Mixed models in R using the lme4 package
Part 1: Linear mixed models with simple, scalar random effects

Douglas Bates

Madison
January 11, 2011

Contents

1 R packages and data in packages

R packages

- Packages incorporate functions, data and documentation.
- You can produce packages for private or in-house use or you can contribute your package to the Comprehensive R Archive Network (CRAN), `http://cran.R-project.org`
- We will be using the *lme4a* package from R-forge. Install it from the *Packages* menu item or with

```
> install.packages("lme4a", repos="http://r-forge.r-project.org")
```

- You only need to install a package once. If a new version becomes available you can update (see the menu item).
- To use a package in an R session you attach it using

```
> require(lme4a)
```

or

```
> library(lme4a)
```

(This usage causes widespread confusion of the terms "package" and "library".)

Accessing documentation

- To be added to CRAN, a package must pass a series of quality control checks. In particular, all functions and data sets must be documented. Examples and tests can also be included.
- The `data` function provides names and brief descriptions of the data sets in a package.

```
> data(package = "lme4a")
```

```
Data sets in package 'lme4a':

Dyestuff                Yield of dyestuff by batch
Dyestuff2               Yield of dyestuff by batch
Pastes                  Paste strength by batch and cask
Penicillin              Variation in penicillin testing
cake                    Breakage angle of chocolate cakes
cbpp                    Contagious bovine pleuropneumonia
sleepstudy              Reaction times in a sleep deprivation study
```

- Use `?` followed by the name of a function or data set to view its documentation. If the documentation contains an example section, you can execute it with the `example` function.

Effects - fixed and random

- Mixed-effects models, like many statistical models, describe the relationship between a *response* variable and one or more *covariates* recorded with it.
- The models we will discuss are based on a *linear predictor* expression incorporating *coefficients* that are estimated from the observed data.
- Coefficients associated with the levels of a categorical covariate are sometimes called the *effects* of the levels.
- When the levels of a covariate are fixed and reproducible (e.g. a covariate `sex` that has levels `male` and `female`) we incorporate them as fixed-effects parameters.
- When the levels of a covariate correspond to the particular observational or experimental units in the experiment we incorporate them as `random effects`.

2 The Dyestuff data and model

The Dyestuff data set

- The `Dyestuff`, `Penicillin` and `Pastes` data sets all come from the classic book *Statistical Methods in Research and Production*, edited by O.L. Davies and first published in 1947.

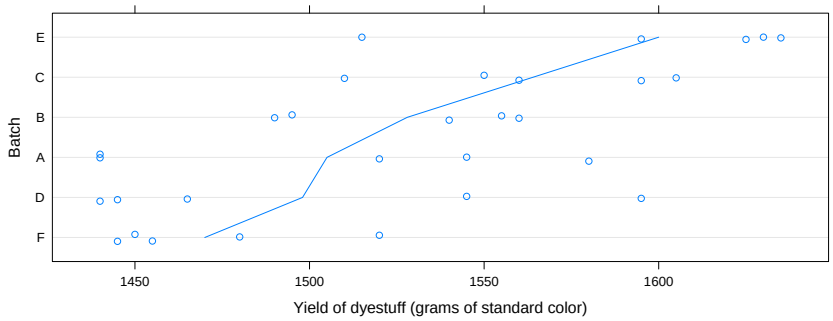
- The `Dyestuff` data are a balanced one-way classification of the `Yield` of dyestuff from samples produced from six `Batch`es of an intermediate product. See `?Dyestuff`.

```
> str(Dyestuff)

'data.frame':        30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 ..
 $ Yield: num  1545 1440 1440 1520 1580 ...

> summary(Dyestuff)

 Batch     Yield
 A:5   Min.   :1440
 B:5   1st Qu.:1469
 C:5   Median :1530
 D:5   Mean   :1528
 E:5   3rd Qu.:1575
 F:5   Max.   :1635
```

The effect of the batches

- To emphasize that `Batch` is categorical, we use letters instead of numbers to designate the levels.
- Because there is no inherent ordering of the levels of `Batch`, we will reorder the levels if, say, doing so can make a plot more informative.
- The particular batches observed are just a selection of the possible batches and are entirely used up during the course of the experiment.
- It is not particularly important to estimate and compare yields from these batches. Instead we wish to estimate the variability in yields due to batch-to-batch variability.
- The `Batch` factor will be used in *random-effects* terms in models that we fit.

Dyestuff data plot



- The line joins the mean yields of the six batches, which have been reordered by increasing mean yield.
- The vertical positions are jittered slightly to reduce overplotting. The lowest yield for batch A was observed on two distinct preparations from that batch.

A mixed-effects model for the dyestuff yield

```
> fm1 <- lmer(Yield ~ 1 + (1|Batch), Dyestuff)
> print(fm1)
```

```
Linear mixed model fit by REML ['merMod']
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
REML criterion at convergence: 319.6543
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1764     42.00
 Residual             2451     49.51
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      19.38    78.8
```

- Fitted model `fm1` has one fixed-effect parameter, the mean yield, and one random-effects term, generating a simple, scalar random effect for each level of `Batch`.

Extracting information from the fitted model

- `fm1` is an object of class `"merMod"`
- There are many *extractor* functions that can be applied to such objects.

```
> fixef(fm1)
```

```
(Intercept)
     1527.5
```

```
> ranef(fm1, drop = TRUE)
```

```
$Batch
        A         B         C         D         E         F
-17.60685   0.39126  28.56223 -23.08454  56.73319 -44.99529
attr(,"class")
[1] "ranef.mer"
```

```
> fitted(fm1)
```

```
 [1] 1509.9 1509.9 1509.9 1509.9 1509.9 1527.9 1527.9 1527.9 1527.9
[10] 1527.9 1556.1 1556.1 1556.1 1556.1 1556.1 1504.4 1504.4 1504.4
[19] 1504.4 1504.4 1584.2 1584.2 1584.2 1584.2 1584.2 1482.5 1482.5
[28] 1482.5 1482.5 1482.5
```

3 Definition of mixed-effects models

Definition of mixed-effects models

- Models with random effects are often written like

$$y_{ij} = \mu + b_i + \epsilon_{ij},\ b_i \sim \mathcal{N}(0, \sigma_b^2), \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, I, j = 1, \ldots, J_i$$

- This scalar notation quickly becomes unwieldy, degenerating into "subscript fests". We will use a vector/matrix notation.

- A mixed-effects model incorporates two vector-valued random variables: the response vector, $\boldsymbol{\mathcal{Y}}$, and the random effects vector, $\boldsymbol{\mathcal{B}}$. We observe the value, $\boldsymbol{y}$, of $\boldsymbol{\mathcal{Y}}$. We do not observe the value of $\boldsymbol{\mathcal{B}}$.

- In the models we will consider, the random effects are modeled as a multivariate Gaussian (or "normal") random variable, $\boldsymbol{\mathcal{B}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\boldsymbol{\theta}$ is a vector of *variance-component parameters*.

Linear mixed models

- The conditional distribution, $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b})$, depends on $\boldsymbol{b}$ only through its mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}=\boldsymbol{b}}$.

- The conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}=\boldsymbol{b}}$, depends on $\boldsymbol{b}$ and on the fixed-effects parameter vector, $\boldsymbol{\beta}$, through a *linear predictor* expression, $\boldsymbol{Zb} + \boldsymbol{X\beta}$. The *model matrices* $\boldsymbol{Z}$ and $\boldsymbol{X}$ are determined from the form of the model and the values of the covariates.

