36-463/663: Multilevel & Hierarchical Models
Fall 2014
HW10 – Due Thu 01 Dec 2016

Announcements

- The datasets we will we will use for this assignment are already installed in the lme package. For example, to use the `Pastes` data set, after getting the lme package with `library(lme4)`, use the command `data(Pastes)`. Then you can use the `pastes` data frame in your workspace as usual, e.g. `str(Pastes)`, `lm(strength ~ batch, data = Pastes)`, etc.
- Please submit this hw as a pdf file on blackboard.

Exercises

1. *Crossed Random Effects.* The `Penicillin` data set [1] in `library(lme4)` is derived from Table 6.6, p. 144, of Davies and Goldsmith (1972), where it is described as coming from an investigation to

   > assess the variability between samples of penicillin by the B. subtilis method. In this test method a bulk-innoculated nutrient agar medium is poured into a Petri dish of approximately 90 mm. diameter, known as a plate. When the medium has set, six small hollow cylinders or pots (about 4 mm. in diameter) are cemented onto the surface at equally spaced intervals. A few drops of the penicillin solutions to be compared are placed in the respective cylinders, and the whole plate is placed in an incubator for a given time. Penicillin diffuses from the pots into the agar, and this produces a clear circular zone of inhibition of growth of the organisms, which can be readily measured. The diameter of the zone is related in a known way to the concentration of penicillin in the solution.

   The variables in the data set are

   - diameter: the diameter of the zone of inhibition due to penicillin
   - plate: the plate (petri dish) in which the diameters are measured
   - sample: one of the six areas in the dish in which the diameter of the zone of inhibition was measured. Each area correponds to a different penicillin solution; the same six solutions are used in each plate.

   Use str(Penicillin), View(Penicillin), etc. to familiarize yourself with the data.

   Because the samples are associated with the same six penicillin solutions in each plate, we will treat the sample and plate factors as crossed: for each plate, the same six different solution samples are tried.

[1] Bates, D.M. (2010). *lme4: Mixed-effects modeling with R.*

(a) Devise a plot (using ggplot2, lattice, or any other tools in R that you like) that shows the data nested within each unique pair of "plate" and "sample" levels, for this completely crossed design.

Provide your plot, and write a short paragraph explaining how to interpret this plot, in terms of the crossed factors "sample" and "plate".

(b) Fit the additive linear model

$$(diameter)_i = \beta_0 + \beta_{plate\ j[i]} + \beta_{sample\ k[i]} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$$

using `lm()` in R, where the β's are all fixed effects, $\beta_{plate\ j}$ is the effect of plate j and $\beta_{sample\ k}$ is the effect of sample k.

Provide a fitted model summary and one paragraph describing the model's fit and interpreation (make any other necessary plots or calculations in R).

(c) In an experiment like this, the effect of plate is not really of interest, so we could model it as a random effect. We might also be more concerned about the variability among the samples, rather than the effect of each particular sample type, so we could model sample as a random effect also. Use `lmer()` to fit a random intercept model to the data, with random effects for plate and sample:

$$\begin{array}{rcll}
\textbf{Level 1:} & & & \\
(diameter)_i & = & \beta_0 + \alpha_{plate\ j[i]} + \alpha_{sample\ k[i]} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) & i = \text{samples} \\
\textbf{Level 2:} & & & \\
\alpha_{plate\ j} & = & 0 + \eta_{plate\ j}, \qquad \eta_{plate\ j} \sim N(0, \tau^2_{plate}) & j = \text{plates} \\
\alpha_{sample\ k} & = & 0 + \eta_{sample\ k}, \qquad \eta_{sample\ k} \sim N(0, \tau^2_{sample}) & k = \text{samples}
\end{array}$$

where β_0 is a fixed effect (grand mean), and the random effects $\alpha_{plate\,j}$ and $\alpha_{samplek}$ are all centered at zero (so all the random effects represent deviations from the grand mean β_0. Since the factors that define the random effects are crossed, this is a *crossed random effects* model.

Provide a fitted model summary and one paragraph describing the model's fit and interpreation (make any other necessary plots or calculations in R).

(d) Use AIC and BIC to assess whether each of the random effects is needed in the mixed effects model. Is there enough variability in "sample" to keep the "sample" random effect? Is there enough variability in "plate" to keep the "plate" random effect?

Provide suitable R output and a sentence or two, to answer these questions.

2. *Nested Random Effects.* The `Pastes` data set in the lme4 package also comes from Davies and Goldsmith (1972, Table 6.5, p. 138). They describe the data as coming from

> deliveries of a chemical paste product contained in casks where, in addition to sampling and testing errors, there are variations in quality between deliveries… As a routine, three casks selected at random from each delivery were sampled and the samples were kept for reference… Ten of the delivery batches were sampled at random and two analytical tests carried out on each of the 30 samples.

(That is to say, two identical tests were performed on each sample.) The variables in the data set are:

- strength: strength of the paste sample in each analytical test.
- batch: which delivery batch the paste sample came from
- cask: which cask within the delivery batch the paste sample came from
- sample: another identifier for the cask and batch that each paste sample came from

Use str(Pastes), View(Pastes), etc. to familiarize yourself with the data.

