Fitting Linear Mixed-Effects Models Using the lme4 Package in R

Douglas Bates

University of Wisconsin - Madison
and R Development Core Team
<Douglas.Bates@R-project.org>

University of Potsdam
August 7, 2008

Outline

Outline

Outline

Organizing and plotting data; simple, scalar random effects

Models for longitudinal data

Singular variance-covariance matrices

Unbalanced, non-nested data sets

Interactions of grouping factors and other covariates

Evaluating the log-likelihood

Outline

Organizing and plotting data; simple, scalar random effects

Models for longitudinal data

Singular variance-covariance matrices

Unbalanced, non-nested data sets

Interactions of grouping factors and other covariates

Evaluating the log-likelihood

Outline

Organizing and plotting data; simple, scalar random effects

Models for longitudinal data

Singular variance-covariance matrices

Unbalanced, non-nested data sets

Interactions of grouping factors and other covariates

Evaluating the log-likelihood

Outline

Organizing and plotting data; simple, scalar random effects

Models for longitudinal data

Singular variance-covariance matrices

Unbalanced, non-nested data sets

Interactions of grouping factors and other covariates

Evaluating the log-likelihood

Web sites associated with the workshop

www.stat.wisc.edu/∼bates/PotsdamGLMM Materials for the course

www.R-project.org Main web site for the R Project

cran.R-project.org Comprehensive R Archive Network primary site

cran.us.R-project.org Main U.S. mirror for CRAN

cran.R-project.org/web/views/Psychometrics.html Psychometrics task view within CRAN

R-forge.R-project.org R-Forge, development site for many public R packages. This is also the URL of the repository for installing the development versions of the `lme4` and `Matrix` packages, if you are so inclined.

lme4.R-forge.R-project.org development site for the `lme4` package

Outline

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.
- The `str` function provides a concise description of the structure of a data set (or any other class of object in R). The `summary` function summarizes each variable according to its class. Both are highly recommended for routine use.
- Entering just the name of the data frame causes it to be printed. For large data frames use the `head` and `tail` functions to view the first few or last few rows.

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.us.R-project.org`
- We will be using the *lme4* package from CRAN. Install it from the *Packages* menu item or with

  ```
  > install.packages("lme4")
  ```
- 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(lme4)
  ```

  or

  ```
  > library(lme4)
  ```

  (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 = "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
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.

Lattice graphics

- One of the strengths of R is its graphics capabilities.
- There are several styles of graphics in R. The style in Deepayan Sarkar's *lattice* package is well-suited to the type of data we will be discussing.
- I will not show every piece of code used to produce the data graphics. The code is available in the script files for the slides (and sometimes in the example sections of the data set's documentation).
- Deepayan's book, *Lattice: Multivariate Data Visualization with R* (Springer, 2008) provides in-depth documentation and explanations of lattice graphics.
- I also recommend Phil Spector's book, *Data Manipulation with R* (Springer, 2008).

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

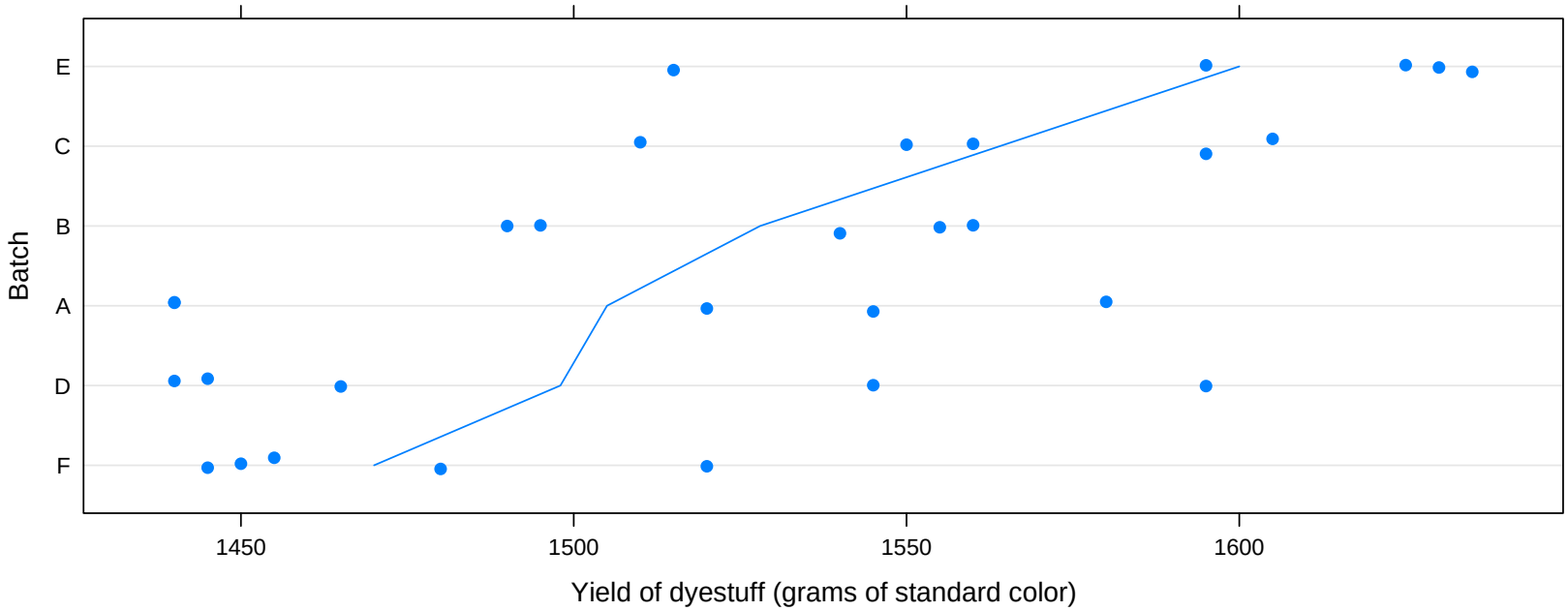
```
> summary(Dyestuff)
```

```
 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
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
```

- 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 `"mer"` (mixed-effects representation).
- 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
-17.60597   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, $\boldsymbol{y}$, 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,

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

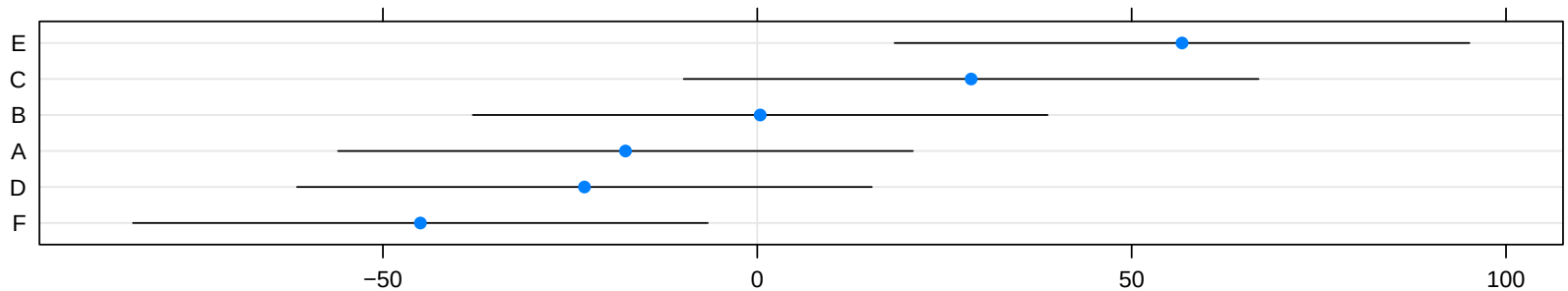
- The scalar σ is the *common scale parameter*; the p-dimensional $\boldsymbol{\beta}$ is the *fixed-effects parameter*; the $n \times p$ $\boldsymbol{X}$ and the $n \times q$ $\boldsymbol{Z}$ are known, fixed *model matrices*; and the $q \times q$ *relative variance-covariance matrix* $\boldsymbol{\Sigma}(\boldsymbol{\theta})$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

Conditional modes 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 anyway?" is that they are BLUPs (Best Linear Unbiased Predictors) but that answer is not informative and the concept does not generalize.
- A better answer is that those values are the conditional means, $\mathsf{E}[\mathcal{B}|\mathcal{Y} = y]$, evaluated at the estimated parameters. 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 we can evaluate the means and standard deviations of the conditional distributions $\mathcal{B}_j|\mathcal{Y}, j = 1, \ldots, q$. 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".



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. The transpose of the `Batch` indicators is

```
> with(Dyestuff, as(Batch, "sparseMatrix"))
```

```
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
```

Formulation of the marginal variance matrix

- In addition to determining $\boldsymbol{Z}$, the random effects terms determine the form and parameterization of the relative variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.
- The parameterization is based on a modified "LDL′" Cholesky factorization

$$\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{S}\boldsymbol{S}'\boldsymbol{T}'$$

  where $\boldsymbol{T}$ is a $q \times q$ unit lower *T*riangular matrix and $\boldsymbol{S}$ is a $q \times q$ diagonal *S*cale matrix with nonnegative diagonal elements.
- $\boldsymbol{\Sigma}$, $\boldsymbol{T}$ and $\boldsymbol{S}$ are all block-diagonal, with blocks corresponding to the random-effects terms.
- The diagonal block of $\boldsymbol{T}$ for a scalar random effects term is the identity matrix, $\boldsymbol{I}$, and the block in $\boldsymbol{S}$ is a nonnegative multiple of $\boldsymbol{I}$.

Verbose fitting, extracting T and S

- 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* 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.
- If you want to see exactly how the parameters $\boldsymbol{\theta}$ generate $\boldsymbol{\Sigma}$, use `expand` to obtain a list with components `sigma`, `T` and `S`. The list also contains a permutation matrix `P` whose role we will discuss later.
- $\boldsymbol{T}$, $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ can be very large but are always highly patterned. The `image` function can be used to examine their structure.

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 $\boldsymbol{\theta}$.

Extract T and S

- As previously indicated, $\boldsymbol{T}$ and $\boldsymbol{S}$ from `fm1` are boring.