- In a *linear mixed model* the conditional distribution is a "spherical" multivariate Gaussian

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}) \sim \mathcal{N}(\boldsymbol{Zb} + \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$$

- The scalar σ is the *common scale parameter*; the dimension of $\boldsymbol{y}$ is n, $\boldsymbol{b}$ is q and $\boldsymbol{\beta}$ is p, hence $\boldsymbol{Z}$ is $n \times q$ and $\boldsymbol{X}$ is $n \times p$.

Simple, scalar random effects terms

- A term like `(1|Batch)` in an `lmer` formula is called a simple, scalar random-effects term.

- The expression on the right of the `"|"` operator (usually just the name of a variable) is evaluated as a factor, called the *grouping factor* for the term.

- Suppose we have k such terms with $n_i, i = 1, \ldots, k$ levels in the ith term's grouping factor. A scalar random-effects term generates one random effect for each level of the grouping factor. If all the random effects terms are scalar terms then $q = \sum_{i=1}^{k} n_i$.

- The model matrix $\boldsymbol{Z}$ is the horizontal concatenation of k matrices. For a simple, scalar term, the ith vertical slice, which has n_i columns, is the indicator columns for the n_i levels of the ith grouping factor.

Structure of the unconditional variance-covariance

- Just as the matrix $\boldsymbol{Z}$ is the horizontal concatenation of matrices generated by individual random-effects terms, the (unconditional) variance-covariance matrix, $\boldsymbol{\Sigma}$, is block-diagonal in k blocks. In other words, the unconditional distributions of random effects from different terms in the model are independent. (However, the conditional distributions, given the observed data, $(\boldsymbol{\mathcal{B}}|\boldsymbol{\mathcal{Y}} = \boldsymbol{y})$, are not independent.)
- If the ith term is a simple, scalar term then the ith diagonal block is a multiple of the identity, $\sigma_i^2 \boldsymbol{I}_{n_i}$.
- This means that unconditional distributions of random effects corresponding to different levels of the grouping factor are independent.

Model matrices for model fm1

- The formula for model `fm1` has a single fixed-effects term, `1`, and one simple, scalar random-effects term, `(1|Batch)`.
- The model matrix, $\boldsymbol{Z}$, whose transpose is stored in a slot called `Zt`, is the matrix of indicators for the six levels of `Batch`.
- The model matrix, $\boldsymbol{X}$, is 30×1. All its elements are unity.

```
> str(as.matrix(model.matrix(fm1)))
```

```
 num [1:30, 1] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:30] "1" "2" "3" "4" ...
  ..$ : chr "(Intercept)"
```

```
> fm1@re@Zt
```

```
6 x 30 sparse Matrix of class "dgCMatrix"
A 1 1 1 1 1 . . . . . . . . . . . . . . . . . . . . . . . . . .
B . . . . . 1 1 1 1 1 . . . . . . . . . . . . . . . . . . . . .
C . . . . . . . . . . 1 1 1 1 1 . . . . . . . . . . . . . . . .
D . . . . . . . . . . . . . . . 1 1 1 1 1 . . . . . . . . . . .
E . . . . . . . . . . . . . . . . . . . . . 1 1 1 1 1 . . . . .
F . . . . . . . . . . . . . . . . . . . . . . . . . . 1 1 1 1 1
```

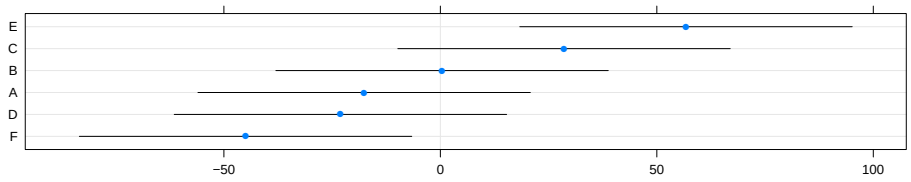
Conditional means of the random effects

- Technically we do not provide "estimates" of the random effects because they are not parameters.
- One answer to the question, "so what are those numbers provided by `ranef` anyway?" is that they are BLUPs (Best Linear Unbiased Predictors) of the random effects. The acronym is attractive but not very informative (what is a "linear unbiased predictor" and in what sense are these the "best"?). Also, the concept does not generalize.

- A better answer is that those values are the conditional means, $\mu_{\mathcal{B}|\mathcal{Y}=\boldsymbol{y}}$, evaluated at the estimated parameter values. Regrettably, we can only evaluate the conditional means for linear mixed models.
- However, these values are also the conditional modes and that concept does generalize to other types of mixed models.

Caterpillar plot for fm1

- For linear mixed models the conditional distribution of the random effects, given the data, written $(\mathcal{B}|\mathcal{Y}=\boldsymbol{y})$, is again a multivariate Gaussian distribution.
- We can evaluate the means and standard deviations of the individual conditional distributions, $(\mathcal{B}_j|\mathcal{Y}=\boldsymbol{y}), j=1,\ldots,q$. We show these in the form of a 95% prediction interval, with the levels of the grouping factor arranged in increasing order of the conditional mean.
- These are sometimes called "caterpillar plots".



REML estimates versus ML estimates

- The default parameter estimation criterion for linear mixed models is restricted (or "residual") maximum likelihood (REML).
- Maximum likelihood (ML) estimates (sometimes called "full maximum likelihood") can be requested by specifying `REML = FALSE` in the call to `lmer`.
- Generally REML estimates of variance components are preferred. ML estimates are known to be biased. Although REML estimates are not guaranteed to be unbiased, they are usually less biased than ML estimates.
- Roughly, the difference between REML and ML estimates of variance components is comparable to estimating σ^2 in a fixed-effects regression by $SSR/(n-p)$ versus SSR/n, where SSR is the residual sum of squares.
- For a balanced, one-way classification like the `Dyestuff` data, the REML and ML estimates of the fixed-effects are identical.

Re-fitting the model for ML estimates

```
> (fm1M <- update(fm1, REML = FALSE))
```

```
Linear mixed model fit by maximum likelihood ['merMod']
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
      AIC       BIC    logLik  deviance
 333.3271  337.5307 -163.6635  327.3271
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1388     37.26
 Residual             2451     49.51
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      17.69   86.33
```

(The extra parentheses around the assignment cause the value to be printed. Generally the results of assignments are not printed.)

Verbose fitting

- When fitting a large model or if the estimates of the variance components seem peculiar, it is a good idea to monitor the progress of the iterations in optimizing the deviance or the REML criterion.
- The optional argument `verbose = TRUE` causes `lmer` to print iteration information during the optimzation of the parameter estimates.
- The quantity being minimized is the *profiled deviance* or the *profiled REML criterion* of the model. The deviance is negative twice the log-likelihood. It is profiled in the sense that it is a function of $\boldsymbol{\theta}$ only — $\boldsymbol{\beta}$ and σ are at their conditional estimates.

Obtain the verbose output for fitting fm1

```
> invisible(update(fm1, verbose = TRUE))
```

```
npt = 3 , n =  1
rhobeg =  0.2 , rhoend =  2e-07
   0.020:   3:      319.672;0.800000
  0.0020:   5:      319.654;0.853320
 0.00020:   7:      319.654;0.848378
 2.0e-05:   9:      319.654;0.848378
 2.0e-06:  11:      319.654;0.848324
 2.0e-07:  13:      319.654;0.848324
At return
 16:     319.65428: 0.848324
```

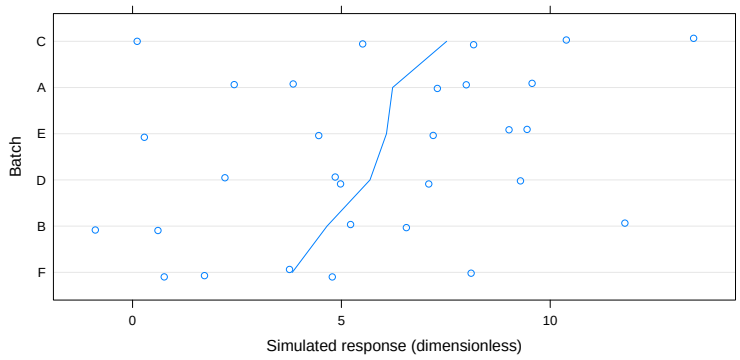
- The lines include a "trust region" radius, the cumulative number of function evaluations, the profiled deviance or profiled REML criterion (i.e. the quantity being minimized) and the components of the parameter vector, $\boldsymbol{\theta}$.

