Fitting Mixed-Effects Models Using the lme4 Package in R

Deepayan Sarkar

Fred Hutchinson Cancer Research Center

18 September 2008

Organizing data in R

- Standard rectangular data sets (columns are variables, rows are observations) are stored in *R* as *data frames*.
- The columns can be *numeric* variables (e.g. measurements or counts) or *factor* variables (categorical data) or *ordered* factor variables. These types are called the *class* of the variable.
- Useful functions to inspect data frames (and many other R objects):
  - `str()` provides concise description of the structure
  - `summary()` summarizes each variable according to its class
  - `head()` and `tail()` extracts the first few or last few rows

R packages

- Packages incorporate functions, data and documentation.
- We will be using the *lme4* package from CRAN, which can be installed from the *Packages* menu item or with
  `> install.packages("lme4")`
- To use the package in an R session, it must be attached; e.g.,
  `> library("lme4")`

Accessing documentation

- All functions and datasets in an *R* package must be documented. Examples and tests are also often included.
- The `data` function provides names and brief descriptions of the data sets in a package.

```
> data(package = "lme4")
```

```
Data sets in package 'lme4':

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

Lattice graphics

- One of the strengths of R is its graphics capabilities.
- There are several styles of graphics in R. Trellis graphics, as implemented in the *lattice* package, is well-suited to the type of data we will be discussing.

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`.

The Dyestuff data set

```
> 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 2 ..
 $ Yield: num  1545 1440 1440 1520 1580 ...
```

```
> summary(Dyestuff)
```

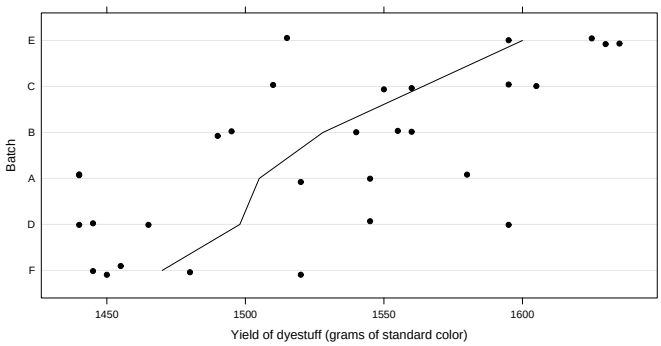
```
 Batch     Yield
 A:5   Min.   :1440
 B:5   1st Qu.:1469
 C:5   Median :1530
 D:5   Mean   :1528
 E:5   3rd Qu.:1575
 F:5   Max.   :1635
```

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.
- 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

```
> dotplot(reorder(Batch, Yield) ~ Yield, Dyestuff,
          type = c("p", "a"), jitter.y = TRUE, ylab = "Batch",
          xlab = "Yield of dyestuff (grams of standard color)")
```



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 yield

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

```
Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
   AIC   BIC logLik deviance REMLdev
 325.7 329.9 -159.8    327.4   319.7
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1763.7   41.996
 Residual             2451.3   49.511
Number of obs: 30, groups: Batch, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      19.38   78.81
```

- The 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 model

- `fm1` is an object of class `"mer"` (mixed-effects representation).
- Many *extractor* functions can be applied to such objects.

```
> fixef(fm1)
```

```
(Intercept)
     1527.5
```

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

```
$Batch
        A         B         C         D         E         F
-17.60597   0.39124  28.56079 -23.08338  56.73033 -44.99302
```

