treatment effects. It arises from interactions (e.g., differences between clones in how they respond to tree species).

Observed differences among treatment means now also depend on which clones happened to be randomly sampled by the researchers.



F--statistic for treatment when there are random effects

$$F = \frac{\text{(variance due to treatment) + (extra variance)}}{\text{(extra variance)}}$$

When a *Random Effect* is present:

$MS_{treatment}$ = variance due to treatment + variance due to interaction + variance due to sampling error

$MS_{residual}$ = variance due to sampling error

$MS_{interaction}$ = variance due to interaction + variance due to sampling error

So to test treatment effect we use

$F = MS_{treatment}/MS_{interaction}$

Attributes of linear mixed-effects models

- Linear models that include both fixed and random effects.
- In R, the formula specifying the model is split into fixed and random parts.
- There is a different error variance for each source.
- Estimation and testing are based on restricted maximum likelihood, which can handle unequal sample size.
- P-values for fixed effects are conservative when designs are unbalanced.
- Implemented in the nlme package in R.

Assumptions of linear mixed-effects models

- Population variation within groups follows a normal distribution with equal variance among groups.
- Groups are randomly sampled from “population” of groups.
- Group population means follow a normal distribution.
- Measurements within groups are independent.
- No carry-over between repeated measurements on the same subject.
- Sphericity assumption: the population variance of the differences between values of the response variable is the same for all pairs of treatments (see below). (Essentially, this is similar to saying the residuals are independent between treatments)

R code – lme4 package, function lmer()

- (formula for **random terms** | **unit** for which these terms apply).
- Starting on the left side of the bar, the formula for a random intercept, by itself, is simply “1”. The formula for a random regression coeficient for a variable x, *without* the corresponding random intercept, is “0 + x”. Random intercepts are included by default, so “x” and “1 + x” are equivalent specifications of both a random slope and a random intercept.
- x and w = fixed effects
- 1|random effect = random intercept
- 0+x|random effect = random slope and no random intercept

same
- x|random effect = random slope and random intercept
- 1+x|random effect = random slope and random intercept

- x-1|random effect = random slope and random intercept are uncorrelated
- 1|x:random effect = random intercept for random effect that is nested within the fixed effect of x
- x*w|random effect = random slope and intercept for the interaction effect of x*w for every random effect
- 1|random effect/x =random effect is nested within x - interaction between random effect and fixed effect
- 1|random effect a:random effect b = random intercept for interaction

Coding in R: example using lme4

NB: some of the fixed effects here may actually be random effects, but it's just an example so don't worry ☺

Example:

Growth rate of toads = whether there are frogs (frog)*where toads come from (region) + temperature + starting size + block + day weighed juvenile toads + individual population

Now with random effects:

lmer(Growth rate of toads = frogs*region + temperature + starting size + (day|block) + (1|population) + (1|day))

1|population = creates random intercept for toads with the same population (6 populations) and their regression lines are shifted up/down by random amount with mean 0

day|block = (day is fixed, block is random) random slope and intercept for the effect of block on each day of weighing. This means the rate at which toads grow depends to some degree (randomly) on when we weighed them.

1|day = creates random intercept for toads weighed on the same day

More complex

lmer(Growth rate of toads = frogs*region + temperature + starting size + (1+day|block) + (frogs*region|population) + (1|day))

Frogs*region|population = random slope and intercept for the interaction effect of frog*region for every population

1 + day|block = random intercept and slope for block are correlated. Use if observations which share levels of both block and frogs and region are more similar than the sum of the two parts

1|region:population = random effect for population, which is nested within region