Estimates of variance components can be zero

- We have been careful to state the variance of the random effects is ≥ 0.
- For some data sets the maximum likelihood or REML estimate, $\widehat{\sigma_b^2}$ ends up as exactly zero. That is, the optimal parameter value is on the boundary of the region of allowable values.
- Box and Tiao (1973) provide simulated data with a structure like the `Dyestuff` data illustrating this.

```
> str(Dyestuff2)

'data.frame':        30 obs. of  2 variables:
 $ Batch: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 2 2 2 2 ..
 $ Yield: num  7.3 3.85 2.43 9.57 7.99 ...
```

Plot of the Dyestuff2 data



- For these data the batch-to-batch variability is not large compared to the within-batch variability.

Fitting the model to Dyestuff2

```
> (fm1A <- lmer(Yield ~ 1 + (1|Batch), Dyestuff2, REML=FALSE))

Linear mixed model fit by maximum likelihood ['merMod']
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff2
     AIC      BIC   logLik deviance
168.8730 173.0766 -81.4365 162.8730
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept)  0.00    0.000
 Residual             13.35    3.653
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)    5.666      0.667   8.494
```

A trivial mixed-effects model is a fixed-effects model

- The mixed model `fm1A` with an estimated variance $\widehat{\sigma_b^2} = 0$ is equivalent to a model with only fixed-effects terms.

```
> summary(lm1 <- lm(Yield ~ 1, Dyestuff2))
```

```
Call:
lm(formula = Yield ~ 1, data = Dyestuff2)
Residuals:
    Min      1Q  Median      3Q     Max
-6.5576 -2.9006 -0.3006  2.4854  7.7684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6656     0.6784   8.352 3.32e-09

Residual standard error: 3.716 on 29 degrees of freedom
```

```
> logLik(lm1)
```

```
'log Lik.' -81.43652 (df=2)
```

Recap of the Dyestuff model

- The model is fit as

```
lmer(formula = Yield ~ 1 + (1 | Batch), data = Dyestuff)
```

- There is one random-effects term, `(1|Batch)`, in the model formula. It is a simple, scalar term for the grouping factor `Batch` with $n_1 = 6$ levels. Thus $q = 6$.
- The model matrix $\boldsymbol{Z}$ is the 30×6 matrix of indicators of the levels of `Batch`.
- The variance-covariance matrix, $\boldsymbol{\Sigma}$, is a nonnegative multiple of the 6×6 identity matrix, $\boldsymbol{I}_6$.
- The fixed-effects parameter vector, $\boldsymbol{\beta}$, is of length $p = 1$. All the elements of the 30×1 model matrix $\boldsymbol{X}$ are unity.

4 Crossed random-effects grouping: Penicillin

The Penicillin data (see also the ?Penicillin description)

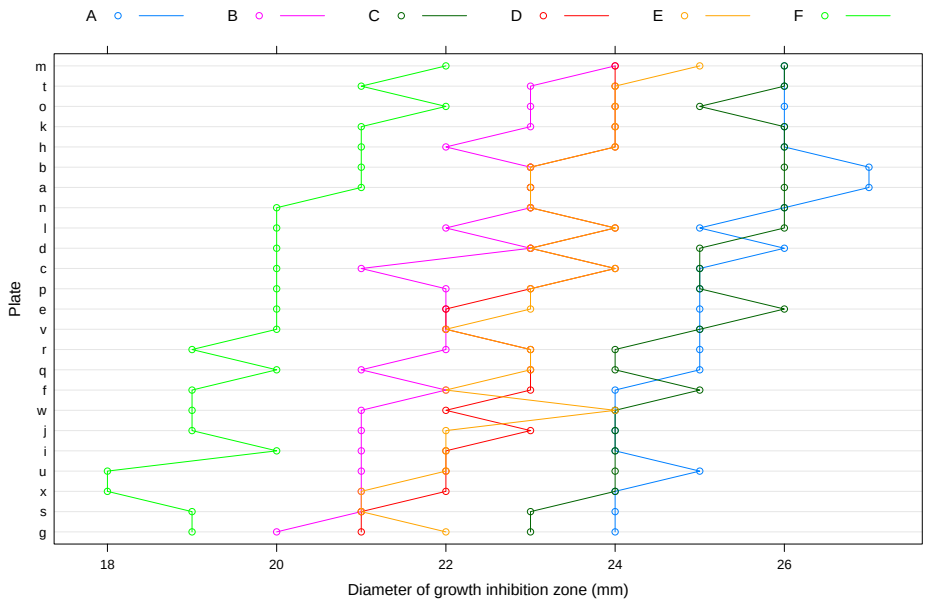
```
> str(Penicillin)
```

```
'data.frame':        144 obs. of  3 variables:
 $ diameter: num  27 23 26 23 23 21 27 23 26 23 ...
 $ plate   : Factor w/ 24 levels "a","b","c","d",..: 1 1 1 1 1 1 2 ..
 $ sample  : Factor w/ 6 levels "A","B","C","D",..: 1 2 3 4 5 6 1 2..
```

```
> xtabs(~ sample + plate, Penicillin)

      plate
sample a b c d e f g h i j k l m n o p q r s t u v w x
     A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     B 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     C 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     D 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     E 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
     F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

- These are measurements of the potency (measured by the diameter of a clear area on a Petri dish) of penicillin samples in a balanced, unreplicated two-way crossed classification with the test medium, `plate`.

Penicillin data plot



Model with crossed simple random effects for Penicillin

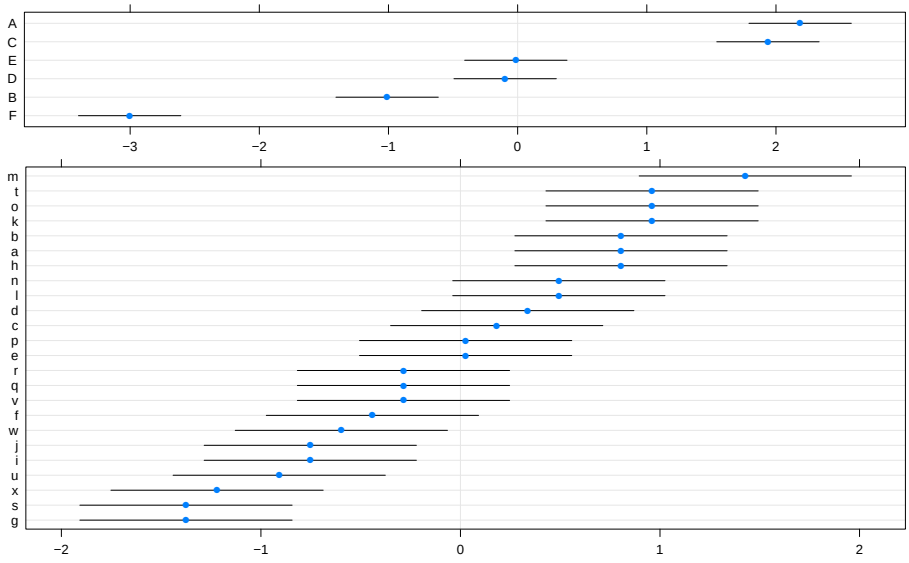
```
> (fm2 <- lmer(diameter ~ 1 + (1|plate) + (1|sample), Penicillin))

