**Mixed Modeling-multilevel modeling**

**http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#fitting**

**Lecture 27—Wednesday, December 1, 2010**

**Topics**

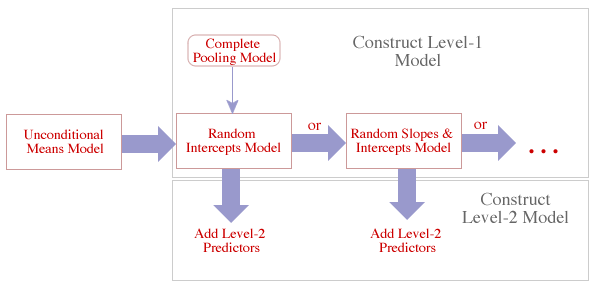
* [Overview of multilevel models](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#overview)
* [The unconditional means model](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#unconmeans)
* [The random intercepts model](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#randomints)
* [The random slopes and intercepts model](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#randomslopeints)
* [Graphs of the basic multilevel models](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#graphs)
* [Fitting the random slopes and intercepts model with the lme4 package](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#fitting)
* [Adding level-2 predictors](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#level2)
  + [Does climate.3 affect the population intercept?](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#intercept)
  + [Does climate.3 affect the population slope?](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#slope)
  + [Does climate.3 affect the population slopes and intercepts?](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#slopeint)
* [Correlated random effects and the centering of predictors](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#correlated)
* [Cited references](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#cited)
* [R code](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#Rcode)

**Overview of multilevel models**

When building a multilevel model there should be a natural progression in the models that one builds. There are four standard preliminary multilevel models that one should always fit and they are the following:

1. the unstructured ordinary least squares (OLS) or complete pooling model,
2. the unconditional means model,
3. the random intercepts model, and
4. the random slopes and intercepts model.

Only once these four models have been fit and the best among them has been selected should the issue of level-2 predictors be considered. Fig. 1 illustrates a general flowchart for fitting these models following the strategy outlined in Singer & Willett (2003). Note: if there is more than one predictor available at level 1, then more complicated models than what are depicted here are possible. These are denoted by … in Fig. 1. For instance, if the basic level-1 model turns out to be quadratic then there could conceivably be random intercepts, random linear coefficients, and random quadratic coefficients.

****

**Fig. 1**  A general flowchart for fitting multilevel models

Arrows in the diagram indicate a logical flow in model fitting. Arrows also point from the simpler to the more complex model, in each case connecting a pair of models that can be compared using a likelihood ratio test.

There are two packages in R for fitting multilevel models. The older and more comprehensive package is nlme, an acronym for **n**on**l**inear **m**ixed **e**ffects models. Its workhorse function is lme whose limitation is that it only fits normal-based models and was not designed to fit mixed models to data that are nonhierarchical. The newer R package for fitting mixed models is lme4 with its workhorse function lmer. It can handle generalized linear mixed effects regression models such as logistic and Poisson regression. It currently lacks some of the nonlinear features of nlme and is not able to fit correlated and heteroscedastic models to the level-1 residuals. Because the lme4 package is under development this situation may eventually change.

Because our primary examples have normally distributed response variables we will focus on the nlme package. As we've seen the primary syntax differences between the lme and lmer functions lie in how the random effects component of the model is specified. The primary reference for the nlme package is Pinheiro and Bates (2000). A book is currently being written describing the lme4 package and a preprint of some of the chapters is available [online](http://lme4.r-forge.r-project.org/book/). I illustrate fitting the four basic models of Fig. 1 using the invertebrates data set we began discussing last time.

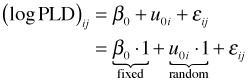
**The unconditional means model**

I begin by fitting the unconditional means (variance components) model. Although this is a model without predictors it does account for the structure in the data. To allow for heterogeneity across level 2 units, it includes a random effect for each level 2 unit (species).

http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/uncond%20mean.gif

where http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/distribution.gif

To understand how the R specification of this model works, I substitute the level-2 equation into the level-1 equation to obtain the composite equation.



This last equation maps directly into R as shown next. The function in the nlme package for fitting linear mixed effects models is lme.

model0 <- lme(fixed=log(PLD)~1, random=~1|species, data=inverts, method="ML")

* The first argument defines the fixed part of the model. It specifies the response and the predictor, in this case only an intercept. Because we will always list the fixed part of the model first, it is unnecessary to include the argument name fixed=, and I will omit it from now on.
* The second argument random= defines the random part of the model. Observe that the response is not specified in the random part so the model statement begins with just a ~. Because the level-2 equations (one in our case) only include random effects for the intercept, the intercept is identified as the only predictor on the right hand side of the model expression in the random= argument. Alternatively in the composite model if we think of u0ias being a coefficient then the predictor it is multiplying is 1. The vertical bar once again separates the model specification from the structural specification. The variable species is listed as defining the level-2 units.
* The third argument data = inverts defines the data set.
* The last argument method="ML" requests that estimates be obtained using full maximum likelihood rather than the default estimation method of REML, restricted (residual) maximum likelihood. This is the appropriate setting for the tests we will carry out today.