```
> fitted(fm1)
```

```
 [1] 1509.9 1509.9 1509.9 1509.9 1509.9 1527.9 1527.9 1527.9
 [9] 1527.9 1527.9 1556.1 1556.1 1556.1 1556.1 1556.1 1504.4
[17] 1504.4 1504.4 1504.4 1504.4 1584.2 1584.2 1584.2 1584.2
[25] 1584.2 1482.5 1482.5 1482.5 1482.5 1482.5
```

Definition of linear mixed-effects models

- A mixed-effects model incorporates two vector-valued random variables: the response, $\mathcal{Y}$, and the random effects, $\mathcal{B}$. We observe the value, $\mathbf{y}$, of $\mathcal{Y}$. We do not observe the value of $\mathcal{B}$.
- In a *linear mixed-effects model* the conditional distribution, $\mathcal{Y}|\mathcal{B}$, and the marginal distribution, $\mathcal{B}$, are independent, multivariate normal (or "Gaussian") distributions,

$$(\mathcal{Y}|\mathcal{B}=\mathbf{b}) \sim \mathcal{N}\left(\mathbf{X}\beta+\mathbf{Z}\mathbf{b},\sigma^2\mathbf{I}\right), \quad \mathcal{B} \sim \mathcal{N}\left(\mathbf{0},\sigma^2\Sigma\right), \quad (\mathcal{Y}|\mathcal{B}) \perp \mathcal{B}.$$

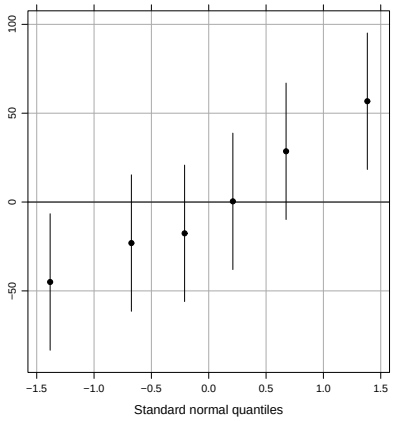
- The scalar σ is the *common scale parameter*; the p-dimensional β is the *fixed-effects parameter*; the $n \times p$ $\mathbf{X}$ and the $n \times q$ $\mathbf{Z}$ are known, fixed *model matrices*; and the $q \times q$ *relative variance-covariance matrix* $\Sigma(\theta)$ is a positive semidefinite, symmetric $q \times q$ matrix that depends on the parameter θ.

Conditional modes of the random effects

- Technically, the reported random effects are not "estimates", because the random effects are not parameters.
- They can be viewed as the conditional means, $\mathrm{E}[\mathcal{B}|\mathcal{Y}=\mathbf{y}]$, evaluated at the estimated parameters. We can only evaluate the conditional means for linear mixed models.
- These values are also the conditional modes and that concept does generalize to other types of mixed models.
- For linear mixed models we can evaluate the conditional standard deviations of these random variables and plot a prediction interval. These intervals can be arranged in a normal probability plot, sometimes called a "caterpillar plot".

Caterpillar plot for fm1

```
> qqmath(ranef(fm1, postVar = TRUE), strip = FALSE)$Batch
```



Mixed-effects model formulas

- In `lmer` the model is specified by the `formula` argument. As in most R model-fitting functions, this is the first argument.
- The model formula consists of two expressions separated by the $\sim$ symbol.
- The expression on the left, typically the name of a variable, is evaluated as the response.
- The right-hand side consists of one or more *terms* separated by '+' symbols.
- A random-effects term consists of two expressions separated by the vertical bar, ('|'), symbol (read as "given" or "by"). Typically, such terms are enclosed in parentheses.
- The expression on the right of the '|' is evaluated as a factor, which we call the *grouping factor* for that term.

Simple, scalar random-effects terms

- In a *simple, scalar* random-effects term, the expression on the left of the '|' is '1'. Such a term generates one random effect (i.e. a scalar) for each level of the grouping factor.
- Each random-effects term contributes a set of columns to $\mathbf{Z}$. For a simple, scalar r.e. term these are the indicator columns for the levels of the grouping factor.

Verbose fitting

- The optional argument `verbose = TRUE` causes `lmer` to print iteration information during the optimization of the parameter estimates.
- The quantity being minimized is the *profiled deviance* of the model. The deviance is negative twice the log-likelihood. It is profiled in the sense that it is a function of θ only — β and σ are at their conditional estimates.

Obtain the verbose output for fitting fm1

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

  0:     319.76562: 0.730297
  1:     319.73553: 0.962418
  2:     319.65736: 0.869480
  3:     319.65441: 0.844020
  4:     319.65428: 0.848469
  5:     319.65428: 0.848327
  6:     319.65428: 0.848324
```

- The first number on each line is the iteration count — iteration 0 is at the starting value for $\boldsymbol{\theta}$.
- The second number is the profiled deviance — the criterion to be minimized at the estimates.
- The third and subsequent numbers are the parameter vector $\boldsymbol{\theta}$.

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.