Linear mixed model fit by REML ['merMod']
Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
   Data: Penicillin
REML criterion at convergence: 330.8606
Random effects:
 Groups   Name        Variance Std.Dev.
 plate    (Intercept) 0.7169   0.8467
 sample   (Intercept) 3.7309   1.9316
 Residual             0.3024   0.5499
Number of obs: 144, groups: plate, 24; sample, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  22.9722     0.8086   28.41
```

Random effects for fm2

- The model for the $n = 144$ observations has $p = 1$ fixed-effects parameter and $q = 30$ random effects from $k = 2$ random effects terms in the formula.

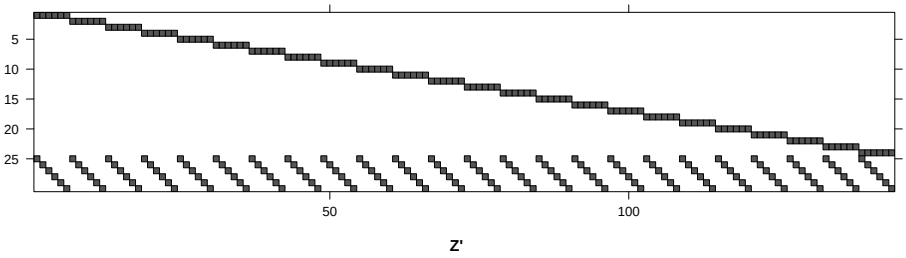
```
> ranef(fm2, drop = TRUE)

$plate
        a         b         c         d         e         f
 0.804547  0.804547  0.181672  0.337391  0.025953 -0.441203
        g         h         i         j         k         l
-1.375516  0.804547 -0.752641 -0.752641  0.960266  0.493109
        m         n         o         p         q         r
 1.427422  0.493109  0.960266  0.025953 -0.285484 -0.285484
        s         t         u         v         w         x
-1.375516  0.960266 -0.908360 -0.285484 -0.596922 -1.219797
$sample
        A         B         C         D         E         F
 2.187058 -1.010476  1.937899 -0.096895 -0.013842 -3.003744

attr(,"class")
[1] "ranef.mer"
```

Prediction intervals for random effects



Model matrix Z for fm2

- Because the model matrix $\boldsymbol{Z}$ is generated from $k = 2$ simple, scalar random effects terms, it consists of two sets of indicator columns.
- The structure of $\boldsymbol{Z}^{\mathsf{T}}$ is shown below. (Generally we will show the transpose of these model matrices - they fit better on slides.)



Models with crossed random effects

- Many people believe that mixed-effects models are equivalent to hierarchical linear models (HLMs) or "multilevel models". This is not true. The `plate` and `sample` factors in `fm2` are crossed. They do not represent levels in a hierarchy.
- There is no difficulty in defining and fitting models with crossed random effects (meaning random-effects terms whose grouping factors are crossed). However, fitting models with crossed random effects can be somewhat slower.
- The crucial calculation in each `lmer` iteration is evaluation of a $q \times q$ sparse, lower triangular, Cholesky factor, $\boldsymbol{L}(\boldsymbol{\theta})$, derived from $\boldsymbol{Z}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Crossing of grouping factors increases the number of nonzeros in $\boldsymbol{L}(\boldsymbol{\theta})$ and causes some "fill-in" of $\boldsymbol{L}$ relative to $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$.

All HLMs are mixed models but not vice-versa

- Even though Raudenbush and Bryk (2002) do discuss models for crossed factors in their HLM book, such models are not hierarchical.
- Experimental situations with crossed random factors, such as "subject" and "stimulus", are common. We can, and should, model such data according to its structure.
- In longitudinal studies of subjects in social contexts (e.g. students in classrooms or in schools) we almost always have partial crossing of the subject and the context factors, meaning that, over the course of the study, a particular student may be observed in more than one class but not all students are observed in all classes. The student and class factors are neither fully crossed nor strictly nested.
- For longitudinal data, "nested" is only important if it means "nested across time". "Nested at a particular time" doesn't count.
- `lme4` handles fully or partially crossed factors gracefully.

Recap of the Penicillin model

- The model formula is

```
diameter ~ 1 + (1 | plate) + (1 | sample)
```

- There are two random-effects terms, `(1|plate)` and `(1|sample)`. Both are simple, scalar random effects terms, with $n_1 = 24$ and $n_2 = 6$ levels, respectively. Thus $q = q_1 n_1 + q_2 n_2 = 30$.
- The model matrix $\boldsymbol{Z}$ is the 144×30 matrix created from two sets of indicator columns.
- The relative variance-covariance matrix, $\boldsymbol{\Sigma}$, is block diagonal in two blocks that are nonnegative multiples of identity matrices.
- The fixed-effects parameter vector, $\boldsymbol{\beta}$, is of length $p = 1$. All the elements of the 144×1 model matrix $\boldsymbol{X}$ are unity.

5 Nested random-effects grouping: Pastes

The Pastes data (see also the ?Pastes description)

```
> str(Pastes)
```

```
'data.frame':        60 obs. of  6 variables:
 $ strength: num  62.8 62.6 60.1 62.3 62.7 63.1 60 61.4 57.5 56.9 ...
 $ batch   : Factor w/ 10 levels "A","B","C","D",..: 1 1 1 1 1 1 2 ..
 $ cask    : Factor w/ 3 levels "a","b","c": 1 1 2 2 3 3 1 1 2 2 ...
 $ sample  : Factor w/ 30 levels "A:a","A:b","A:c",..: 1 1 2 2 3 3 ..
 $ bb      : Factor w/ 10 levels "E","J","I","B",..: 9 9 9 9 9 9 4 ..
  ..- attr(*, "scores")= num [1:10(1d)] 62.3 59.3 62.1 59.7 55.9 ...
  .. ..- attr(*, "dimnames")=List of 1
  .. .. ..$ : chr  "A" "B" "C" "D" ...
 $ ss      : Factor w/ 30 levels "A:b","A:a","A:c",..: 2 2 1 1 3 3 ..
  ..- attr(*, "scores")= num [1:30(1d)] 5 9 5 4 2 9 7 4 10 3 ...
  .. ..- attr(*, "dimnames")=List of 1
  .. .. ..$ : chr  "E:b" "I:a" "E:a" "D:a" ...
```