The summary output from the model is displayed below.

summary(model0)

Linear mixed-effects model fit by **maximum likelihood**  
Data: inverts   
     **AIC**      BIC    **logLik**  
451.2372 461.3906 -222.6186

**Random effects:**  
Formula: ~1 | species  
        (Intercept)  Residual  
StdDev:   0.8198609 0.4529645

**Fixed effects:** log(PLD) ~ 1   
               Value Std.Error  DF  t-value p-value  
(Intercept) 2.888758 0.1008001 144 28.65829       0

Standardized Within-Group Residuals:  
        Min          Q1         Med         Q3        Max   
-2.15822703 -0.58084922 -0.05855462 0.55220371 2.38062759

Number of Observations: 218  
Number of Groups: 74

I've highlighted in bold the quantities of interest to us.

* In the first line we're told that maximum likelihood is the estimation method that was used. Thus the AIC and log-likelihood that appear on lines 3 and 4 are correct and can be used to make comparisons between models with different fixed effects (or random effects).
* The next section lists the estimates from the random effects part of the model. Recall that we don't actually estimate the random effects, we instead estimate the parameters of their distribution. For the unconditional means model these are tau2and sigma2. Notice that in the line where the numerical estimates appears the label is StdDev, so in fact what's displayed here are τ (Intercept) and σ (Residual). Because it is the squares of these quantities that are interpretable, I'll wait to comment further on these numbers.
* In the Fixed effects section is the reported value of the intercept, its estimated standard error, and a Wald test that tests whether its value is significantly different from zero.

Because the primary reason for fitting the unconditional means model is to obtain the variance components, I turn to these next. Variance components are extracted from the model with the VarCorr function.

VarCorr(model0)

species = pdLogChol(1)   
            Variance  StdDev   
(Intercept) 0.6721719 0.8198609  
Residual    0.2051768 0.4529645

What's reported in the first column are the squares of the values reported in the summary table, so these are variances. From the output we determine that http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tauhat2.gif= 0.672 and http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/sigmahat2.gif= 0.205. From this we learn a couple of things.

1. Because http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tauhat2.gif> 0, heterogeneity is present. Put another way we have evidence of a correlation structure (as opposed to independent observations).
2. Furthermore because the variability between species is nearly three times the variability within species, level-2 modeling in which predictors are included to explain differences between species is likely to be productive.

We can use the output from VarCorr to calculate the intraclass correlation coefficient. (See [lecture 26](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture26.htm#correlation).)

correlation

Unfortunately VarCorr has not produced a matrix, but a table, and the elements of the table are actually character values.

VarCorr(model0)[1,1]

[1] "0.6721719"

As a result our calculations will turn out to be rather awkward. I need to convert each character value to a numeric value before I can do arithmetic on it.

tau.sq <- as.numeric(VarCorr(model0)[1,1])

sigma.sq <- as.numeric(VarCorr(model0)[2,1])

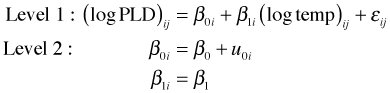
tau.sq/(tau.sq+sigma.sq)

[1] 0.76614

So the correlation between observations coming from the same species is 0.766, a sizeable value and far from 0. The implication is that the independence assumption of ordinary linear regression is not valid with these data.

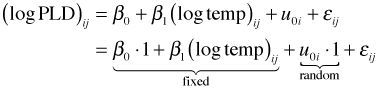
**The random intercepts model**

The random intercepts model enhances the unconditional means model by including a level-1 predictor. We add the variable log(temp) to the model.



For completeness I've included equations at level 2 for all level-1 parameters. In the random intercepts model the slopes are assumed to not vary across species, so I set the level-2 equation for the slope for each species equal to a constant.

The composite equation equivalent of the multilevel random intercepts model is shown next. It's obtained by plugging the level-2 equations into the level-1 equation.