```
> efm1 <- expand(fm1)
> efm1$S
```

```
6 x 6 diagonal matrix of class "ddiMatrix"
     [,1]    [,2]    [,3]    [,4]    [,5]    [,6]
[1,] 0.84823       .       .       .       .       .
[2,]       . 0.84823       .       .       .       .
[3,]       .       . 0.84823       .       .       .
[4,]       .       .       . 0.84823       .       .
[5,]       .       .       .       . 0.84823       .
[6,]       .       .       .       .       . 0.84823
```

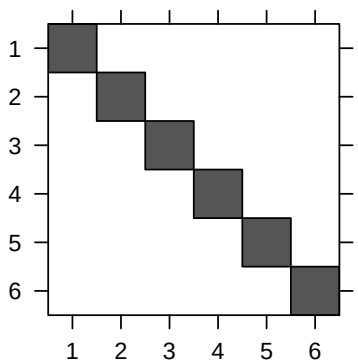
```
> efm1$T
```

```
6 x 6 sparse Matrix of class "dtCMatrix"
[1,] 1 . . . . .
[2,] . 1 . . . .
[3,] . . 1 . . .
[4,] . . . 1 . .
[5,] . . . . 1 .
[6,] . . . . . 1
```

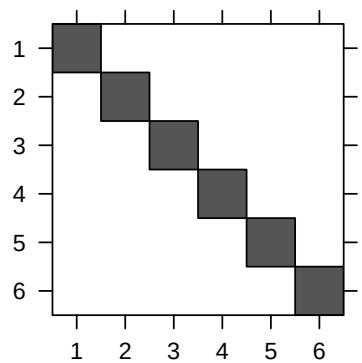
Reconstructing Σ

```
> (fm1S <- tcrossprod(efm1$T %*% efm1$S))
```

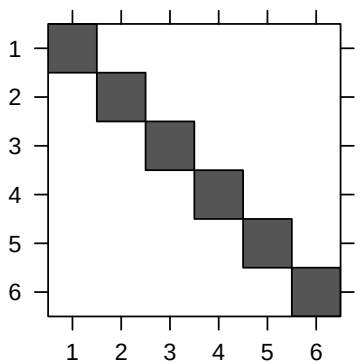
```
6 x 6 sparse Matrix of class "dsCMatrix"
[1,] 0.71949 .       .       .       .       .
[2,] .       0.71949 .       .       .       .
[3,] .       .       0.71949 .       .       .
[4,] .       .       .       0.71949 .       .
[5,] .       .       .       .       0.71949 .
[6,] .       .       .       .       .       0.71949
```



T



S



Σ

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
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
```

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

Recap of the Dyestuff model

- The model is fit as

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

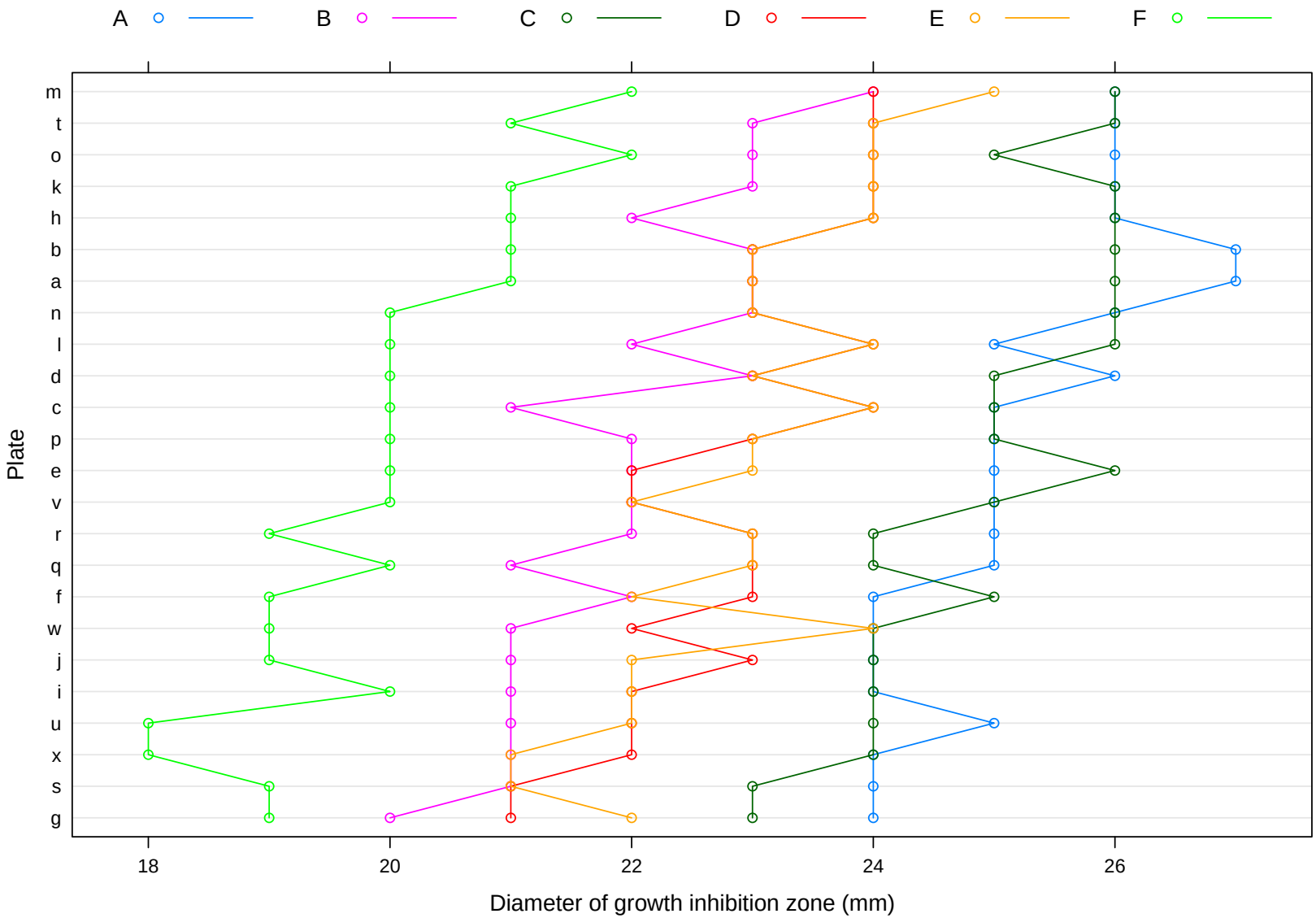
- There is one random-effects term, `(1|Batch)`, in the model formula. It is a simple, scalar term for the grouping factor `Batch` with $n_1 = 6$ levels. Thus $q = 6$.
- The model matrix $\boldsymbol{Z}$ is the 30×6 matrix of indicators of the levels of `Batch`.
- The relative variance-covariance matrix, $\boldsymbol{\Sigma}$, 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.

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 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
```

- 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

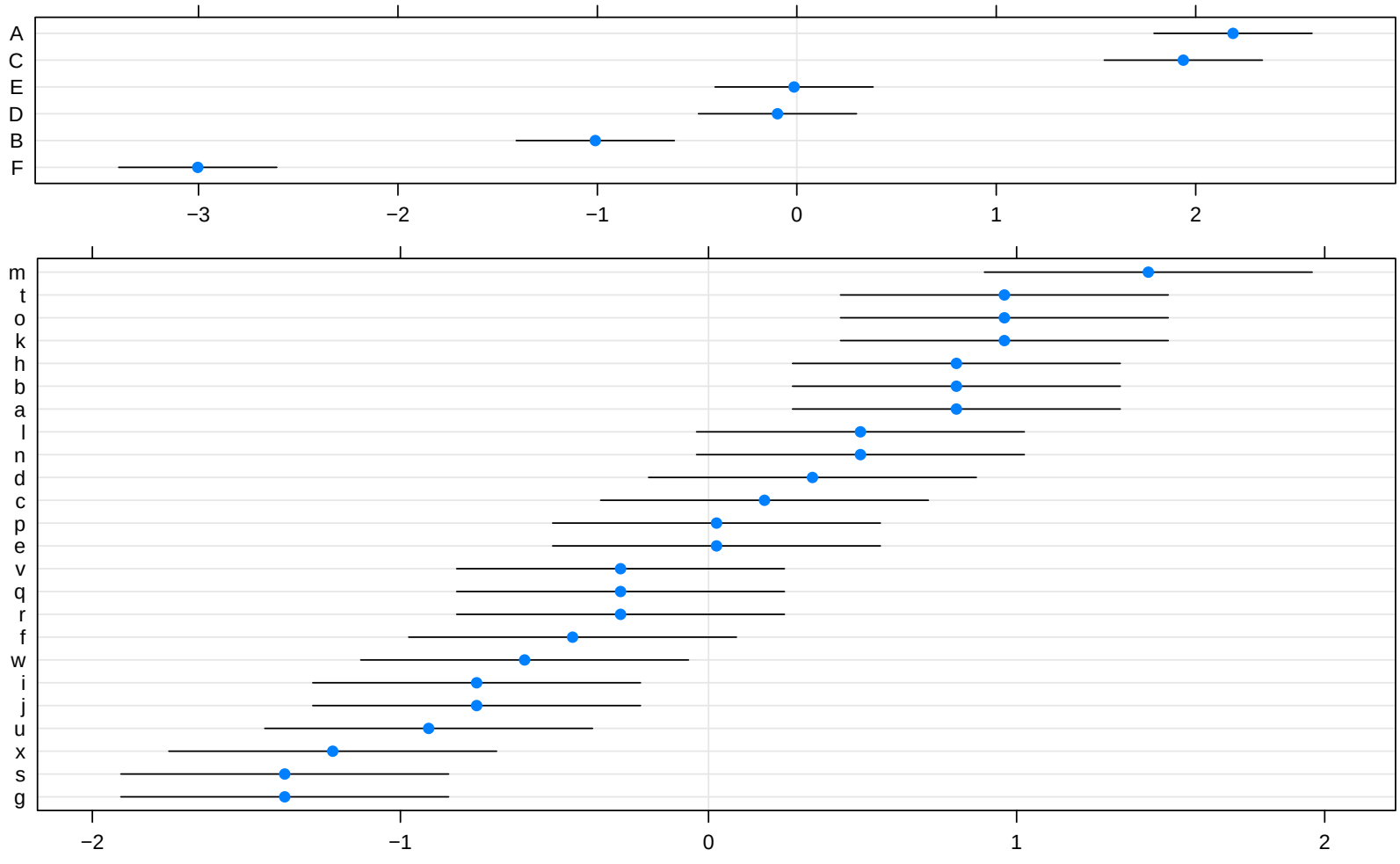
```
> (fm2 <- lmer(diameter ~ 1 + (1 | plate) + (1 | sample),
+     Penicillin))
Linear mixed model fit by REML
Formula: diameter ~ 1 + (1 | plate) + (1 | sample)
   Data: Penicillin
   AIC   BIC logLik deviance REMLdev
 338.9 350.7 -165.4    332.3   330.9
Random effects:
 Groups   Name        Variance 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.