```
> xtabs(~ batch + sample, Pastes, sparse = TRUE)
```

```
10 x 30 sparse Matrix of class "dgCMatrix"
A 2 2 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
B . . . 2 2 2 . . . . . . . . . . . . . . . . . . . . . . . . . .
C . . . . . . 2 2 2 . . . . . . . . . . . . . . . . . . . . . . .
D . . . . . . . . . 2 2 2 . . . . . . . . . . . . . . . . . . . .
E . . . . . . . . . . . . 2 2 2 . . . . . . . . . . . . . . . . .
F . . . . . . . . . . . . . . . 2 2 2 . . . . . . . . . . . . . .
G . . . . . . . . . . . . . . . . . . . 2 2 2 . . . . . . . . . .
H . . . . . . . . . . . . . . . . . . . . . . 2 2 2 . . . . . . .
I . . . . . . . . . . . . . . . . . . . . . . . . . 2 2 2 . . .
J . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2 2 2
```

Structure of the Pastes data

- The `sample` factor is nested within the `batch` factor. Each sample is from one of three casks selected from a particular batch.
- Note that there are 30, not 3, distinct samples.

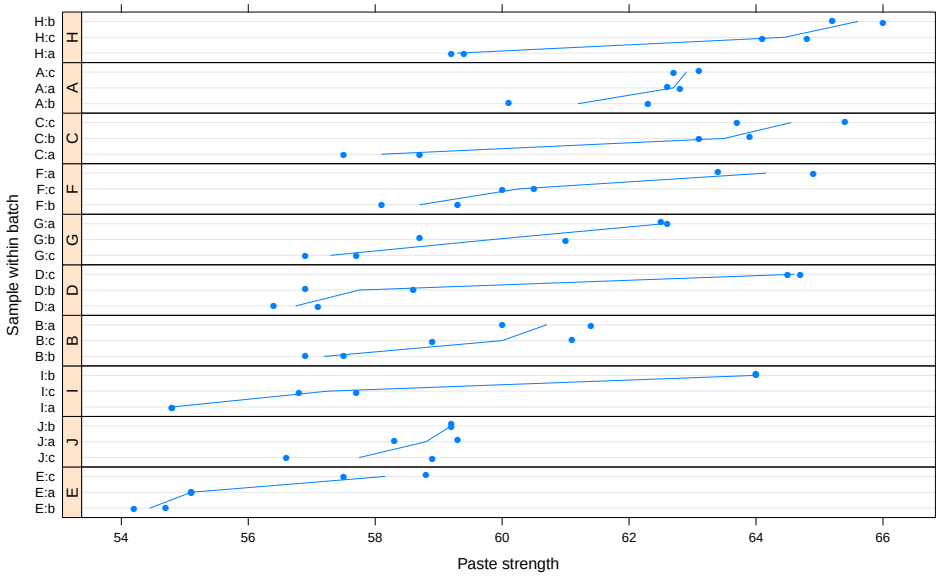
- We can label the casks as 'a', 'b' and 'c' but then the `cask` factor by itself is meaningless (because cask 'a' in batch 'A' is unrelated to cask 'a'in batches 'B', 'C', ...). The `cask` factor is only meaningful within a `batch`.
- Only the `batch` and `cask` factors, which are apparently crossed, were present in the original data set. `cask` may be described as being nested within `batch` but that is not reflected in the data. It is *implicitly nested*, not explicitly nested.
- You can save yourself a lot of grief by immediately creating the explicitly nested factor. The recipe is

```
> Pastes <- within(Pastes, sample <- factor(batch:cask))
```

Avoid implicitly nested representations

- The `lme4` package allows for very general model specifications. It does not require that factors associated with random effects be hierarchical or "multilevel" factors in the design.
- The same model specification can be used for data with nested or crossed or partially crossed factors. Nesting or crossing is determined from the structure of the factors in the data, not the model specification.
- You can avoid confusion about nested and crossed factors by following one simple rule: ensure that different levels of a factor in the experiment correspond to different labels of the factor in the data.
- Samples were drawn from 30, not 3, distinct casks in this experiment. We should specify models using the `sample` factor with 30 levels, not the `cask` factor with 3 levels.

Pastes data plot

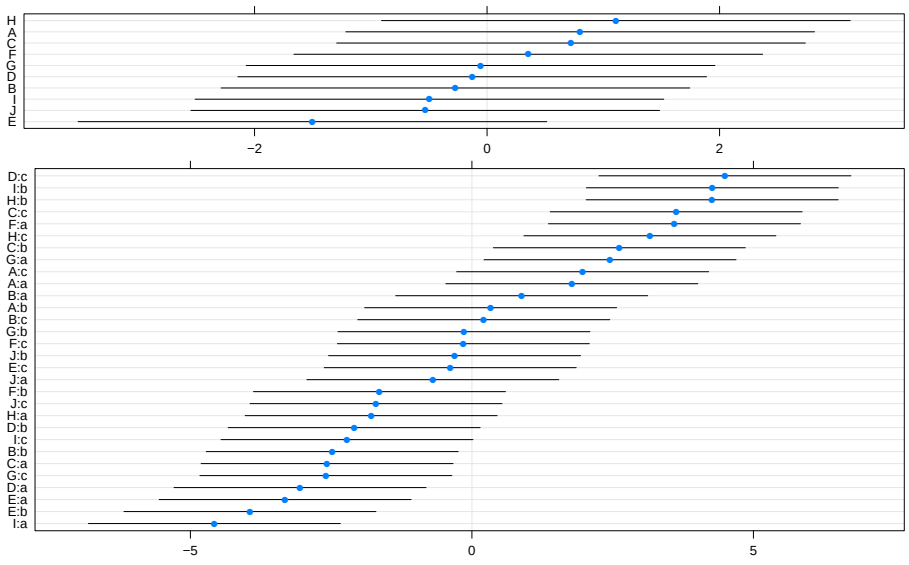


A model with nested random effects

```
> (fm3 <- lmer(strength ~ 1 + (1|batch) + (1|sample), Pastes))
```

```
Linear mixed model fit by REML ['merMod']
Formula: strength ~ 1 + (1 | batch) + (1 | sample)
   Data: Pastes
REML criterion at convergence: 246.9907
Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 8.434    2.9041
 batch    (Intercept) 1.657    1.2874
 Residual             0.678    0.8234
Number of obs: 60, groups: sample, 30; batch, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.6769   88.72
```

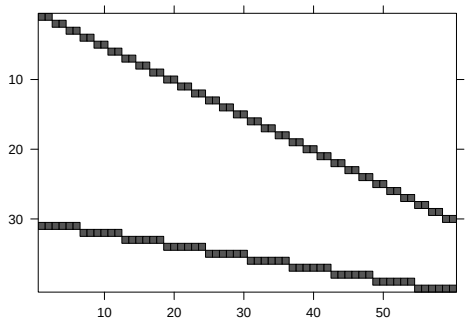
Random effects from model fm3



Batch-to-batch variability is low compared to sample-to-sample.

Dimensions and relationships in fm3

- There are $n = 60$ observations, $p = 1$ fixed-effects parameter, $k = 2$ simple, scalar random-effects terms ($q_1 = q_2 = 1$) with grouping factors having $n_1 = 30$ and $n_2 = 10$ levels.
- Because both random-effects terms are simple, scalar terms, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is block-diagonal in two diagonal blocks of sizes 30 and 10, respectively. $\boldsymbol{Z}$ is generated from two sets of indicators.



Eliminate the random-effects term for batch?

- We have seen that there is little batch-to-batch variability beyond that induced by the variability of samples within batches.
- We can fit a reduced model without that term and compare it to the original model.
- Somewhat confusingly, model comparisons from likelihood ratio tests are obtained by calling the `anova` function on the two models. (Put the simpler model first in the call to `anova`.)
- Sometimes likelihood ratio tests can be evaluated using the REML criterion and sometimes they can't. Instead of learning the rules of when you can and when you can't, it is easiest always to refit the models with `REML = FALSE` before comparing.

Comparing ML fits of the full and reduced models

```
> fm3M <- update(fm3, REML = FALSE)
> fm4M <- lmer(strength ~ 1 + (1|sample),
+              Pastes, REML = FALSE)
> anova(fm4M, fm3M)