From the composite model we can directly construct the R code for fitting this model.

model1 <- lme(log(PLD)~log(temp), random=~1|species, data=inverts, method="ML")

This is identical to the unconditional means model specification except that we now have a predictor in the fixed effects portion of the model. Recall the R convention that if a predictor is included in regression model, the intercept is included automatically. Thus an equivalent way of writing the fixed part of the random intercepts model would be

fixed=log(PLD)~1+log(temp)

The model summary is shown below. The only new information we obtain is the estimated coefficient of the level-1 predictor log(temp), shown here to be negative and significantly different from zero.

summary(model1)

Linear mixed-effects model fit by maximum likelihood  
Data: inverts   
     AIC      BIC    logLik  
239.5597 253.0977 -115.7798

Random effects:  
Formula: ~1 | species  
       (Intercept)  Residual  
StdDev:  0.8891414 0.2115642

**Fixed effects**: log(PLD) ~ log(temp)   
               Value Std.Error   DF   t-value p-value  
(Intercept) 6.164731 0.17954444 143  34.33541       0  
log(temp)  -1.161889 0.05168823 143 -22.47879       0  
Correlation:   
          (Intr)  
log(temp) -0.811

Standardized Within-Group Residuals:  
        Min          Q1         Med         Q3        Max   
-3.08734705 -0.45362977 -0.04267881 0.50275266 3.62335848

Number of Observations: 218  
Number of Groups: 74

The first use we can make of the random intercepts model is in assessing how important it was to account for the structure of the data in our regression model. To do so we need to first fit a model in which the structure is ignored. That would be an ordinary linear regression model with log(temp) as the predictor. In the context of multilevel models this model is sometimes called the complete pooling model.

model.OLS <- lm(log(PLD)~log(temp), data=inverts)

We can test for the importance of structure using a likelihood ratio test. The degrees of freedom for the test is just the difference in the number of estimated parameters. The random intercepts model estimates one additional parameter http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau2.gif.

logLik(model.OLS)

'log Lik.' -273.5342 (df=3)

logLik(model1)

'log Lik.' -115.7798 (df=4)

LR1 <- 2\*(logLik(model1)-logLik(model.OLS))[1]

1-pchisq(LR1,1)

[1] 0

Clearly the difference in the log-likelihoods is gigantic. The AIC tells an identical story. The random intercepts model is a huge improvement over the ordinary least squares model. Clearly it is important that we account for data structure in fitting our models.

Technically the likelihood ratio test we just carried out is incorrect. The null hypothesis we are testing asserts that the variance of the random effects is zero. Because variances are non-negative, a value of zero is a boundary value. One of the conditions for the chi-squared approximation to hold (called a regularity condition) is that the value being tested does not lie on the boundary of the parameter space. This condition fails here and hence the chi-squared distribution of the likelihood ratio statistic is suspect.

A number of solutions have been proposed. One that is easy to implement is the following (Verbeke and Molenberghs 2000, p. 64–73) and I describe it for the general case. Suppose the models in question differ by one random effect. Thus the goal is to compare two nested models such that one model includes *q* random effects and the second model includes one additional random effect for a total of *q* + 1 random effects. The likelihood ratio statistic in this case has a distribution that is approximately a 50:50 mixture of two chi-squared distributions, one with *q* degrees of freedom and one with *q* + 1.

Thus to compare the two models above we should use a mixture of a http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/chisqzero.gifand a http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/chisqone.gifrandom variable with equal weights 0.5. A http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/chisqzero.gifrandom variable gives a probability mass of 1 to the value 0 and a mass of 0 to every other value. If LR is the observed value of the likelihood ratio statistic, then the *p*-value is calculated as follows.

http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/pval.gif

Thus in our example above the *p*-value of our test is calculated as follows.

.5\*0 + .5\*(1-pchisq(LR1,1))

[1] 0

so in this instance our conclusion does not change.

The second use we can make of the random intercepts model is to assess the importance of the level-1 predictor. From the Wald test that appears in the summary table we already know that the level-1 predictor is statistically significant. We can obtain the likelihood ratio test version of this test by comparing the random intercepts model to the unconditional means model, using R's anova function.

anova(model1,model0)

       Model df      AIC      BIC    logLik   Test  L.Ratio p-value  
model1     1  4 239.5597 253.0977 -115.7798   
model0     2  3 451.2372 461.3906 -222.6186 1 vs 2 213.6775  <.0001

We can also use these two models to obtain a pseudo-R2 statistic for the importance of the level-1 predictor. The formula is shown below.

http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/level1%20r2.gif

From the form of the pseudo-R2 statistic we see it can be interpreted as the fraction of the original level-1 variance that has been explained by the addition of the level-1 predictor . It's called a pseudo-R2 statistic because the different variance terms that appear in the multilevel models are not entirely independent of each other. A predictor added at one level can also have ramifications at other levels making interpretation of these statistics difficult. To calculate the pseudo-R2 statistic for the invertebrates random intercepts model we need the variance components.