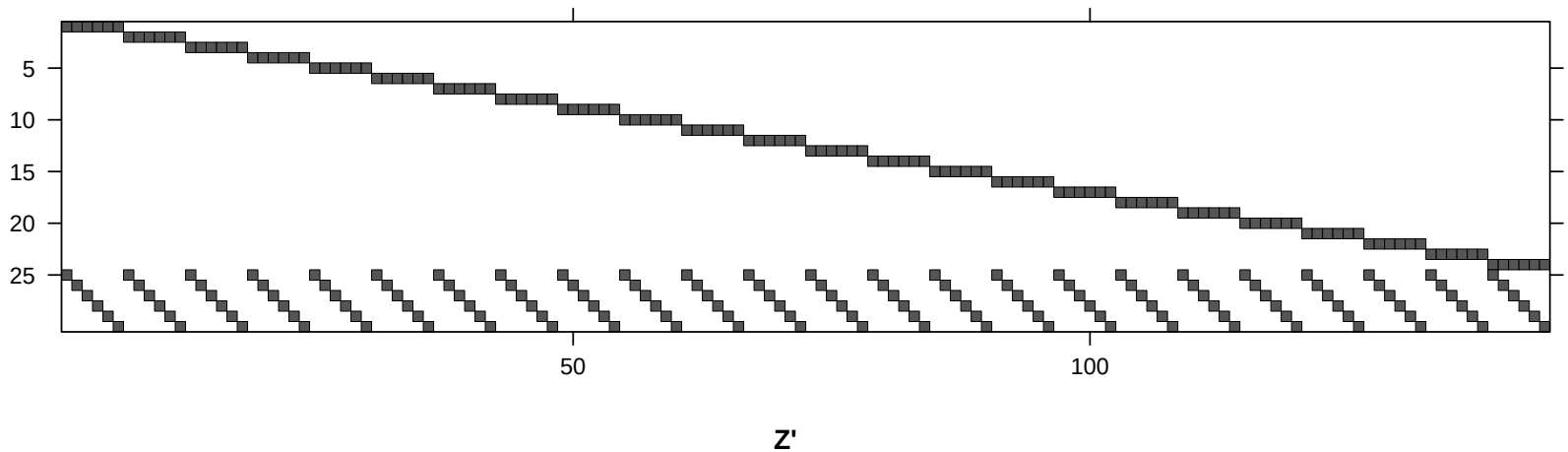
```
> fixef(fm2)

(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



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 $\boldsymbol{Z}'$ 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 the sparse, lower triangular, Cholesky factor, $\boldsymbol{L}(\boldsymbol{\theta})$, that satisfies

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})' = \boldsymbol{P}(\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})' + \boldsymbol{I}_q)\boldsymbol{P}'$$

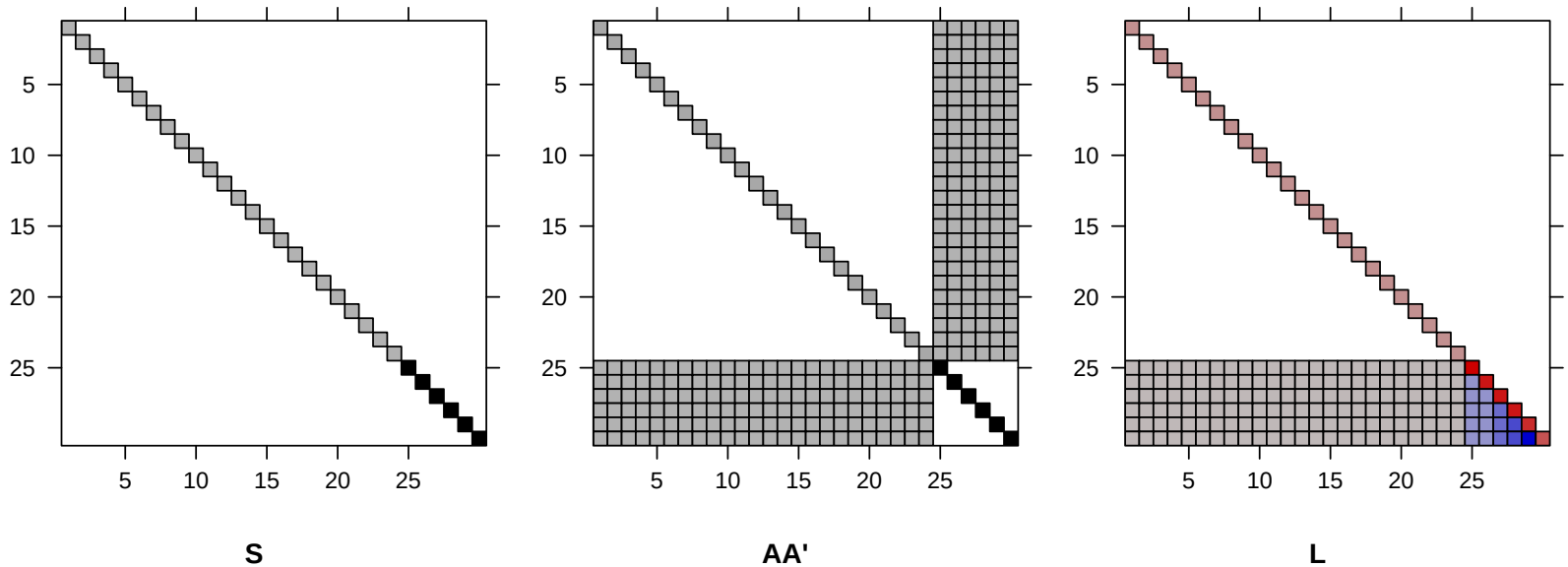
from $\boldsymbol{A}(\boldsymbol{\theta})' = \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})$. Crossing of grouping factors increases the number of nonzeros in $\boldsymbol{A}\boldsymbol{A}'$ and also causes some "fill-in" when creating $\boldsymbol{L}$ from $\boldsymbol{A}$.

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 (partial crossing) 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.

Images of some of the $q \times q$ matrices for fm2

- Because both random-effects terms are scalar terms, $\boldsymbol{T}$ is a block-diagonal matrix of two blocks, both of which are identity matrices. Hence $\boldsymbol{T} = \boldsymbol{I}_q$.
- For this model it is also the case that $\boldsymbol{P} = \boldsymbol{I}_q$.
- $\boldsymbol{S}$ consists of two diagonal blocks, both of which are multiples of an identity matrix. The multiples are different.



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 $\boldsymbol{Z}$ 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 matrices $\boldsymbol{AA}'$ and $\boldsymbol{L}$ show the crossing of the factors. $\boldsymbol{L}$ has some fill-in relative to $\boldsymbol{AA}'$.
- 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.

The Pastes data (see also the ?Pastes description)

```
> str(Pastes)

'data.frame': 60 obs. of  4 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 2 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 4 4