Data: Pastes
Models:
fm4M: strength ~ 1 + (1 | sample)
fm3M: strength ~ 1 + (1 | batch) + (1 | sample)
     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
fm4M  3 254.40 260.69 -124.20   248.40
fm3M  4 255.99 264.37 -124.00   247.99 0.4072      1     0.5234
```

p-values of LR tests on variance components

- The likelihood ratio is a reasonable criterion for comparing these two models. However, the theory behind using a χ^2 distribution with 1 degree of freedom as a reference distribution for this test statistic does not apply in this case. The null hypothesis is on the boundary of the parameter space.
- Even at the best of times, the p-values for such tests are only approximate because they are based on the asymptotic behavior of the test statistic. To carry the argument further, all results in statistics are based on models and, as George Box famously said, "All models are wrong; some models are useful."

LR tests on variance components (cont'd)

- In this case the problem with the boundary condition results in a p-value that is larger than it would be if, say, you compared this likelihood ratio to values obtained for data simulated from the null hypothesis model. We say these results are "conservative".
- As a rule of thumb, the p-value for the χ^2 test on a simple, scalar term is roughly twice as large as it should be.
- In this case, dividing the p-value in half would not affect our conclusion.

Updated model, REML estimates

```
> (fm4 <- update(fm4M, REML = TRUE))
```

```
Linear mixed model fit by REML ['merMod']
Formula: strength ~ 1 + (1 | sample)
   Data: Pastes
REML criterion at convergence: 247.6484
Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 9.977    3.1586
 Residual             0.678    0.8234
Number of obs: 60, groups: sample, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.5864   102.4
```

Recap of the analysis of the Pastes data

- The data consist of $n = 60$ observations on $n_1 = 30$ samples nested within $n_2 = 10$ batches.
- The data are labelled with a `cask` factor with 3 levels but that is an implicitly nested factor. Create the explicit factor `sample` and ignore `cask` from then on.
- Specification of a model for nested factors is exactly the same as specification of a model with crossed or partially crossed factors — provided that you avoid using implicitly nested factors.
- In this case the `batch` factor was inert — it did not "explain" substantial variability in addition to that attributed to the `sample` factor. We therefore prefer the simpler model.
- At the risk of "beating a dead horse", notice that, if we had used the `cask` factor in some way, we would still need to create a factor like `sample` to be able to reduce the model. The `cask` factor is only meaningful within `batch`.

This is all very nice, but ...

- These methods are interesting but the results are not really new. Similar results are quoted in *Statistical Methods in Research and Production*, which is a very old book.
- The approach described in that book is actually quite sophisticated, especially when you consider that the methods described there, based on observed and expected mean squares, are for hand calculation — in pre-calculator days!
- Why go to all the trouble of working with sparse matrices and all that if you could get the same results with paper and pencil? The one-word answer is *balance*.
- Those methods depend on the data being balanced. The design must be completely balanced and the resulting data must also be completely balanced.
- Balance is fragile. Even if the design is balanced, a single missing or questionable observation destroys the balance. Observational studies (as opposed to, say, laboratory experiments) cannot be expected to yield balanced data sets.
- Also, the models involve only simple, scalar random effects and do not incorporate covariates.

6 Incorporating fixed-effects terms: classroom

Structure of the classroom data

- The `classroom` data are a cross-section of students within classes within schools. The `mathgain` variable is the difference in mathematics achievement scores in grade 1 and kindergarten.
- These data are quite unbalanced. The distribution of the number of students observed per classroom is

```
> xtabs( ~ xtabs(~ classid, classroom))
```

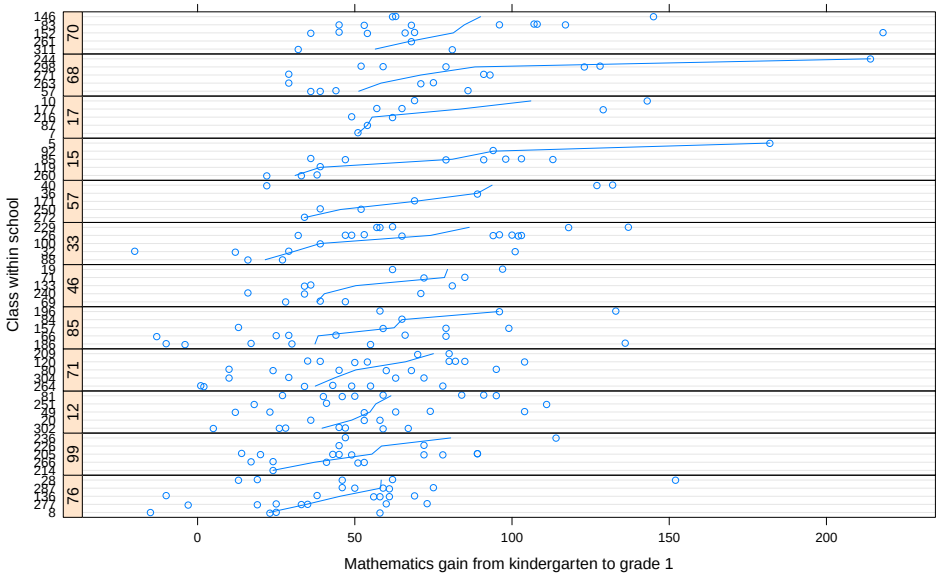
```
xtabs(~classid, classroom)
 1  2  3  4  5  6  7  8  9 10
42 53 53 61 39 31 14 13  4  2
```

- Similarly, the distribution of the number of classes observed per school is

```
> table(xtabs(~ schoolid,
+             unique(subset(classroom, select = c(classid, schoolid)))))
```

```
 1  2  3  4  5  9
13 34 26 21 12  1
```

Twelve schools, each with 5 classrooms



Simple, "unconditional" model for the classroom data

```
> (fm5 <- lmer(mathgain ~ 1 + (1|classid) + (1|schoolid), classroom))

Linear mixed model fit by REML ['merMod']
Formula: mathgain ~ 1 + (1 | classid) + (1 | schoolid)
   Data: classroom
REML criterion at convergence: 11768.76
Random effects:
 Groups   Name        Variance Std.Dev.
 classid  (Intercept)   99.23   9.961
 schoolid (Intercept)   77.49   8.803
 Residual             1028.23  32.066
Number of obs: 1190, groups: classid, 312; schoolid, 107

Fixed effects:
            Estimate Std. Error t value
(Intercept)   57.427      1.443   39.79
```

Some comments on the "unconditional" model

- In the multilevel modeling literature a model such as `fm5` that does not incorporate fixed-effects terms for demographic characteristics of the student, class or school, is called an "unconditional" model.
- Notice that the dominant level of variability is the residual variability. It is unlikely that random effects for both classes and schools are needed when modeling these data.
- We have seen in Exercises 2 that there seem to be trends with respect to the `minority` factor and the `mathkind` score but no overall trends with respect to `sex`.
- A coefficient for a continuous covariate, such as `mathkind`, or for fixed, reproducible levels of a factor like `sex` or `minority` is incorporated in the fixed-effects terms.

Model-building approach

- Note that these unbalanced data have, for the most part, very few classes per school (sometimes as few as 1) and very few students per class (also sometimes as few as 1). Under these circumstances, it is optimistic to expect to be able to partition the variability across students, classes and schools.
- We should consider adding fixed-effects terms and perhaps removing one of the random-effects terms.
- We will start by incorporating fixed-effects terms then revisit the need for both random-effects terms.
- We will begin with the fixed-effects terms adopted as a final model in chapter 4 of West, Welch and Gałecki (2007).
- For brevity, we only display the output of model fits as this contains enough information to reconstruct the call to `lmer`.

A model with fixed-effects terms

```
Linear mixed model fit by REML ['merMod']
Formula: mathgain ~ 1 + mathkind + minority + sex + ses + housepov + (1 |      classid) + (1 | schoolid)
   Data: classroom
REML criterion at convergence: 11378.06
Random effects:
 Groups   Name        Variance Std.Dev.
 classid  (Intercept)  81.56    9.031
 schoolid (Intercept)  77.76    8.818
 Residual             734.42   27.100
Number of obs: 1190, groups: classid, 312; schoolid, 107