VarCorr(model1)

species = pdLogChol(1)   
            Variance  StdDev   
(Intercept) 0.7905725 0.8891414  
Residual    0.0447594 0.2115642

VarCorr(model0)

species = pdLogChol(1)   
            Variance  StdDev   
(Intercept) 0.6721719 0.8198609  
Residual    0.2051768 0.4529645

Pulling out the relevant level-1 variances (labeled Residual in the output) we have the following.

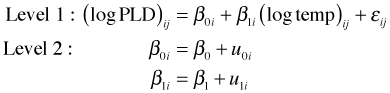
(as.numeric(VarCorr(model0)[2,1]) - as.numeric(VarCorr(model1)[2,1]))/ as.numeric(VarCorr(model0)[2,1])

[1] 0.7818496

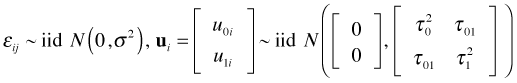
So 78% of the variability of log(PLD) at level 1 (the individual level) is explained by its linear relationship to log(temp). This calculation also highlights why this statistic is called a pseudo-R2 statistic. Observe from the variance component table that while http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/sigma2.gifdecreased (from 0.205 to 0.045) as a result of adding log(temp) to the model at level 1, http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau2.gifwent up (from 0.67 to 0.79). Thus if we were to calculate an R2 at level 2 here, it would be negative! Because of the inter-relationship between the variances at different levels it is difficult to take the pseudo-R2 calculation too seriously here.

**The random slopes and intercepts model**

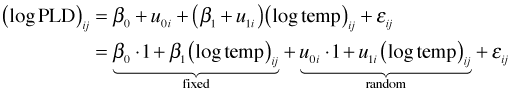
The random slopes and intercepts model adds a second equation at level 2 to the random intercepts model by allowing slopes to vary across species. The multilevel model formulation is shown below.



where



The equivalent composite equation, obtained by plugging the two level-2 equations into the level-1 equation, is shown next.



From this we can write down the R formulation of the model.

model2 <- lme(log(PLD)~1+log(temp), random=~1+log(temp)|species, data=inverts, method="ML")

or equivalently

model2 <- lme(log(PLD)~log(temp), random=~log(temp)|species, data=inverts, method="ML")

The summary output is slightly more complicated because there are now three parameters estimated in the random part of the model.

summary(model2)

Linear mixed-effects model fit by maximum likelihood  
 Data: inverts   
       AIC      BIC    logLik  
  210.4372 230.7441 -99.21859

Random effects:  
 Formula: ~log(temp) | species  
 Structure: General positive-definite, Log-Cholesky parametrization  
            StdDev    Corr    
(Intercept) 2.0544351 (Intr)  
log(temp)   0.5301851 -0.92   
Residual    0.1520061

Fixed effects: log(PLD) ~ log(temp)   
                Value  Std.Error  DF   t-value p-value  
(Intercept)  7.088667 0.29554222 143  23.98529       0  
log(temp)   -1.454417 0.08465557 143 -17.18040       0  
 Correlation:   
          (Intr)  
log(temp) -0.942

Standardized Within-Group Residuals:  
        Min          Q1         Med          Q3         Max   
-2.56320123 -0.36679646 -0.01297876  0.41001700  3.64570155

Number of Observations: 218  
Number of Groups: 74

To assess whether including random slopes has improved the model we should compare this model with the random intercepts model.

anova(model2,model1)

       Model df      AIC      BIC     logLik   Test  L.Ratio p-value  
model2     1  6 210.4372 230.7441  -99.21859   
model1     2  4 239.5597 253.0977 -115.77985 1 vs 2 33.12252  <.0001

So the addition of random slopes is a significant improvement.

Again this test is technically not correct. Instead to compare these two models we should use a mixture of http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/chisqone.gifand http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/chisqtwo.gifrandom variables with equal weights 0.5.

pvalue

Thus in our example above a more honest *p*-value for our test is calculated as follows.

LR.out <- anova(model2,model1)

.5\*(1-pchisq(LR.out$L.Ratio[2],1)) + .5\*(1-pchisq(LR.out$L.Ratio[2],2))

[1] 3.642662e-08

which is still highly significant. I examine the variance components next.