> 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

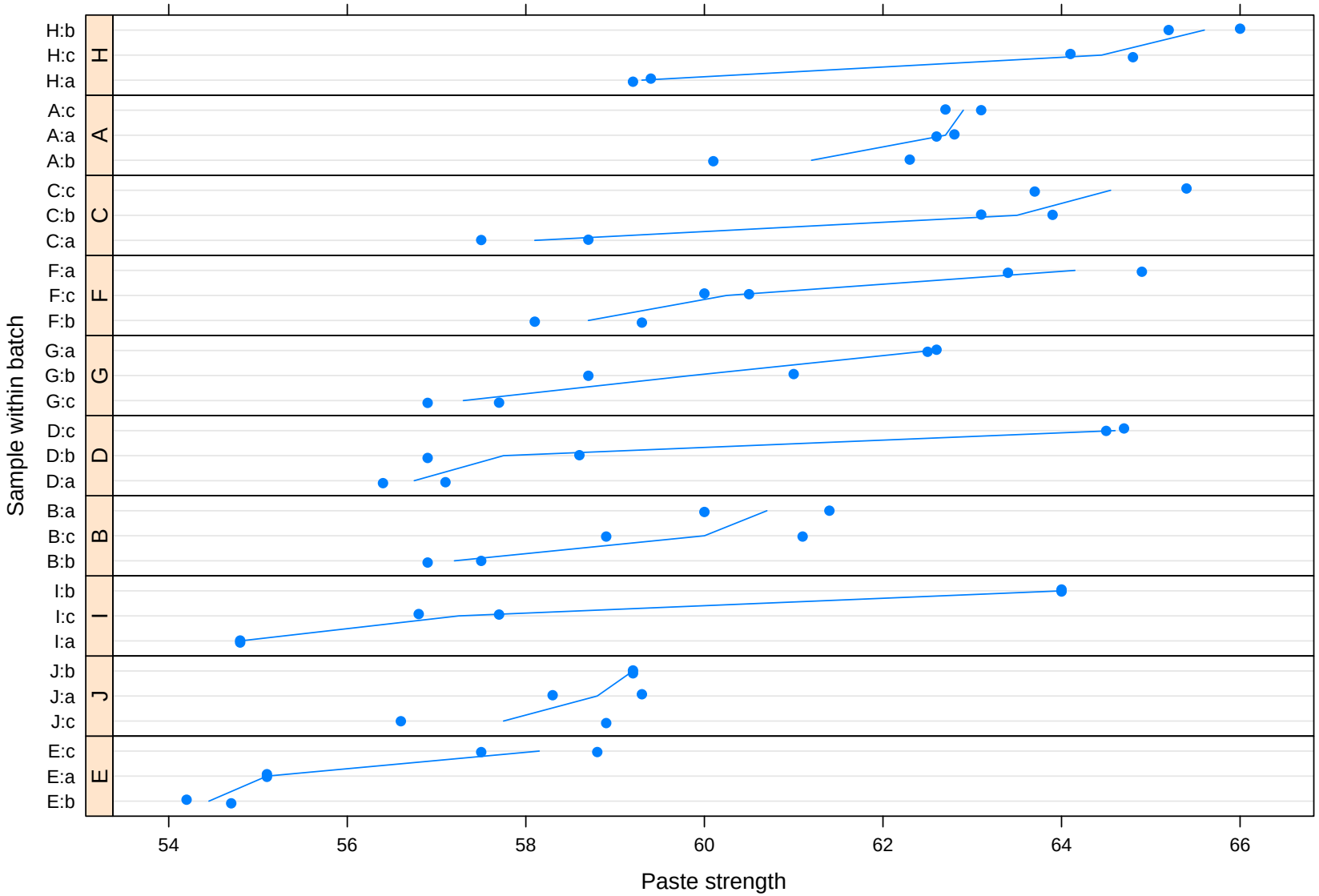
- The `sample` factor is nested within the `batch` factor. Each sample is from one of three casks selected from a particular batch.
- Note that there are 30, not 3, distinct samples.
- We can label the casks as 'a', 'b' and 'c' but then the `cask` factor by itself is meaningless (because cask 'a' in 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 <- (batch:cask)[drop = TRUE])
```

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.

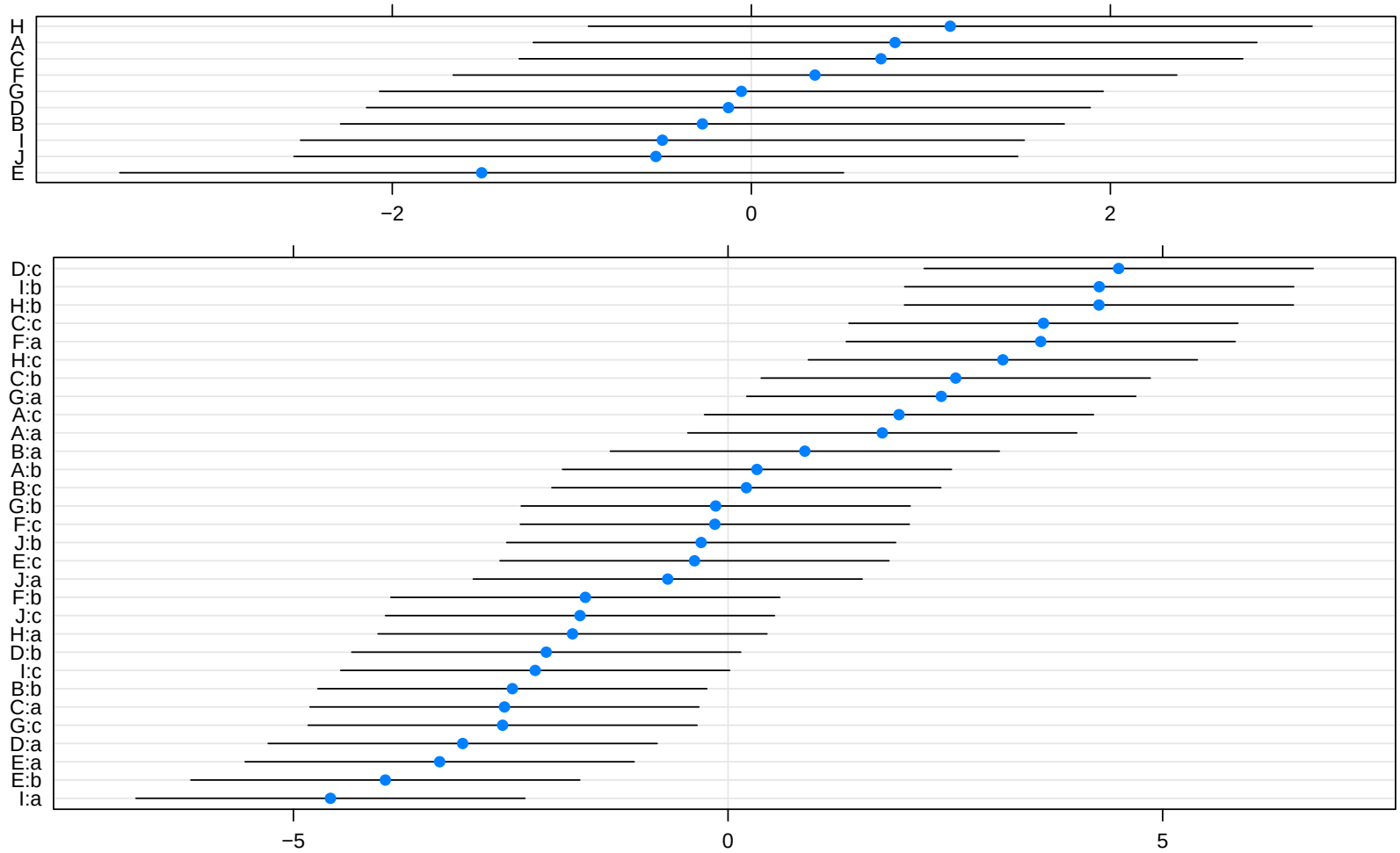
Pastes data plot



A model with nested random effects