Re-fitting the model for ML estimates

```
> update(fm1, REML = FALSE)

Linear mixed model fit by maximum likelihood
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff
   AIC   BIC logLik deviance REMLdev
 333.3 337.5 -163.7    327.3   319.7
Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 1388.1   37.258
 Residual             2451.3   49.511
Number of obs: 30, groups: Batch, 6

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

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 $\mathbf{Z}$ is the 30×6 matrix of indicators of the levels of `Batch`.
- The fixed-effects parameter vector, $\boldsymbol{\beta}$, is of length $p = 1$. All the elements of the 30×1 model matrix $\mathbf{X}$ are unity.

The Penicillin data

- Potency (measured as 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`.

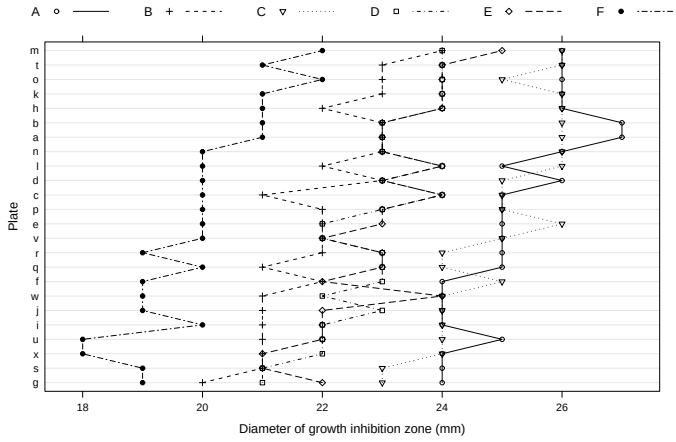
```
> str(Penicillin)

'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 2 2 2
 $ sample  : Factor w/ 6 levels "A","B","C","D",..: 1 2 3 4 5 6 1 2 3 4

> 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
```

Penicillin data plot



Model with crossed simple random effects for Penicillin

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

```
Linear mixed model fit by REML
Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
   Data: Penicillin
   AIC   BIC logLik deviance REMLdev
 338.9 350.7 -165.4    332.3   330.9
Random effects:
 Groups   Name        Variance Std.Dev.
 plate    (Intercept) 0.71691  0.84671
 sample   (Intercept) 3.73030  1.93140
 Residual             0.30242  0.54992
Number of obs: 144, groups: plate, 24; sample, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  22.9722     0.8085   28.41
```

Fixed and 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.

```
> fixef(fm2)
```

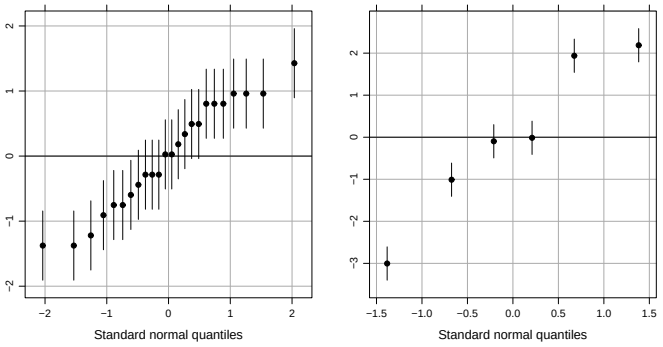
```
(Intercept)
     22.972
```

```
> 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
 2.187057 -1.010476  1.937898 -0.096895 -0.013842 -3.003742
```

Prediction intervals for random effects



Likelihood ratio tests

- Nested models can be compared by the `anova` function
- It is safer to use `REML = FALSE` when doing LR tests

```
> fm2ML <- lmer(diameter ~ 1 + (1|plate) + (1|sample),
                data = Penicillin, REML = FALSE)
> fm3ML <- lmer(diameter ~ 1 + (1|plate),
                data = Penicillin, REML = FALSE)
> anova(fm3ML, fm2ML)
```

```
Data: Penicillin
Models:
fm3ML: diameter ~ 1 + (1 | plate)
fm2ML: diameter ~ 1 + (1 | plate) + (1 | sample)
      Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm3ML  3  617.71  626.62 -305.86
fm2ML  4  340.19  352.07 -166.09 279.52      1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Models with crossed random effects

- All hierarchical linear models (HLMs) or "multilevel models" are mixed models but not vice-versa
- 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).
- Crossing of random effects can affect the speed with which a model can be fit.

Models with crossed random effects

- Experimental situations with crossed random factors, such as "subject" and "stimulus", are common. We can and should model such data according to its structure.
- The `lme4` package is different from most other software for fitting mixed models in that it handles fully crossed and partially crossed random effects 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 ($q_1 = q_2 = 1$) 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 $\mathbf{Z}$ is the 144×30 matrix created from two sets of indicator columns.
- The fixed-effects parameter vector, $\boldsymbol{\beta}$, is of length $p = 1$. All the elements of the 144×1 model matrix $\mathbf{X}$ are unity.

More complex models

- Models with covariates for random effects
- Generalized linear mixed models (`glmer`)
- Nonlinear mixed models (`nlmer`)
- See `help(lmer)`