VarCorr(model2)

species = pdLogChol(log(temp))   
            Variance   StdDev    Corr   
(Intercept) 4.22070229 2.0544348 (Intr)  
log(temp)   0.28109623 0.5301851 -0.92   
Residual    0.02310587 0.1520061

From the output we see that http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tauhat0sq.gif= 4.22, http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tauhat1sq.gif= 0.281, and http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/sigmahat2.gif= 0.023. Notice that http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tauhat01.gifis not reported. Instead an estimate of the correlation is shown. We can calculate the covariance from the formula http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau01%20formula.gifas follows .

as.numeric(VarCorr(model2)[2,3]) \* as.numeric(VarCorr(model2)[1,2]) \* as.numeric(VarCorr(model2)[2,2])

[1] -1.002092

The level-2 variance components are not directly comparable to each other because http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tauhat1sq.gifneeds to be scaled by the predictor log(temp). Any change in the scale of temp will affect the reported value of http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tauhat1sq.gif. Thus although it's tempting at this point to say that there is greater variance in the intercept random effects than in the slope random effects (and thus directly address one of the research questions of the investigators) that conclusion does not follow. In truth the variance relationship changes with the value of log(temp). Furthermore there is a nonzero covariance between the random effects which means the two variances are not independent of each other.

The large reported negative correlation between the random effects, –0.92, can potentially lead to estimation problems down the road (as we'll eventually see). A standard way to fix this problem is to center the predictor. Centering the predictor generally means subtracting off its mean. I refit the model after centering log(temp) and check the correlation. I use the I( ) function so that the centering arithmetic is carried out before the model is fit. Notice that the centering must be done everywhere the variable log(temp) appears, including where it appears in the random argument.

model2.5 <- lme(log(PLD)~I(log(temp)-mean(log(temp))), random=~I(log(temp)-mean(log(temp)))|species, data=inverts, method="ML")

VarCorr(model2.5)

species = pdLogChol(I(log(temp) - mean(log(temp))))   
                               Variance   StdDev    Corr    
(Intercept)                    0.81092486 0.9005137 (Intr)  
I(log(temp) - mean(log(temp))) 0.28109620 0.5301851 -0.444  
Residual                       0.02310586 0.1520061

Observe that the correlation has been reduced by half. The random slopes and intercepts model is typically the baseline upon which the level-2 model is built. We'll start that process next.

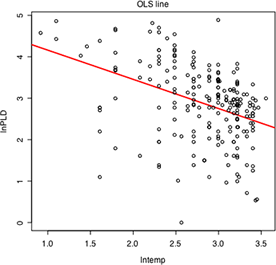
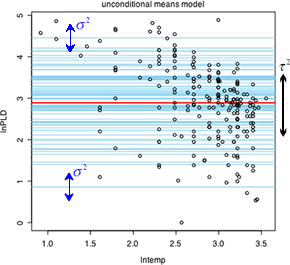
**Graphs of the basic multilevel models**

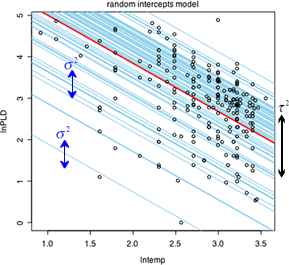
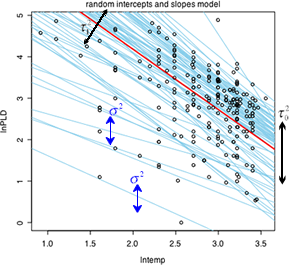
As we've seen for these data any of the models that account for the hierarchical structure of the data set are a huge improvement over an unstructured model. It's worth noting that in the random slopes and intercepts model we've accounted for the structure by estimating only three additional parameters over an ordinary linear regression model. These three parameters, the two random effect variances and the correlation between the random effects, define the distribution of the random effects.

Graphs of the four models we've fit thus far are shown in Fig. 2.

* http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau2.gifand http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau02.gifrefer to the variability in the intercepts of the subject-specific regression lines about the intercept of the population-average line (the line determined by the estimated fixed effects). This describes the spread of the lines in the vertical direction (keeping their slopes the same).
* http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau12.gifis the variability of the individual slopes about the slope of the population-average line. It can be thought of as describing the extent to which individual lines are rotated around the population-average line.
* http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/sigma2.gifis the average variability of the observations about their own individual species-level regression lines. It has the same interpretation as in ordinary regression.

The remaining parameter http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau01.gif, or equivalently ρ, is not depicted in the figure.