Code

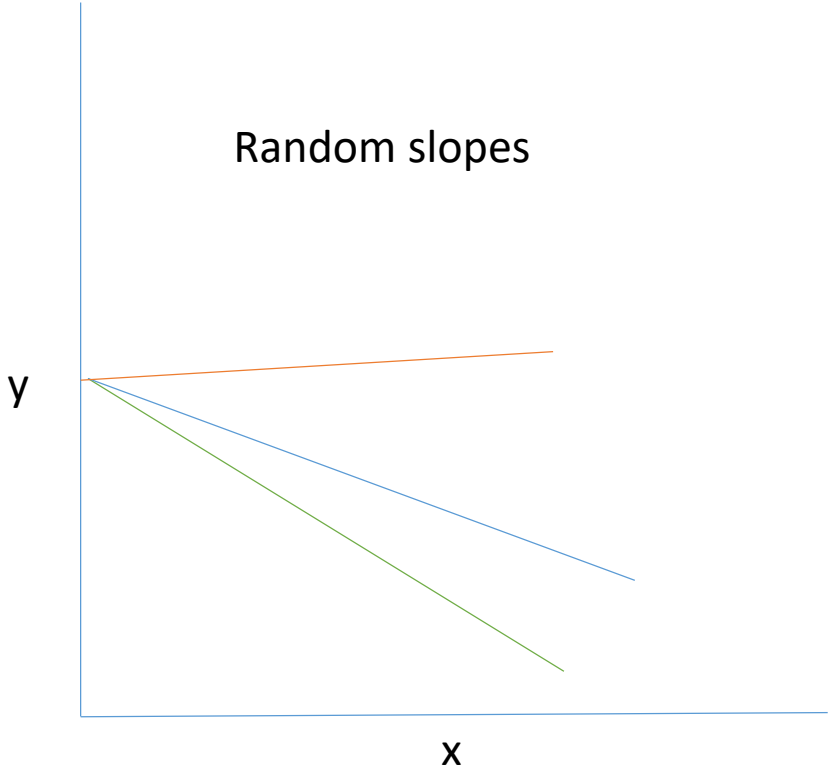
1|random factor = random intercept

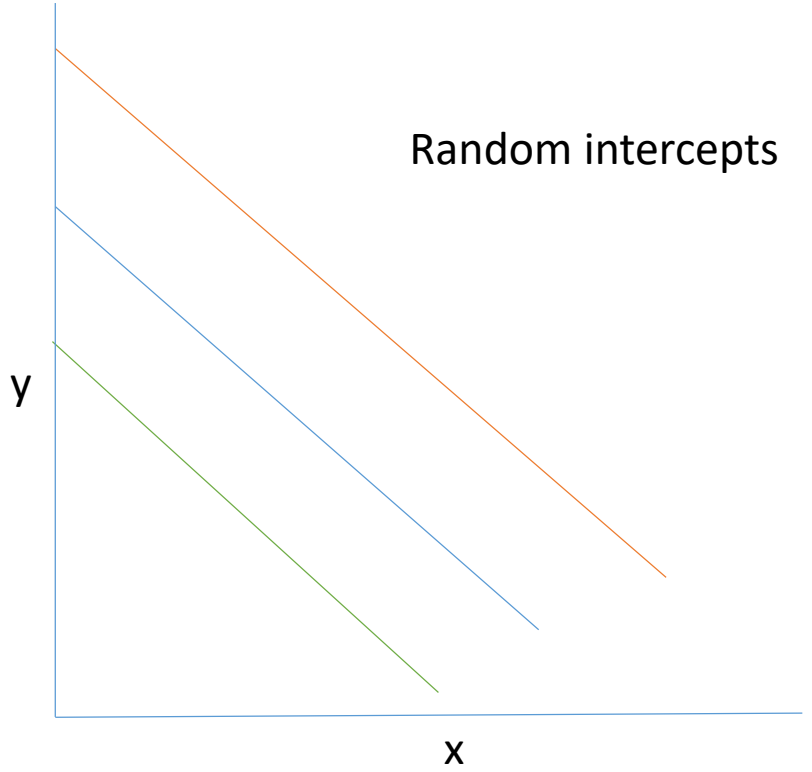
Fixed factor|random factor = random slope – random effect varies the slope of the fixed factor???

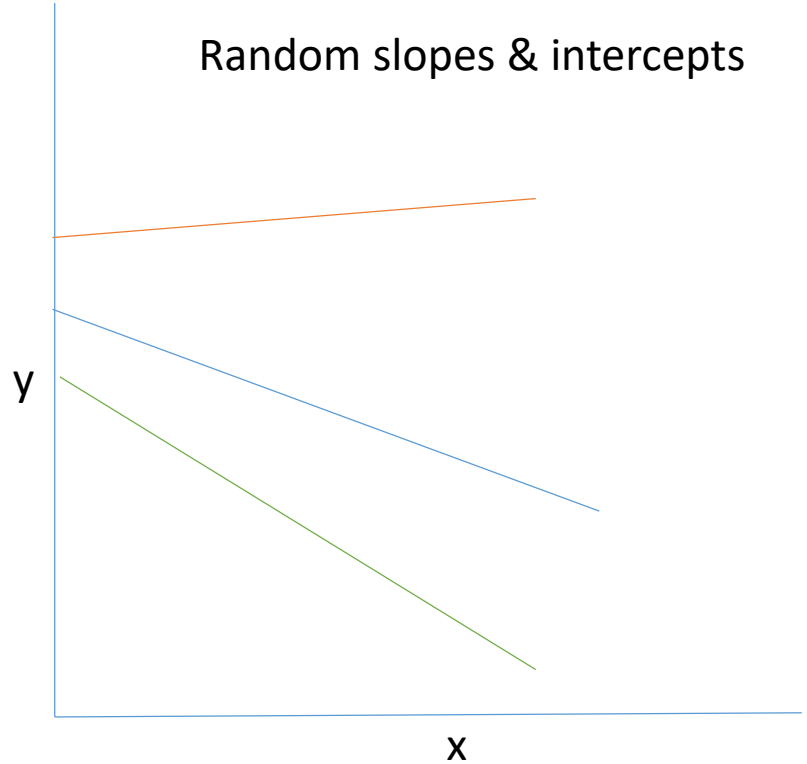
How do I know whether my random effect should have random slopes and/or intercepts in my model? - For continuous x variable