```
> (fm3 <- lmer(strength ~ 1 + (1 | batch) + (1 | sample),
+     Pastes))
Linear mixed model fit by REML
Formula: strength ~ 1 + (1 | batch) + (1 | sample)
   Data: Pastes
 AIC   BIC logLik deviance REMLdev
 255 263.4 -123.5    248.0     247
Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 8.43378  2.90410
 batch    (Intercept) 1.65691  1.28721
 Residual             0.67801  0.82341
Number of obs: 60, groups: sample, 30; batch, 10
Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.0533     0.6768   88.73
```

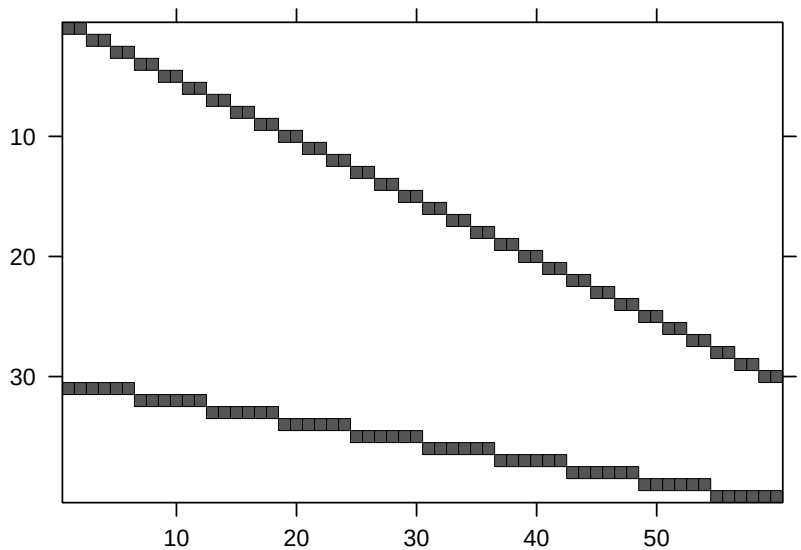
Random effects from model fm3



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

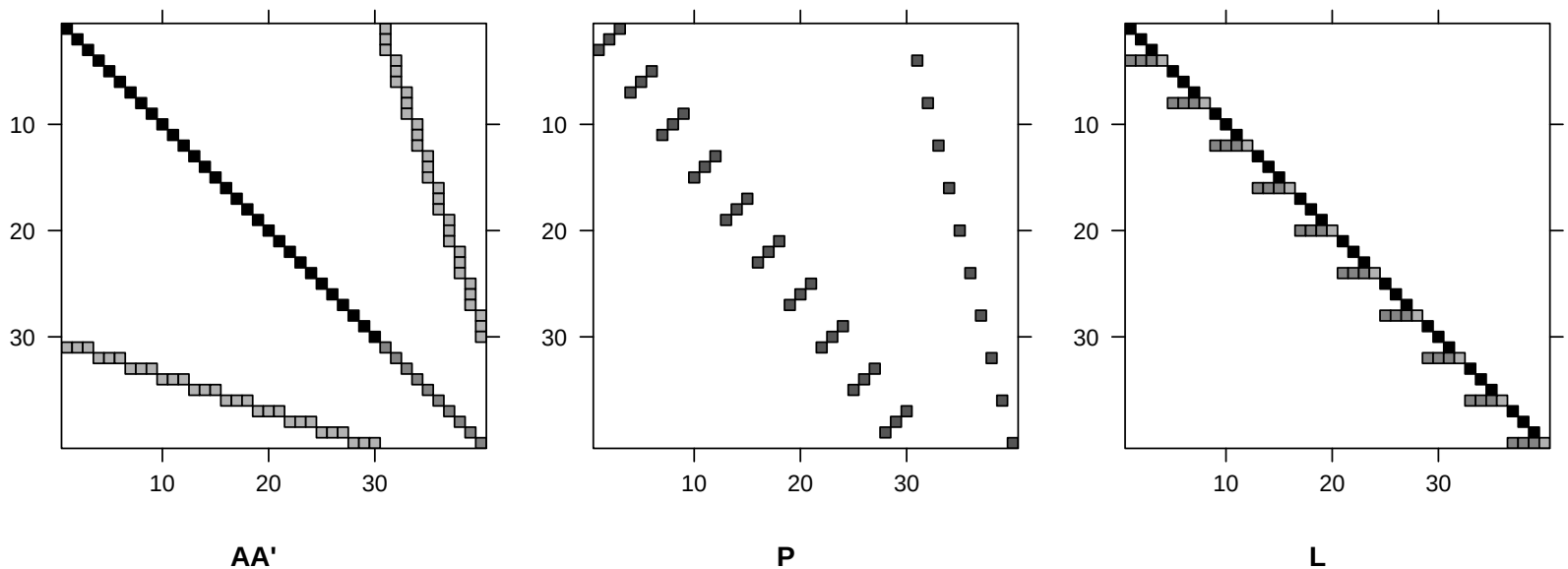
Dimensions and relationships in fm3

- 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 scalar terms, $\boldsymbol{T} = \boldsymbol{I}_{40}$ and $\boldsymbol{S}$ is block-diagonal in two diagonal blocks of sizes 30 and 10, respectively. $\boldsymbol{Z}$ is generated from two sets of indicators.



Images of some of the $q \times q$ matrices for fm3

- The permutation $\boldsymbol{P}$ has two purposes: reduce fill-in and "post-order" the columns to keep nonzeros near the diagonal.
- In a model with strictly nested grouping factors there will be no fill-in. The permutation $\boldsymbol{P}$ is chosen for post-ordering only.



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  Chisq Chi Df Pr(>Chisq)
fm4M  3  254.40  260.69 -124.20
fm3M  4  255.99  264.37 -124.00 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 region 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 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
Formula: strength ~ 1 + (1 | sample)
   Data: Pastes
   AIC   BIC logLik deviance REMLdev
 253.6 259.9 -123.8    248.4   247.6
Random effects:
 Groups   Name        Variance Std.Dev.
 sample   (Intercept) 9.97622  3.15852
 Residual             0.67803  0.82342
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 $q_1 = 30$ samples nested within $q_2 = 10$ 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`.

Recap of simple, scalar random-effects terms

- For the `lmer` function (and also for `glmer` and `nlmer`) 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 $\boldsymbol{I}_{n_i}$ to $\boldsymbol{T}$. If all random effects terms are scalar terms then $\boldsymbol{T} = \boldsymbol{I}$.
- Such a term contributes a diagonal block $c_i\boldsymbol{I}_{n_i}$ to $\boldsymbol{S}$. The multipliers c_i can be different for different terms. The term contributes exactly one element (which is c_i) to $\boldsymbol{\theta}$.

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.

A large observational data set

- A major 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
 $ dept    : Factor w/ 106 levels "AERO","AFAM",..: 43 43 43 43 43 43 4
 $ id      : Factor w/ 54711 levels "900000001","900000002",..: 12152 1
 $ 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
 $ 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.

Outline

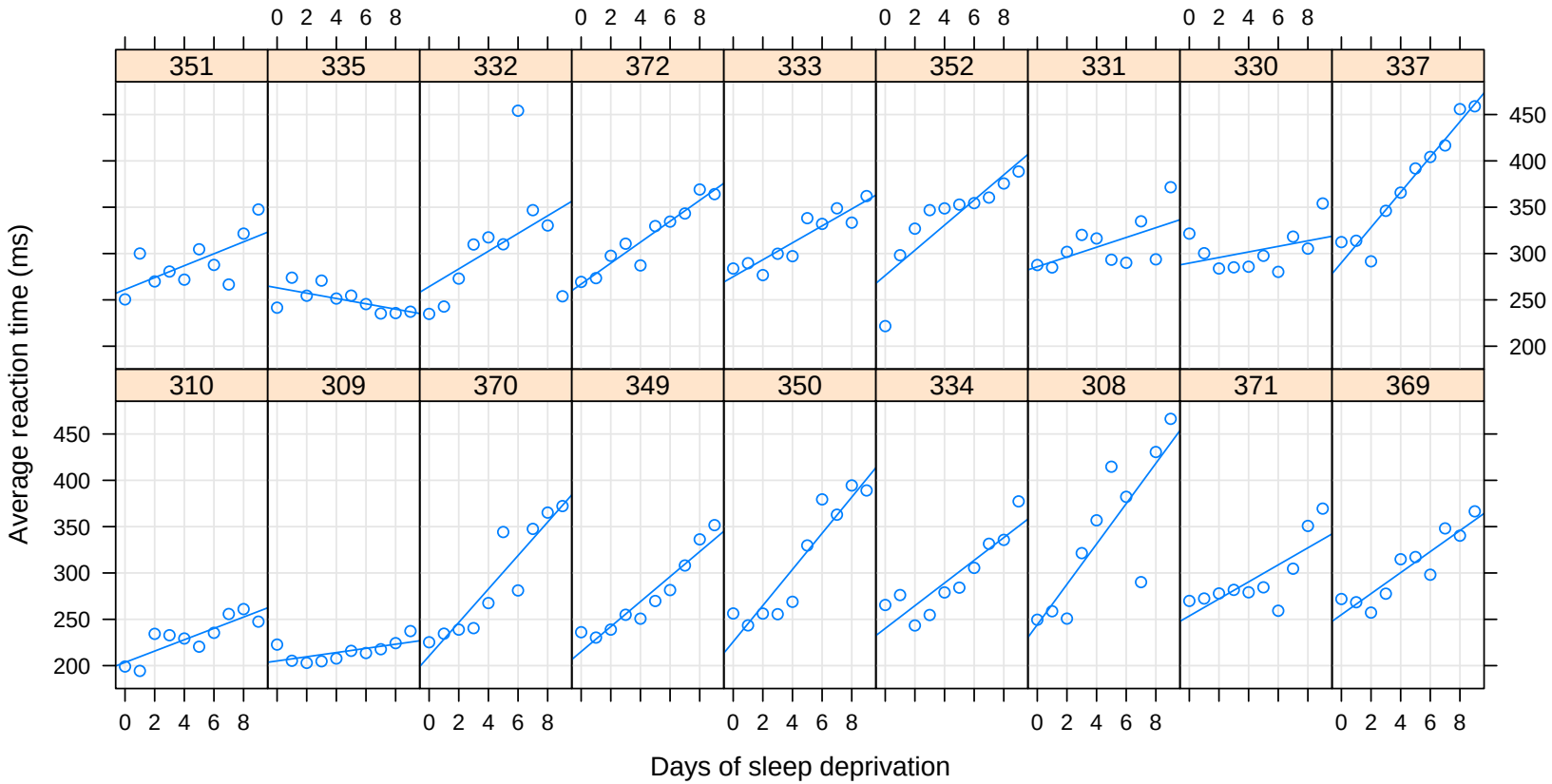
Simple longitudinal data

- *Repeated measures* data consist of measurements of a response (and, perhaps, some covariates) on several *experimental* (or observational) *units*.
- Frequently the experimental (observational) unit is `Subject` and we will refer to these units as "subjects". However, the methods described here are not restricted to data on human subjects.
- *Longitudinal* data are repeated measures data in which the observations are taken over time.
- We wish to characterize the response over time within subjects and the variation in the time trends between subjects.
- Frequently we are not as interested in comparing the particular subjects in the study as much as we are interested in modeling the variability in the population from which the subjects were chosen.

Sleep deprivation data

- This laboratory experiment measured the effect of sleep deprivation on cognitive performance.
- There were 18 subjects, chosen from the population of interest (long-distance truck drivers), in the 10 day trial. These subjects were restricted to 3 hours sleep per night during the trial.
- On each day of the trial each subject's reaction time was measured. The reaction time shown here is the average of several measurements.
- These data are *balanced* in that each subject is measured the same number of times and on the same occasions.

Reaction time versus days by subject



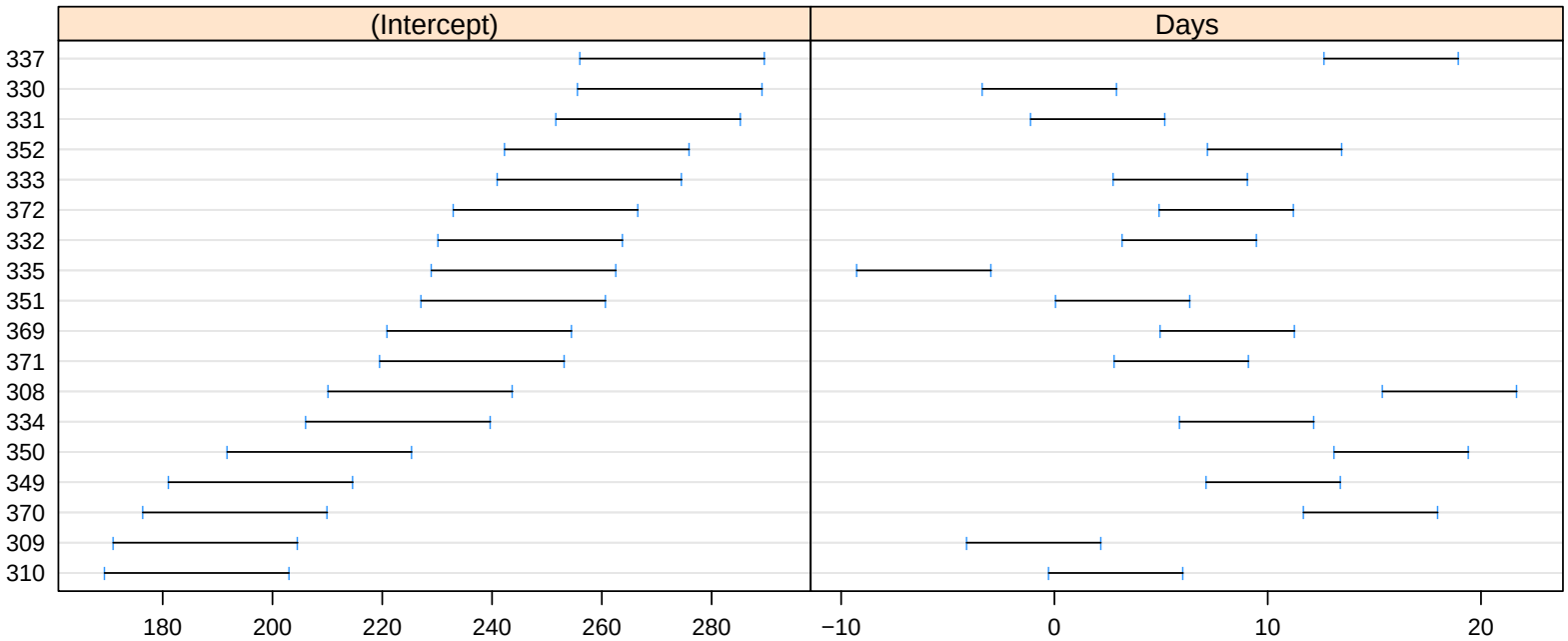
Comments on the sleep data plot

- The plot is a "trellis" or "lattice" plot where the data for each subject are presented in a separate panel. The axes are consistent across panels so we may compare patterns across subjects.
- A reference line fit by simple linear regression to the panel's data has been added to each panel.
- The aspect ratio of the panels has been adjusted so that a typical reference line lies about 45° on the page. We have the greatest sensitivity in checking for differences in slopes when the lines are near $\pm 45^\circ$ on the page.
- The panels have been ordered not by subject number (which is essentially a random order) but according to increasing intercept for the simple linear regression. If the slopes and the intercepts are highly correlated we should see a pattern across the panels in the slopes.

Assessing the linear fits

- In most cases a simple linear regression provides an adequate fit to the within-subject data.
- Patterns for some subjects (e.g. 350, 352 and 371) deviate from linearity but the deviations are neither widespread nor consistent in form.
- There is considerable variation in the intercept (estimated reaction time without sleep deprivation) across subjects – 200 ms. up to 300 ms. – and in the slope (increase in reaction time per day of sleep deprivation) – 0 ms./day up to 20 ms./day.
- We can examine this variation further by plotting confidence intervals for these intercepts and slopes. Because we use a pooled variance estimate and have balanced data, the intervals have identical widths.
- We again order the subjects by increasing intercept so we can check for relationships between slopes and intercepts.

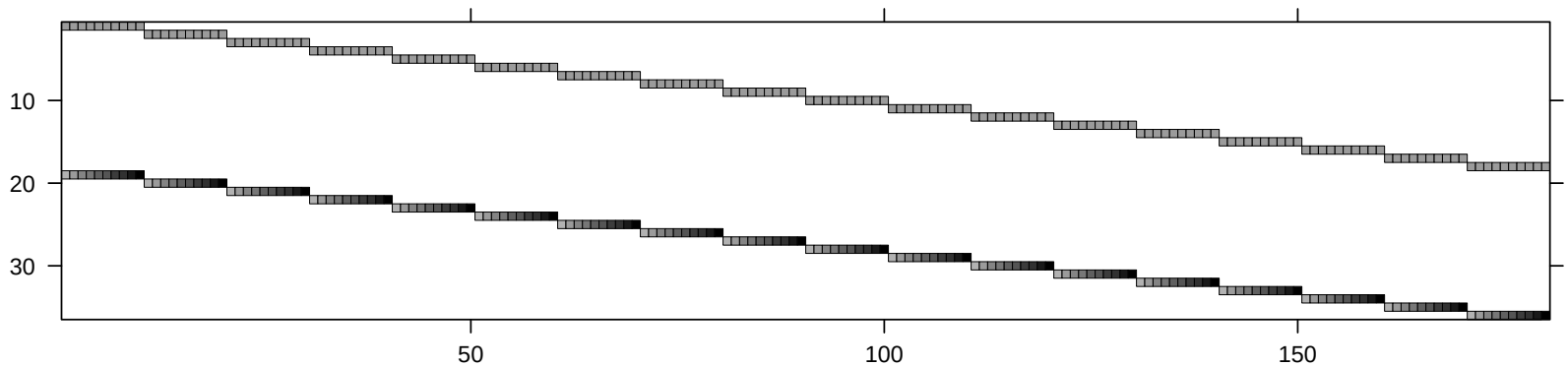
95% conf int on within-subject intercept and slope



These intervals reinforce our earlier impressions of considerable variability between subjects in both intercept and slope but little evidence of a relationship between intercept and slope.

A preliminary mixed-effects model

- We begin with a linear mixed model in which the fixed effects $[\beta_1, \beta_2]'$ are the representative intercept and slope for the population and the random effects $\boldsymbol{b}_i = [b_{i1}, b_{i2}]', i = 1, \ldots, 18$ are the deviations in intercept and slope associated with subject i.
- The random effects vector, $\boldsymbol{b}$, consists of the 18 intercept effects followed by the 18 slope effects.



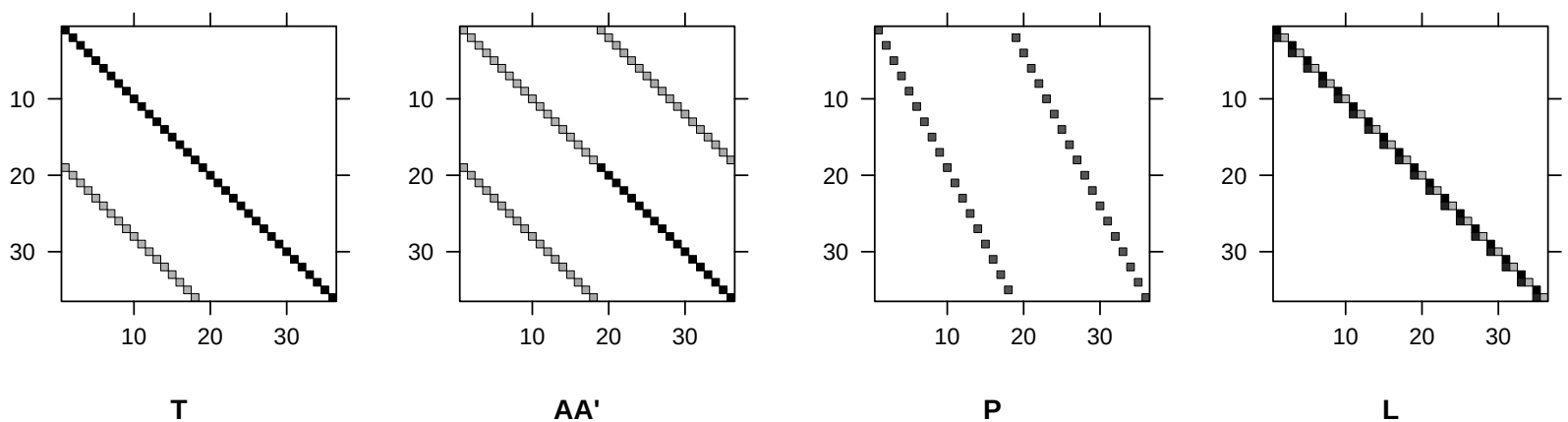
Fitting the model