**Fig. 2**  The OLS (complete pooling) model and the three basic multilevel models for the larval development data set.

**Fitting the random slopes and intercepts model with the lme4 package**

Using the lme4 package to fit multilevel models follows the same protocol that we used for random intercepts models. There is no random argument. Instead the random terms appear as regressors using the notation (x|group). Just as with lme, if the coefficient is specified as random it is assumed that the intercept should also be random. Thus the notation (x|group) implies (1+x|group). The random slopes and intercepts model is fit as follows.

detach(package:nlme)

library(lme4)

model2.lmer <- lmer(log(PLD)~log(temp)+ (log(temp)|species), data=inverts, REML=FALSE)

The default probability family for lmer is normal (gaussian). As was discussed previously, the nlme package and the lme4 package share functions with the same name so it is necessary to detach one before using the other. It's worth noting that the estimates, log-likelihood, and other statistics returned by lmer don't match what we obtained previously with lme.

summary(model2.lmer)

Linear mixed model fit by maximum likelihood   
Formula: log(PLD) ~ log(temp) + (log(temp) | species)   
   Data: inverts   
   AIC   BIC logLik deviance REMLdev  
 228.4 248.7 -108.2    216.4   223.1  
Random effects:  
 Groups   Name        Variance Std.Dev. Corr     
 species  (Intercept) 2.187731 1.47910           
          log(temp)   0.046322 0.21522  -1.000   
 Residual             0.041248 0.20309           
Number of obs: 218, groups: species, 74

Fixed effects:  
            Estimate Std. Error t value  
(Intercept)  6.50414    0.23158   28.09  
log(temp)   -1.26940    0.05938  -21.38

Correlation of Fixed Effects:  
          (Intr)  
log(temp) -0.918

The source of the problem is the high correlation between the random slopes and random intercepts, reported here to be –1.000. This has prevented lmer from finding the correct solution, although lme was able to. This is not an indictment of lmer but is really an indication that the uncentered model is currently ill-conditioned and poorly formulated. We'll address this issue further [below](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#correlated).

detach(package:lme4)

library(nlme)

**Adding level-2 predictors**

The climate.3 variable in the data set is a level-2 predictor. It varies only at the species level and is a categorical variable with three levels.

table(inverts$climate.3)

polar temperate tropical   
   29       152       37

This last display is not overly informative because it includes in its tabulation the multiple observations from the same species. To get a count of how many species there are from each geographic region we can do the following.

table(tapply(inverts$climate.3, inverts$species, function(x) x[1]))

1  2  3   
8 53 13

* The generic function, function(x) x[1], takes the first element from the vector x.
* In the tapply function call we group the climate observations by species (second argument) but the generic function then only takes the first one. The result is that we get one observation per species and hence a correct count of the totals when the outside table function is applied to the result.

Observe that the data are very unbalanced. Most of species in the data set are classified as temperate. The climate.3 variable is a factor variable in R. By default dummy variables have been constructed alphabetically that contrast temperate and tropical species with polar species.

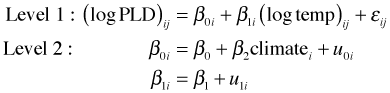
contrasts(inverts$climate.3)

          temperate tropical  
polar             0        0  
temperate         1        0  
tropical          0        1

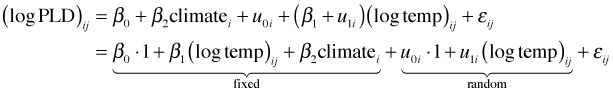
To include climate.3 in the model we have to decide in what way we think it alters the basic level-1 relationship between log(PLD) and log(temp). Does it modify the intercept, the slope, or both? In general theory should guide us in answering this question. Because we as outsiders don't have theory to guide us, I'll try all three possibilities.

**Does climate.3 affect the population intercept?**

If climate.3 affects the intercept it should be entered into the level-2 equation for the intercept.



The corresponding composite model is shown below. In it we see that climate is part of the fixed effect portion of the model.



To estimate this model in R we start with the random slopes and intercepts formulation and just add climate to the fixed part of the model.

model3 <- lme(log(PLD)~log(temp) + climate.3, random=~log(temp)|species, data=inverts, method='ML')

summary(model3)

Linear mixed-effects model fit by maximum likelihood  
 Data: inverts   
       AIC      BIC    logLik  
  213.8831 240.9590 -98.94154

Random effects:  
 Formula: ~log(temp) | species  
 Structure: General positive-definite, Log-Cholesky parametrization  
            StdDev    Corr    
(Intercept) 2.0229353 (Intr)  
log(temp)   0.5313966 -0.916  
Residual    0.1519112

Fixed effects: log(PLD) ~ log(temp) + climate.3   
                       Value Std.Error  DF    t-value p-value  
(Intercept)         6.889221 0.4181549 143  16.475284  0.0000  
log(temp)          -1.452980 0.0852618 143 -17.041387  0.0000  
climate.3temperate  0.194004 0.3313066  71   0.585573  0.5600  
climate.3tropical   0.308108 0.3860982  71   0.798005  0.4275  
 Correlation:   
                   (Intr) lg(tm) clmt.3tm  
log(temp)          -0.672                  
climate.3temperate -0.704  0.019           
climate.3tropical  -0.591 -0.004  0.749