Fixed effects:
             Estimate Std. Error t value
(Intercept) 285.05797   11.02077  25.866
mathkind     -0.47086    0.02228 -21.133
minorityY    -7.75587    2.38499  -3.252
sexF         -1.23459    1.65743  -0.745
ses           5.23971    1.24497   4.209
housepov    -11.43920    9.93736  -1.151
```

Where are the p-values?!!

- The first thing that most users notice is that there are no p-values for the fixed-effects coefficients! Calculating a p-value for $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ is not as straightforward as it may seem. The ratio called a "t value" in the output does not have a Student's T distribution under the null hypothesis.
- For simple models fit to small, balanced data sets one can calculate a p-value. Not so for unbalanced data. When the number of groups and observations are large, approximations don't matter — you can consider the ratio as having a standard normal distribution.

- The only time that you can calculate an "exact" p-value and the difference between this and the standard normal dist'n is important is for small, balanced data sets, which are exactly the cases that appear in text books. People get very, very upset if the values calculated by the software don't agree perfectly with the text book answers.
- Here, just say a coefficient is "significant" if $|t| > 2$.

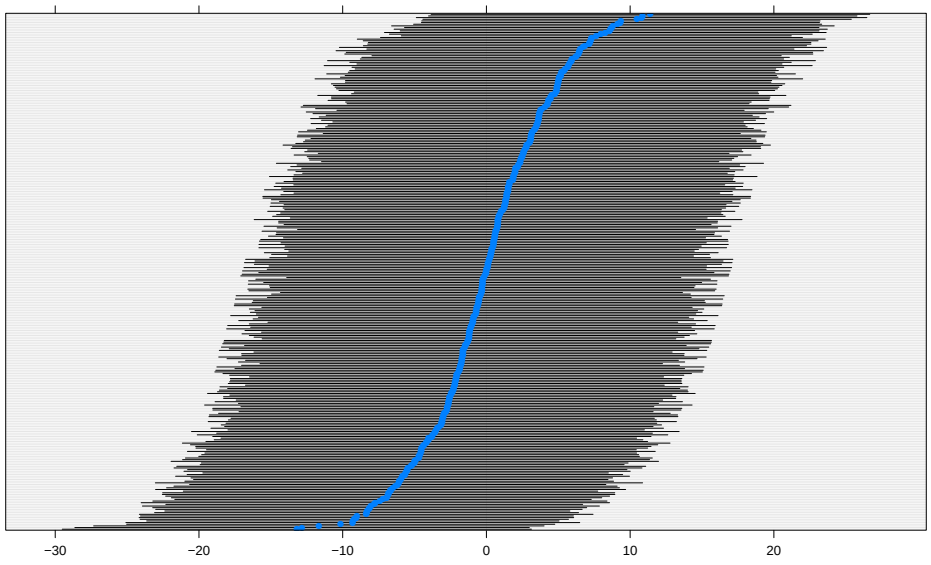
Removing the insignificant term for sex

```
Linear mixed model fit by REML ['merMod']
Formula: mathgain ~ 1 + mathkind + minority + ses + housepov + (1 | classid) +      (1 | schoolid)
   Data: classroom
REML criterion at convergence: 11381.46
Random effects:
 Groups   Name        Variance Std.Dev.
 classid  (Intercept)  81.1     9.005
 schoolid (Intercept)  77.6     8.809
 Residual             734.5    27.101
Number of obs: 1190, groups: classid, 312; schoolid, 107

Fixed effects:
             Estimate Std. Error t value
(Intercept) 284.70353   11.00868  25.862
mathkind     -0.47137    0.02227 -21.170
minorityY    -7.78049    2.38387  -3.264
ses           5.25701    1.24456   4.224
housepov    -11.50245    9.92827  -1.159
```

Removing the insignificant term for housepov

```
Linear mixed model fit by REML ['merMod']
Formula: mathgain ~ mathkind + minority + ses + (1 | classid) + (1 | schoolid)
   Data: classroom
REML criterion at convergence: 11389.22
Random effects:
 Groups   Name        Variance Std.Dev.
 classid  (Intercept)  82.84    9.102
 schoolid (Intercept)  75.04    8.662
 Residual             734.61   27.104
Number of obs: 1190, groups: classid, 312; schoolid, 107

Fixed effects:
             Estimate Std. Error t value
(Intercept) 282.41932   10.84061  26.052
mathkind     -0.47032    0.02225 -21.137
minorityY    -8.29086    2.33888  -3.545
ses           5.36462    1.24067   4.324
```

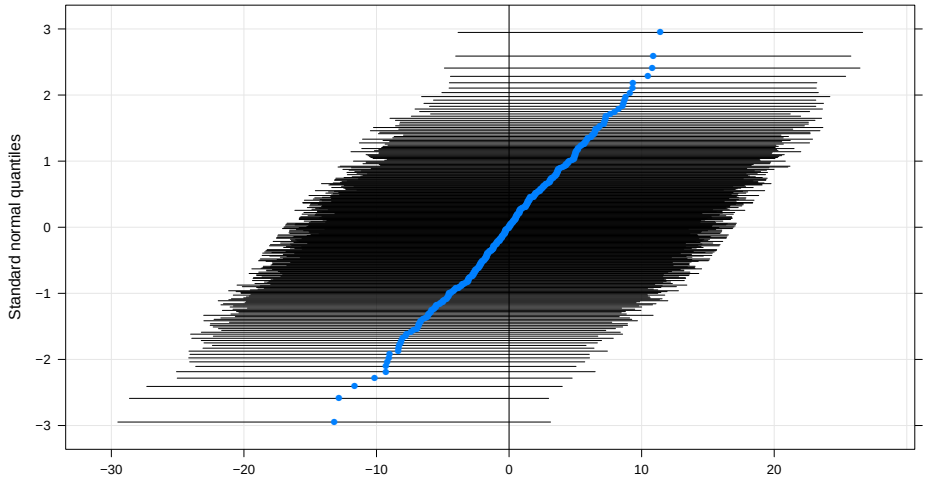
Prediction intervals on random effects for class



Normal probability plot of random effects for class

With many levels of the grouping factor, use a normal probability plot of the prediction intervals for the random effects.

```
> qqmath(ranef(fm8, post = TRUE))$classid
```



Refit without random effects for class

```
Linear mixed model fit by maximum likelihood ['merMod']
Formula: mathgain ~ mathkind + minority + ses + (1 | schoolid)
   Data: classroom
     AIC      BIC   logLik deviance
11415.50 11445.99 -5701.75 11403.50
Random effects:
 Groups   Name        Variance Std.Dev.
```

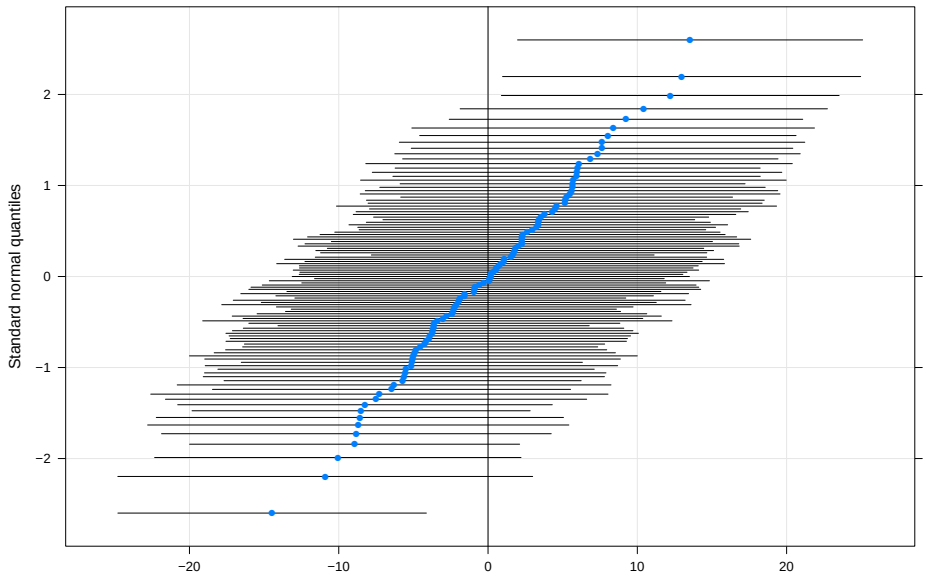


Figure 1: Normal probability plot of random effects for school