```
> (fm1 <- lmer(Reaction ~ Days + (Days | Subject),
+     sleepstudy))
```

```
Linear mixed model fit by REML
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance REMLdev
 1756 1775 -871.8     1752    1744
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.095  24.7405
          Days         35.071   5.9221  0.065
 Residual             654.944  25.5919
Number of obs: 180, groups: Subject, 18
Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

Terms and matrices

- The term `Days` in the formula generates a model matrix $\boldsymbol{X}$ with two columns, the intercept column and the numeric `Days` column. (The intercept is included unless suppressed.)
- The term `(Days|Subject)` generates a vector-valued random effect (intercept and slope) for each of the 18 levels of the `Subject` factor.



A model with uncorrelated random effects

- The data plots gave little indication of a systematic relationship between a subject's random effect for slope and his/her random effect for the intercept. Also, the estimated correlation is quite small.
- We should consider a model with uncorrelated random effects. To express this we use two random-effects terms with the same grouping factor and different left-hand sides. In the formula for an `lmer` model, distinct random effects terms are modeled as being independent. Thus we specify the model with two distinct random effects terms, each of which has `Subject` as the grouping factor. The model matrix for one term is intercept only (`1`) and for the other term is the column for `Days` only, which can be written `0+Days`. (The expression `Days` generates a column for `Days` and an intercept. To suppress the intercept we add `0+` to the expression; `-1` also works.)

A mixed-effects model with independent random effects

```
Linear mixed model fit by REML
Formula: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance REMLdev
 1754 1770 -871.8     1752    1744
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 627.577  25.0515
 Subject  Days         35.852   5.9876
 Residual             653.594  25.5655
Number of obs: 180, groups: Subject, 18
Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.885   36.51
Days          10.467      1.559    6.71

Correlation of Fixed Effects:
     (Intr)
Days -0.184
```

Comparing the models

- Model `fm1` contains model `fm2` in the sense that if the parameter values for model `fm1` were constrained so as to force the correlation, and hence the covariance, to be zero, and the model were re-fit, we would get model `fm2`.
- The value `0`, to which the correlation is constrained, is not on the boundary of the allowable parameter values.
- In these circumstances a likelihood ratio test and a reference distribution of a χ^2 on 1 degree of freedom is suitable.

```
> anova(fm2, fm1)
```

```
Data: sleepstudy
Models:
fm2: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
fm1: Reaction ~ Days + (Days | Subject)
    Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm2  5 1762.05 1778.01 -876.02
fm1  6 1763.99 1783.14 -875.99 0.0609      1      0.805
```

Conclusions from the likelihood ratio test

- Because the large p-value indicates that we would not reject `fm2` in favor of `fm1`, we prefer the more parsimonious `fm2`.
- This conclusion is consistent with the AIC (Akaike's Information Criterion) and the BIC (Bayesian Information Criterion) values for which "smaller is better".
- We can also use a Bayesian approach, where we regard the parameters as themselves being random variables, is assessing the values of such parameters. A currently popular Bayesian method is to use sequential sampling from the conditional distribution of subsets of the parameters, given the data and the values of the other parameters. The general technique is called *Markov chain Monte Carlo* sampling.
- The `lme4` package has a function called `mcmcsamp` to evaluate such samples from a fitted model. At present, however, there seem to be a few "infelicities", as Bill Venables calls them, in this function.