Standardized Within-Group Residuals:  
        Min          Q1         Med          Q3         Max   
-2.50718015 -0.36428111 -0.01874428  0.40493319  3.66881086

Number of Observations: 218  
Number of Groups: 74

Because climate.3 is a categorical variable with more than two categories, the Wald tests shown in the summary output do not provide a test of the construct climate.3. For that we need to carry out a likelihood ratio test.

anova(model2,model3)

       Model df      AIC      BIC    logLik   Test  L.Ratio p-value  
model2     1  6 210.4372 230.7441 -99.21859                          
model3     2  8 213.8831 240.9590 -98.94154 1 vs 2 0.554096   0.758

So climate.3 is not a significant predictor of the intercepts. We can further quantify the extent to which climate has explained the between-species variability in their intercepts by computing a pseudo-R2 statistic.

VarCorr(model2)

species = pdLogChol(log(temp))   
              Variance    StdDev Corr   
  
log(temp)   0.28109623 0.5301851 -0.92   
Residual    0.02310587 0.1520061

VarCorr(model3)

species = pdLogChol(log(temp))   
              Variance    StdDev Corr   
(Intercept) 4.09225978 2.0229335 (Intr)  
log(temp)   0.28238248 0.5313967 -0.916  
Residual    0.02307705 0.1519113

Because climate.3 was entered in the level-2 intercept equation we should look to the intercept variance component, http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/tau02.gif, for whatever effect it is having. From the output we can see the effect has been minimal. Formally we compute the statistic as follows.

http://www.unc.edu/courses/2010fall/ecol/563/001/images/lectures/lecture27/level2%20r2.gif

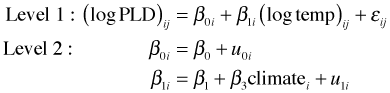
(as.numeric(VarCorr(model2)[1,1]) - as.numeric(VarCorr(model3)[1,1])) / as.numeric(VarCorr(model2)[1,1])

[1] 0.03043155

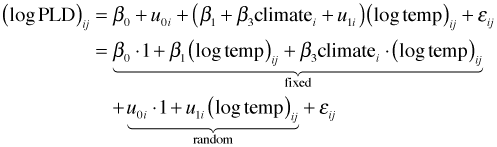
So only 3% of the species-level variation in the intercepts is explained by climate. This is consistent with the fact that adding climate.3 to the level-2 equation for the intercept did not significantly improve the model.

**Does climate.3 affect the population slope?**

To see if climate.3 affects the slope we enter it into the level-2 equation for the slope.



or in composite form



In R we specify this model by entering a term for the interaction between climate.3 and log(temp). Thus we need to include the term log(temp):climate.3  in the fixed part of the model expression.

model4<-lme(log(PLD)~log(temp) + log(temp):climate.3, random=~log(temp)|species, data=inverts, method='ML')

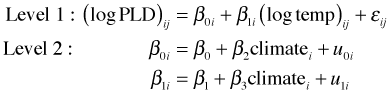
anova(model4,model2)

       Model df      AIC      BIC    logLik   Test   L.Ratio p-value  
model4     1  8 213.9420 241.0179 -98.97099   
model2     2  6 210.4372 230.7441 -99.21859 1 vs 2 0.4951964  0.7807

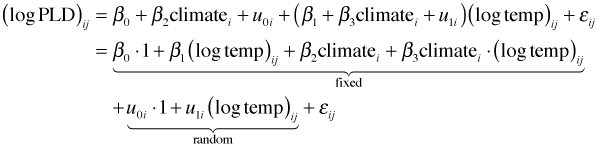
The likelihood ratio test fails to find the effect of climate on the slope to be statistically significant.

**Does climate.3 affect the population slopes and intercepts?**

If climate.3 affects both the slope and intercept it should be entered into both of the level-2 equations.



or in composite form



From the composite form the fixed part of the model should include log(temp), climate.3, and the interaction between the two. We can obtain this in R as follows.

fixed=log(PLD)~log(temp)+climate.3+log(temp):climate.3

or using R's shortcut notation

fixed=log(PLD)~log(temp)\*climate.3

where the \* notation automatically generates all three terms (plus the intercept).

model5<-lme(log(PLD)~log(temp)\*climate.3, random=~log(temp)|species, data=inverts, method='ML')

anova(model5,model2)

       Model df      AIC      BIC    logLik   Test  L.Ratio p-value  
model5     1 10 206.4266 240.2715 -93.21330   
model2     2  6 210.4372 230.7441 -99.21859 1 vs 2 12.01058  0.0173

So compared to a model without climate, we find evidence for a significant climate effect on the slopes and intercepts. This is a puzzling result. We've already seen that individually climate had no effect on the intercepts and no effect on the slopes, but when we let it affect them both simultaneously the effect became significant. To try to understand this further I compare the current model against each of the two previous models, a model in which climate is only allowed to affect the slopes (model4) and a model in which climate is only allowed to affect the intercepts (model3).