```
 schoolid (Intercept)  97.87    9.893
 Residual             789.14   28.092
Number of obs: 1190, groups: schoolid, 107

Fixed effects:
             Estimate Std. Error t value
(Intercept) 282.30407   10.90141  25.896
mathkind     -0.47045    0.02237 -21.026
minorityY    -7.79580    2.35039  -3.317
ses           5.51955    1.24921   4.418
```

Check if random effects for class are significant

```
> fm8M <- update(fm8, REML = FALSE)
> anova(fm9M, fm8M)
```

```
Data: classroom
Models:
fm9M: mathgain ~ mathkind + minority + ses + (1 | schoolid)
fm8M: mathgain ~ mathkind + minority + ses + (1 | classid) + (1 | schoolid)
     Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
fm9M  6 11416 11446 -5701.7    11404
fm8M  7 11406 11441 -5695.8    11392 11.967      1  0.0005415
```

- Contrary to what we saw in the plots, the random-effects term for `classid` is significant even in the presence of the `schoolid` term
- Part of the reason for this inconsistency is our incorporating 312 random effects at a "cost" of 1 parameter. In some way we are undercounting the number of degrees of freedom added to the model with this term.

A large observational data set

- A large U.S. university (not mine) provided data on the grade point score (`gr.pt`) by student (`id`), instructor (`instr`) and department (`dept`) from a 10 year period. I regret that I cannot make these data available to others.
- These factors are unbalanced and partially crossed.

```
> str(anon.grades.df)

'data.frame':   1721024 obs. of  9 variables:
 $ instr   : Factor w/ 7964 levels "10000","10001",..: 1 1 1 1 1 1 1 1 1 ...
 $ dept    : Factor w/ 106 levels "AERO","AFAM",..: 43 43 43 43 43 43 43 43 43 ...
 $ id      : Factor w/ 54711 levels "900000001","900000002",..: 12152 1405 23882 18875 18294 20922 4150 13540 549
 $ nclass  : num  40 29 33 13 47 49 37 14 21 20 ...
 $ vgpa    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ rawai   : num   2.88 -1.15 -0.08 -1.94  3.00 ...
 $ gr.pt   : num  4 1.7 2 0 3.7 1.7 2 4 2 2.7 ...
 $ section : Factor w/ 70366 levels "19959 AERO011A001",..: 18417 18417 18417 18417 9428 18417 18417 9428 9428 94
 $ semester: num  19989 19989 19989 19989 19972 ...
```

A preliminary model

```
Linear mixed model fit by REML
Formula: gr.pt ~ (1 | id) + (1 | instr) + (1 | dept)
   Data: anon.grades.df
     AIC     BIC   logLik deviance REMLdev
 3447389 3447451 -1723690  3447374 3447379
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.3085   0.555
 instr    (Intercept) 0.0795   0.282
 dept     (Intercept) 0.0909   0.301
 Residual             0.4037   0.635
Number of obs: 1685394, groups: id, 54711; instr, 7915; dept, 102

Fixed effects:
            Estimate Std. Error t value
(Intercept)   3.1996     0.0314     102
```

Comments on the model fit

- $n = 1685394$, $p = 1$, $k = 3$, $n_1 = 54711$, $n_2 = 7915$, $n_3 = 102$, $q_1 = q_2 = q_3 = 1$, $q = 62728$
- This model is sometimes called the "unconditional" model in that it does not incorporate covariates beyond the grouping factors.
- It takes less than an hour to fit an "unconditional" model with random effects for student (`id`), instructor (`inst`) and department (`dept`) to these data.
- Naturally, this is just the first step. We want to look at possible time trends and the possible influences of the covariates.
- This is an example of what "large" and "unbalanced" mean today. The size of the data sets and the complexity of the models in mixed modeling can be formidable.

7 A model fit to a large data set

A model fit to a large data set (by today's standards)

- Harold Doran recently fit a linear mixed model to the annual achievement test results for the last 4 years in one of the United States. There were $n = 5212017$ observations on a total of $n_1 = 1876788$ students and $n_2 = 47480$ teachers.
- The models had simple, scalar random effects for student and for teacher resulting in $q = 1924268$ (i.e. nearly 2 million!)
- There were a total of $p = 29$ fixed-effects parameters.
- At present Harold needed to fit the model to a subset and only evaluate the conditional means for all the students and teachers but we should be able to get around that limitation and actually fit the model to all these responses and random effects.
- I don't know of other software that can be used to fit a model this large.

Size of the decomposition for this large model

- The limiting factor on the memory size in such a model is the Cholesky factor $\boldsymbol{L}(\boldsymbol{\theta})$.
- In this case the `x` slot is itself over 1GB in size and the `i` slot is over 0.5 GB.
- These are close to an inherent limit on atomic R objects (the range of an index into an atomic object cannot exceed 2^{31}.

```
> str(L)
```

```
Formal class 'dCHMsimpl' [package "Matrix"] with 10 slots
  ..@ x       : num [1:174396181] 1.71 2.16 1.4 1.32 2.29 ...
  ..@ p       : int [1:1924269] 0 2 4 5 7 9 10 12 14 15 ...
  ..@ i       : int [1:174396181] 0 2 1 2 2 3 5 4 5 5 ...
  ..@ nz      : int [1:1924268] 2 2 1 2 2 1 2 2 1 2 ...
  ..@ nxt     : int [1:1924270] 1 2 3 4 5 6 7 8 9 10 ...
  ..@ prv     : int [1:1924270] 1924269 0 1 2 3 4 5 6 7 8 ...
  ..@ colcount: int [1:1924268] 2 2 1 2 2 1 2 2 1 2 ...
  ..@ perm    : int [1:1924268] 1922843 1886519 134451 1921046 1893309 183471 1912388 1888309 196670 1922626 ...
  ..@ type    : int [1:4] 2 1 0 1
  ..@ Dim     : int [1:2] 1924268 192426
```

Recap of simple, scalar random-effects terms

- For `lmer` a simple, scalar random effects term is of the form `(1|F)`.
- The number of random effects generated by the ith such term is the number of levels, n_i, of `F` (after dropping "unused" levels — those that do not occur in the data. The idea of having such levels is not as peculiar as it may seem if, say, you are fitting a model to a subset of the original data.)
- Such a term contributes n_i columns to $\boldsymbol{Z}$. These columns are the indicator columns of the grouping factor.

- Such a term contributes a diagonal block $\sigma_i^2 \boldsymbol{I}_{n_i}$ to $\boldsymbol{\Sigma}$. The multipliers σ_i^2 can be different for different terms. The term contributes exactly one element (which happens to be σ_i/σ) to $\boldsymbol{\theta}$.