Likelihood ratio tests on variance components

- As for the case of a covariance, we can fit the model with and without the variance component and compare the fit quality.
- As mentioned previously, the likelihood ratio is a reasonable test statistic for the comparison but the "asymptotic" reference distribution of a χ^2 does not apply because the parameter value being tested is on the boundary.
- The p-value computed using the χ^2 reference distribution should be conservative (i.e. greater than the p-value that would be obtained through simulation).

```
> fm3 <- lmer(Reaction ~ Days + (1 | Subject), sleepstudy)
> anova(fm3, fm2)
```

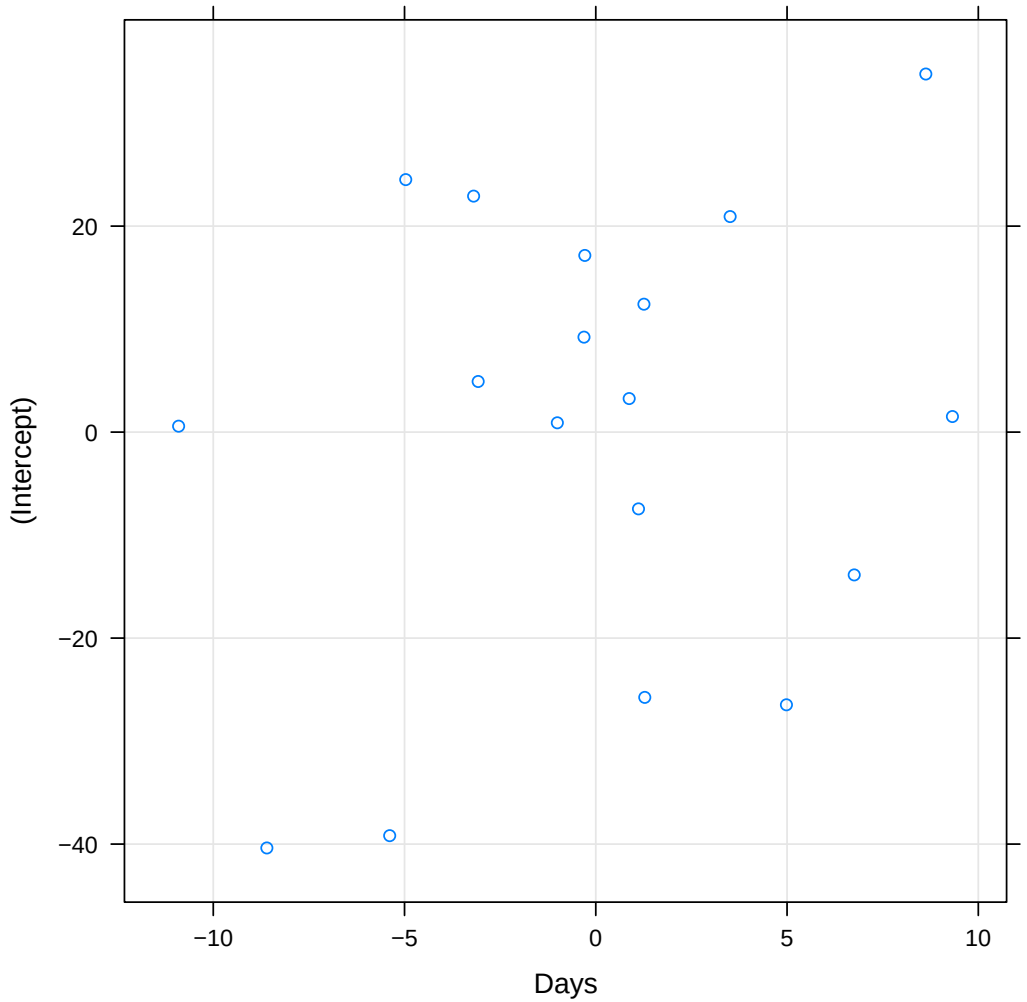
```
Data: sleepstudy
Models:
fm3: Reaction ~ Days + (1 | Subject)
fm2: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
    Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm3  4 1802.10 1814.87 -897.05
fm2  5 1762.05 1778.01 -876.02 42.053      1  8.885e-11
```

Conditional modes of the random effects

```
> (rr2 <- ranef(fm2))
```

```
$Subject
    (Intercept)        Days
308   1.5138200   9.3232135
309 -40.3749105  -8.5989183
310 -39.1816682  -5.3876346
330  24.5182907  -4.9684965
331  22.9140346  -3.1938382
332   9.2219311  -0.3084836
333  17.1560765  -0.2871973
334  -7.4515945   1.1159563
335   0.5774094 -10.9056435
337  34.7689482   8.6273639
349 -25.7541541   1.2806475
350 -13.8642120   6.7561993
351   4.9156063  -3.0750415
352  20.9294539   3.5121076
369   3.2587507   0.8730251
370 -26.4752098   4.9836365
371   0.9055257  -1.0052631
372  12.4219020   1.2583667
```

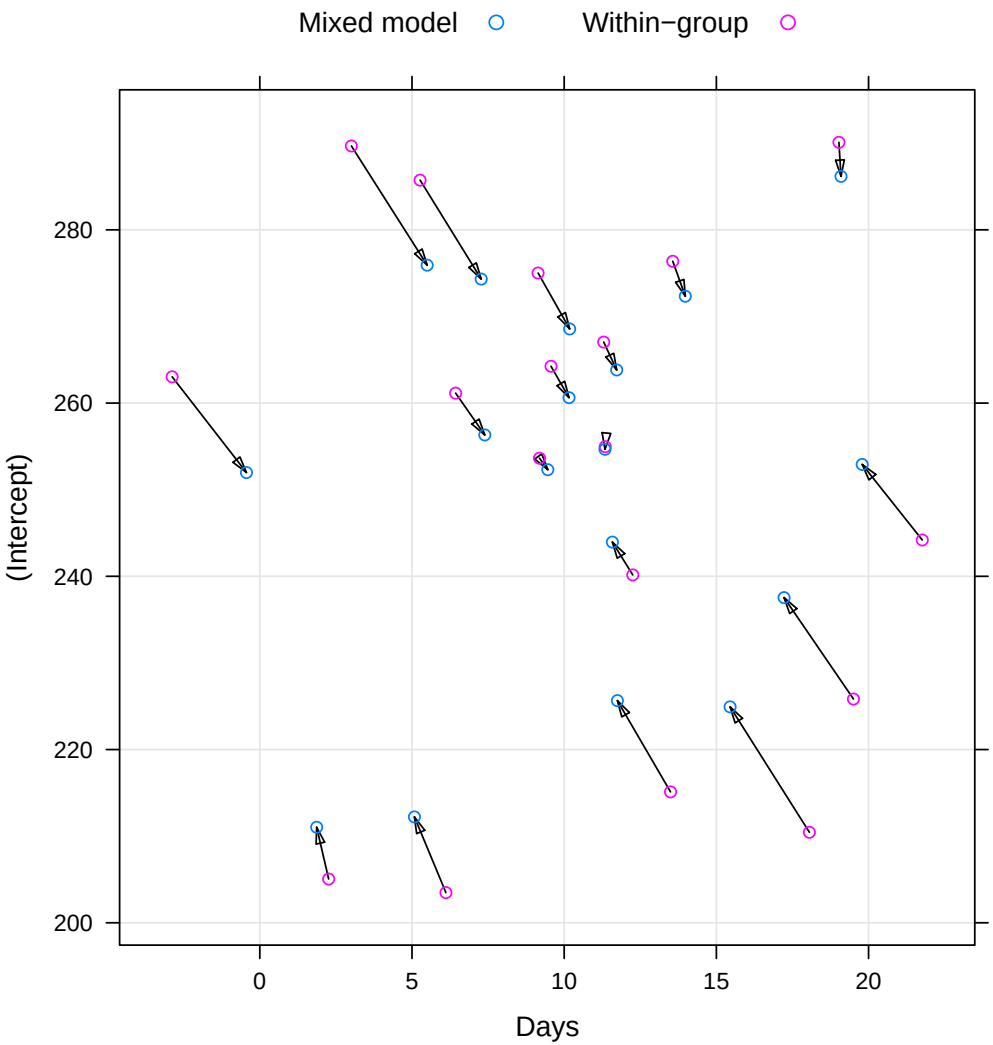
Scatterplot of the conditional modes



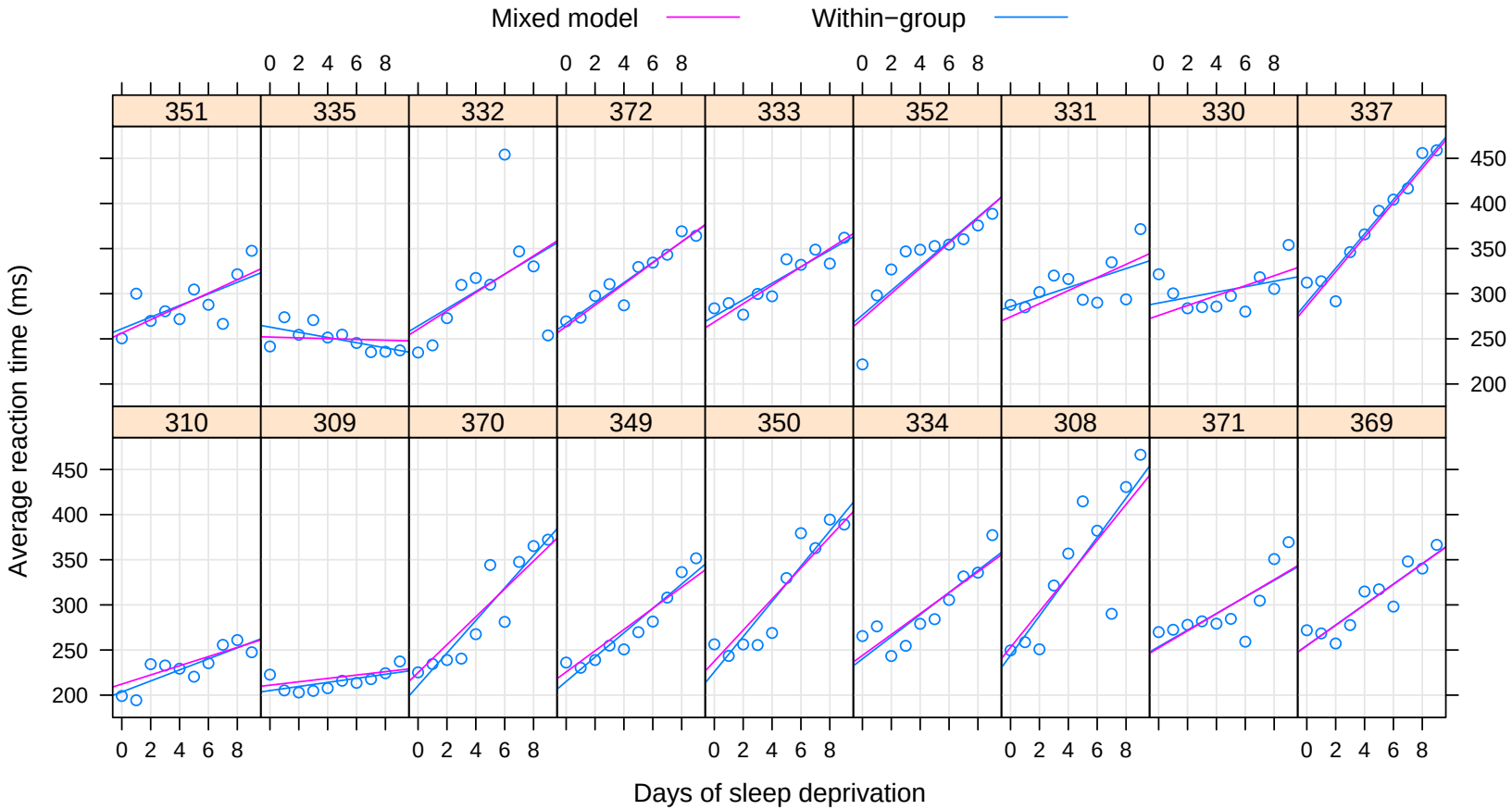
Comparing within-subject coefficients

- For this model we can combine the conditional modes of the random effects and the estimates of the fixed effects to get conditional modes of the within-subject coefficients.
- These conditional modess will be "shrunken" towards the fixed-effects estimates relative to the estimated coefficients from each subject's data. John Tukey called this "borrowing strength" between subjects.
- Plotting the shrinkage of the within-subject coefficients shows that some of the coefficients are considerably shrunken toward the fixed-effects estimates.
- However, comparing the within-group and mixed model fitted lines shows that large changes in coefficients occur in the noisy data. Precisely estimated within-group coefficients are not changed substantially.