#test of effect on slopes if already affects intercepts

anova(model5,model4)

       Model df      AIC      BIC    logLik   Test  L.Ratio p-value  
model5     1 10 206.4266 240.2715 -93.21330   
model4     2  8 213.9420 241.0179 -98.97099 1 vs 2 11.51538  0.0032

#test of effect on intercepts if already affects slopes

anova(model5,model3)

       Model df      AIC      BIC    logLik   Test  L.Ratio p-value  
model5     1 10 206.4266 240.2715 -93.21330   
model3     2  8 213.8831 240.9590 -98.94154 1 vs 2 11.45648  0.0033

So now we find that climate is having a significant effect on both the slopes and intercepts, the exact opposite of the conclusion we drew before. So what's going on?

**Correlated random effects and the centering of predictors**

The basic problem is that the intercept and slope random effects in the random slope and intercept models we've been fitting are highly correlated. The correlation has been reported as part of the VarCorr output. For instance, in model2 the correlation was reported to be –0.92.

VarCorr(model2)

species = pdLogChol(log(temp))   
            Variance   StdDev    Corr    
(Intercept) 4.22070369 2.0544351 (Intr)  
log(temp)   0.28109629 0.5301851 -0.92   
Residual    0.02310586 0.1520061

A negative correlation between slopes and intercepts is to be expected but an extremely large correlation between the random effects can lead to convergence problems in multilevel models. The correlation here while large, is not large enough to cause numerical problems, but it is large enough to cause interpretation problems. In fact it leads to a complete confounding of the intercept and slope effects.

When one adds a level-2 predictor to a level-2 equation one is trying to account for some of the variability that is currently being soaked up by the random effects. Thus in a real sense the random effects and the level-2 fixed effects are in competition with one another. Adding a level-2 predictor to one level-2 equation but not the other when the random effects are highly correlated is a little like poking at a water balloon. It just shifts some of the variance to the other term. As a result because of the correlation of the random effects it's impossible to determine where exactly the effect of a level-2 predictor lies.

The solution is to center the predictor. Centering means subtracting off a constant value from the predictor. A judicious choice of centering constant can nearly uncorrelate the random slopes and intercepts. For these data the choice of constant that does this is log(35). To carry out the centering within the model rather than first creating a centered variable in the original data frame, we need to wrap the arithmetic by the I( ) function. Notice that the centered predictor is now the predictor everywhere in the model, not the original predictor log(temp). In particular the centered predictor, not the original predictor, also appears in the random argument.

model6<-lme(log(PLD)~I(log(temp)-log(35)), random=~I(log(temp)-log(35))|species, data=inverts, method="ML")

VarCorr(model6)

species = pdLogChol(I(log(temp) - log(35)))   
                       Variance   StdDev    Corr    
(Intercept)            0.65098009 0.8068334 (Intr)  
I(log(temp) - log(35)) 0.28109608 0.5301850 -0.005  
Residual               0.02310587 0.1520062

Observe that the correlation has been reduced to near zero. We'll discuss additional ramifications of centering in [lecture 28](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture28.htm#centering).

**Cited references**

* Pinheiro, J. C. & Bates, D. M. (2000) *Mixed-Effects Models in S and S-Plus.* (Springer-Verlag, New York).
* Singer, J. D. & Willett, J. B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* (Oxford University Press, Oxford, UK).
* Verbeke, G. & Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* (Springer-Verlag, New York).

**R Code**

A compact collection of most of the R code displayed in this document appears [here](http://www.unc.edu/courses/2010fall/ecol/563/001/notes/lecture27%20Rcode.html).

[Course Home Page](http://www.unc.edu/courses/2010fall/ecol/563/001/index.html)

|  |
| --- |
| Jack Weiss *Phone:* (919) 962-5930 *E-Mail:* jack\_weiss@unc.edu *Address:* Curriculum in Ecology, Box 3275, University of North Carolina, Chapel Hill, 27599 Copyright © 2010 Last Revised--December 14, 2010 URL: [http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm](http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture27.htm#lecture27) |