Random intercepts to account for variation where individuals are sampled repeatedly
Random slopes to account for variation in group responses or 'within-individual' variation

Colour = different level of random effect (e.g. three populations)





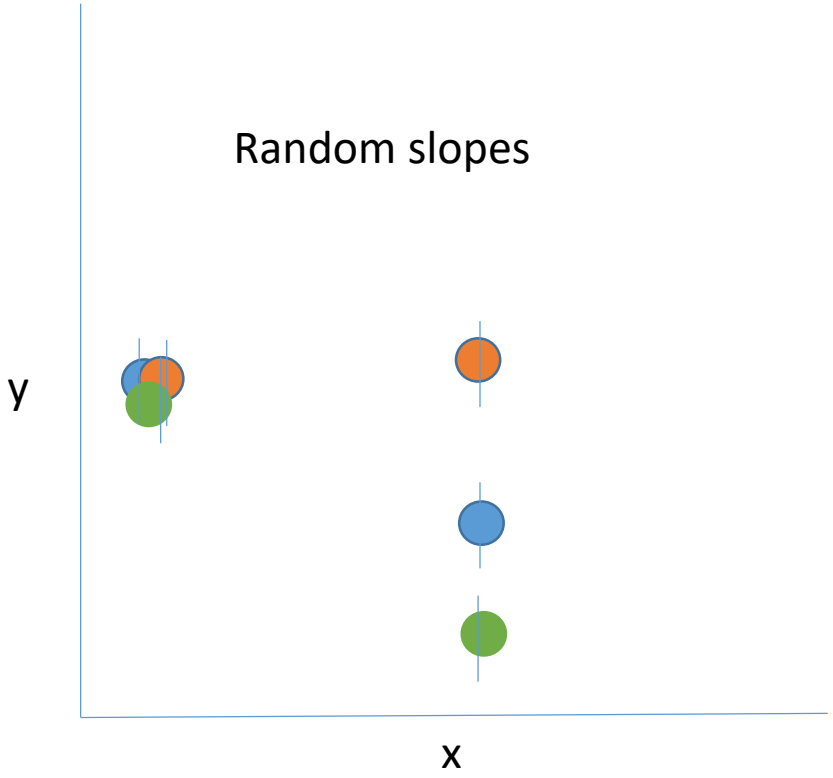


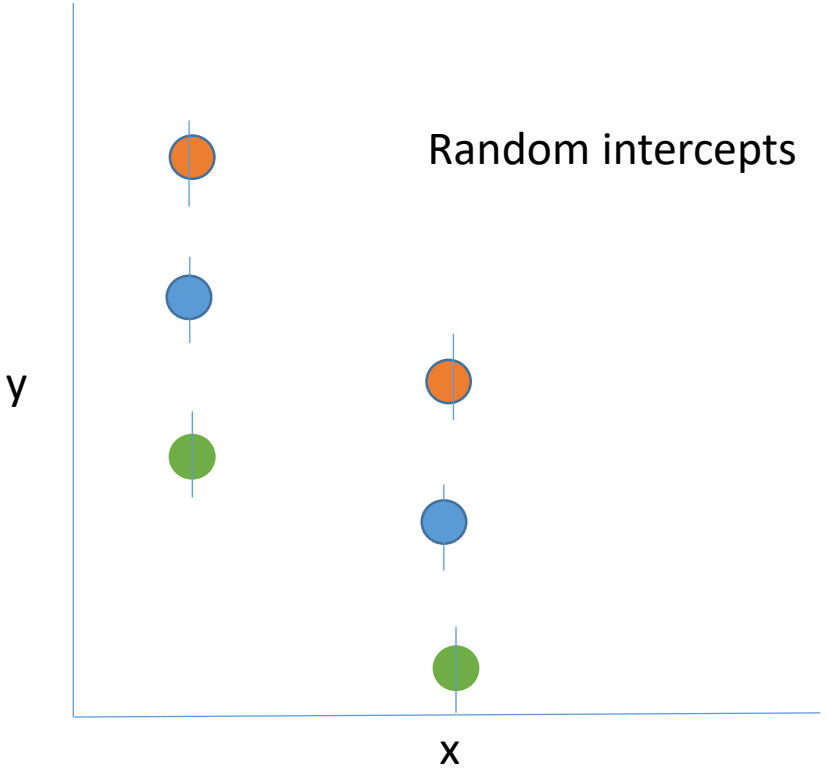
How do I know whether my random effect should have random slopes and/or intercepts in my model? - For categorical x variable