Note that the casks are *nested* within the delivery batches; each sample can be in one and only one cask, and each cask can be in one and only one delivery batch. Also, the cask labels are not unique; they are repeated within each batch. *If we did not know that cask was nested within batch, this might lead us to believe that the cask and batch factors were crossed instead of nested!*

A random-intercepts model for this data with casks nested within batches might look like this:

$$
\begin{aligned}
&\textbf{Level 1:} \\
(strength)_i &= \alpha_j[i] + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad i = \text{(sample within cask)} \\
&\textbf{Level 2:} \\
\alpha_j &= \beta_{k[j]} + \eta_{2j}, \quad \eta_{2j} \sim N(0, \tau^2) \quad j = \text{(cask within delivery batch)} \\
&\textbf{Level 3:} \\
\beta_k &= \mu_0 + \eta_{3k}, \quad \eta_{3k} \sim N(0, \omega^2) \quad k = \text{(delivery batch)}
\end{aligned}
$$

Here, the α's are random effects at level 1, the β's are random effects at level 2, and μ_0 is a fixed effect (grand mean).

(a) Make a tree diagram (similar to the tree diagram we used for the Minnesota Radon data, or for the Oxide data) showing how each pair of observations is nested within each cask, and the casks are nested within each delivery batch. Label the branches of the tree with the appropriate parameters from the multi-level model above.

(b) Now devise a plot (using ggplot2, lattice, or any other R tools you like) to show the raw data in each cask, and each cask within each sample.

Provide the plot, and write a short paragraph explaining how to interpret this plot, and how it is different from the plot you drew in problem (1a).

(c) We want to build a random intercept model for this data. In `lmer()`, the "batch" random effect can be incorporated into the model as usual: `(1|batch)`. For the "cask" random effect, if we just used a term like `(1|cask)`, the effect would be to make batch and cask *crossed* instead of `nested`—since there are only 3 cask labels, we would only get one random effect for each label, rather than one for each cask within each batch. Instead, we can use the term `(1|cask:batch)` to force `lmer()` to generate separate random effects for each casket within each batch.

Provide a fitted model employing a random effect for batch and for casket nested within batch, and write a short paragraph describing the model's fit and interpretation (make any other necessary plots or calculations in R).

(d) The `sample` variable in the `Pastes` dataset provides unique labels for each cask within each batch. So, the model `lmer(strength ~ 1 + (1|sample) + (1|batch))` should produce the same fit and parameter estimates as your model in part (2c). Fit this model and verify that it produces the same answer as your answer for part (2c). *(If your answers aren't the same, go back and fix your model for part (2c))!*

Provide a summary of the fitted model as well as evidence that this model is identical in fit to the model in problem (2c).

(e) Use AIC and BIC to assess whether each of the random effects is needed in the mixed effects model. Is there enough variability in "batch" to keep the "batch" random effect? Is there enough variability in "cask:batch" within "batch" to keep the "cask" nested with "batch" random effect?

Provide suitable R output and a sentence or two, to answer these questions.

3. *Fake Data Testing.* In problem (2e), you used AIC and BIC to decide whether to keep the batch random effect. In this problem we will construct a fake data test to confirm (or disconfirm!) the approximate inference you made with AIC and BIC. Our approach is as follows.

Let $m_1, m_2, \ldots, m_B$ be the mean strength in each of the B batches, let $\overline{m} = \frac{1}{B}\sum_{b=1}^{B} m_b$ be their mean, and let $T(Pastes) = \frac{1}{B-1}\sum_{b=1}^{B}(m_b - \overline{m})^2$, the sample variance of the batch means. This will be our test statistic: if it is "small" we do not need the batch random effect; if it is "large" we do need the batch random effect. For 1000 fake data sets simulated from the H_0 model (without the batch random effect) we will calculate $T(fake.Pastes)$ and compare the observed value $T(Pastes)$ to the distribution of simulated $T(fake.Pastes)$ values, to determin if $T(Pastes)$ is "small" or "large".

In what follows we will use the `sim()` function from `library(arm)` to generate the fake data, since it is a little simpler and a little faster than JAGS.

(a) Compute $T(Pastes)$. Hint: You can get the vector of batch means using a combination of `split()`, `sapply()` and `mean()`; or by looking at `coef()` for the linear model `strength ~ batch - 1`.

(b) Use the `sim()` function from `library(arm)` to generate 1000 sets of plausible parameter values[2] from the model that omits the "batch" random effect. If `lmer2 <- lmer(strength ~ 1 + (1|batch:cask))` is the model that omits the batch random effect, for example, then `params <- sim(lmer2, n.sims=1000)` will generate the 1000 parameter sets you need.

(c) Now use the `fitted()` function to generate 1000 fake data sets of 60 observations each, from the 1000 sets of parameter values you just produced. Continuing from the previous part, `fake.data <- fitted(params,lmer2)` will do the trick.

(d) Now find $T($`fake.data[,i]`$)$ for each column i of the matrix of simulated fake data sets, and make a histogram of the result. Plot $T(Pastes)$ as a vertical red line on the histogram. Compute the fraction of $T($`fake.data[,i]`$)$ that are larger than $T(Pastes)$; this is the simulation p-value for the test of H_0. Does this fake-data test reject H_0? Explain why. Is this result consistent with the AIC and BIC results?

Provide your R code for parts (a), (b) and (c), and the graph, p-value and answers to the questions, for part (d).

[2] These are actually draws from a posterior distribution, calculated more efficiently than JAGS can do.