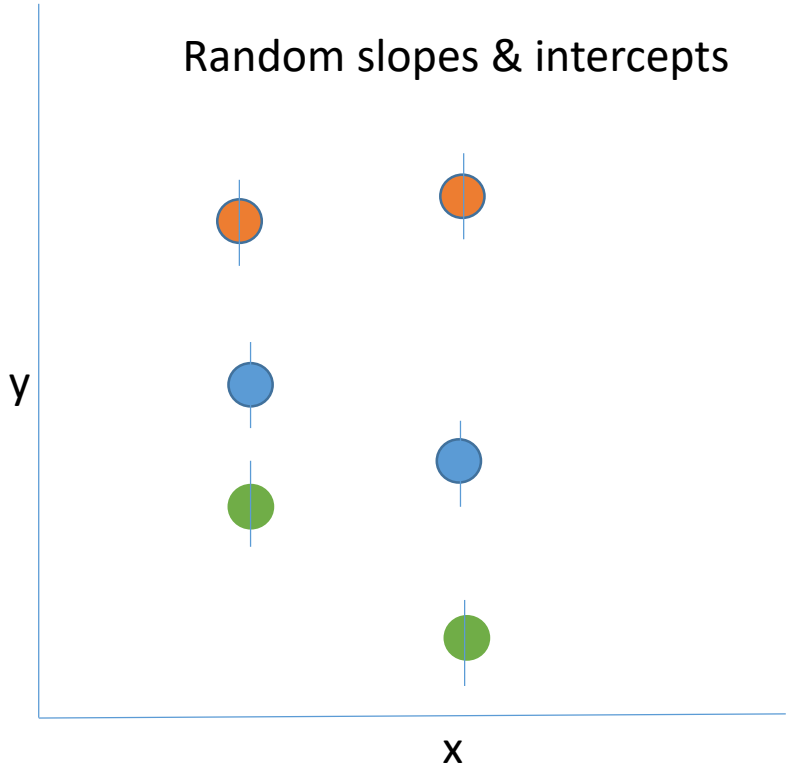
Random intercepts to account for variation where individuals are sampled repeatedly

Random slopes to account for variation in group responses or 'within-individual' variation

Colour = different level of random effect (e.g. three populations)







How do I know whether my random effect should have random slopes and/or intercepts in my model?

- Shielzeth and Forstmeier 2009 suggest you should always do both random slopes and intercepts when you can. Using both random intercepts and slopes reduces the indicidence of type 1 and type 2 errors and reduces the chance of overconfident estimates (unrealistically low standard error, SE). However fitting random slopes requires relatively large sample sizes for model convergence, especially if the data set contains many groups with only a few observations. Therefore use both untless the model doesn't converge, in which case fitting a random intercept only is erereable to not including the random variable at all (Grueber et al 2011).
- Random intercepts to account for variation where individuals are sampled repeatedly
- Random slopes to account for variation in group responses or 'within-individual' variation

Extra reading:

- Zuur 2009 – Mixed Effects Models and Extensions in Ecology with R – free online text book with theory and R code examples
- Grueber et al 2011, Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology.
- Harrison et al 2018, A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ
- Schielzeth and Forstmeier 2009. Conclusions beyond support: overconfident estimates in mixed models. Behavioural Ecology.
- Zuur and leno 2016. A protocol for conducting and presenting results of regression-type analyses. Methods in ecology and evolution.

THE END ☺

Hope that was helpful! If you have any questions or comments, please email: roseanna.gamlen.greene@gmail.com