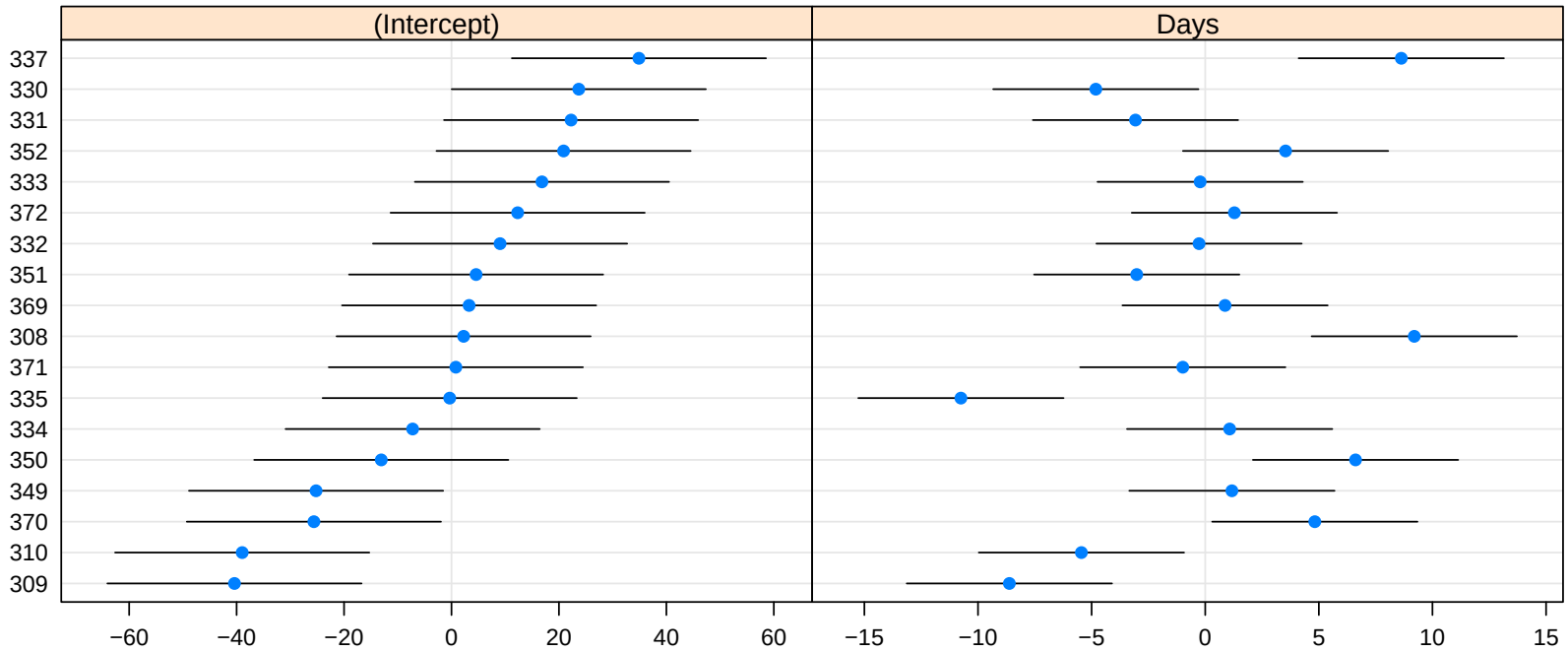
Estimated within-group coefficients and BLUPs



Observed and fitted



Plot of prediction intervals for the random effects



Each set of prediction intervals have constant width because of the balance in the experiment.

Conclusions from the example

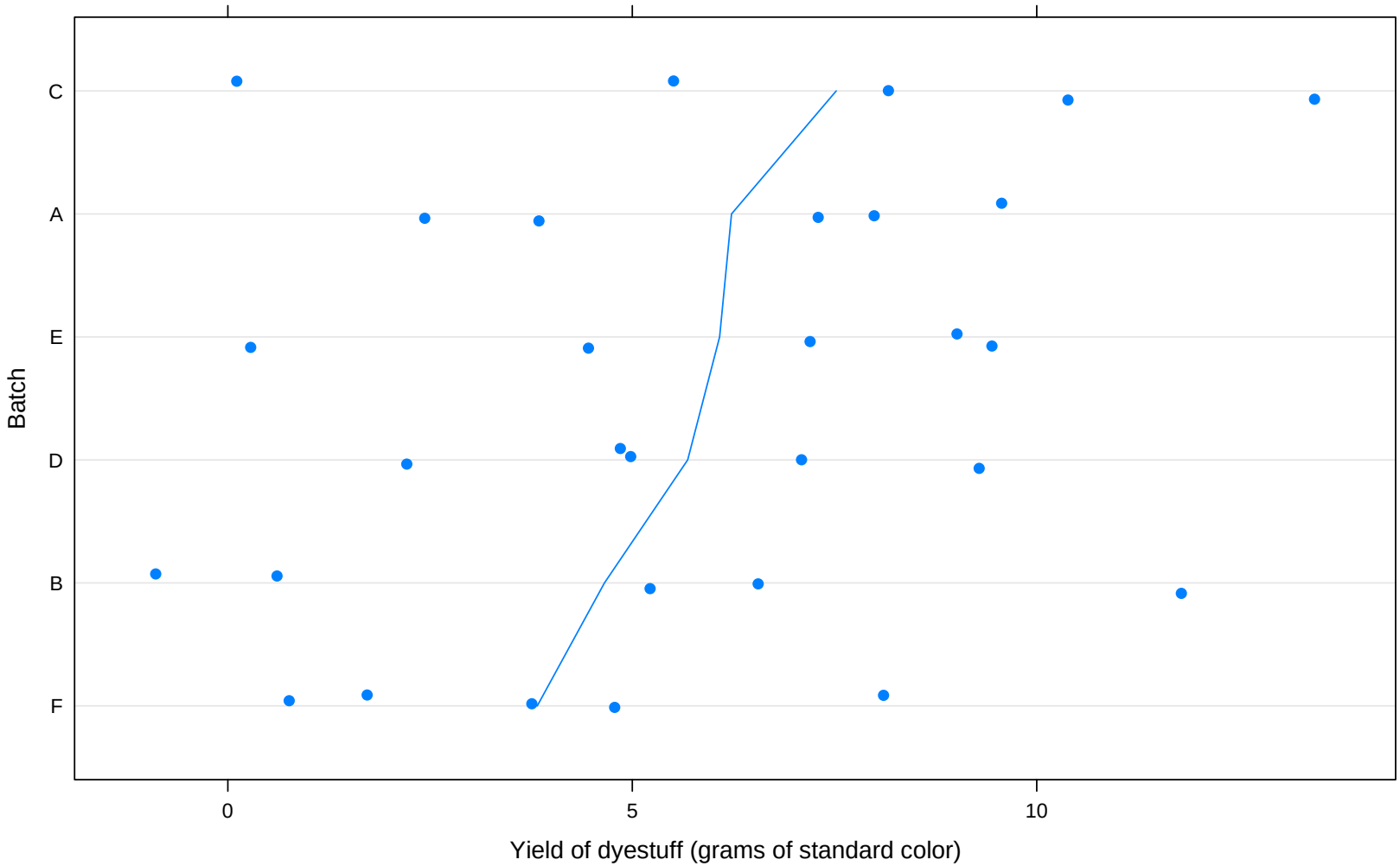
- Carefully plotting the data is enormously helpful in formulating the model.
- It is relatively easy to fit and evaluate models to data like these, from a balanced designed experiment.
- We consider two models with random effects for the slope and the intercept of the response w.r.t. time by subject. The models differ in whether the (marginal) correlation of the vector of random effects per subject is allowed to be nonzero.
- The "estimates" (actually, the conditional modes) of the random effects can be considered as penalized estimates of these parameters in that they are shrunk towards the origin.
- Most of the prediction intervals for the random effects overlap zero.

Outline

Random-effects variances estimated as zero?

- The estimated variance of the random effects for a given term must be non-negative. It can be shown that they must be also be finite. However, it is possible for the estimate to be exactly zero, indicating an overspecified model. In these cases we should respecify the model to have fewer random effects.
- Because the estimates of these variances are printed explicitly, it is obvious to us when such a situation occurs (although you need to watch for estimated variances that are very close to but not exactly zero).
- Box and Tiao in "Bayesian Inference in Statistical Analysis" (Addison-Wesley, 1973) provide simulated data similar to the `Dyestuff` data but with much lower batch-to-batch variability.

Dyestuff2 data plot



Model fit for Dyestuff2

```
> lmer(Yield ~ 1 + (1 | Batch), Dyestuff2)

Linear mixed model fit by REML
Formula: Yield ~ 1 + (1 | Batch)
   Data: Dyestuff2
   AIC   BIC logLik deviance REMLdev
 167.8 172.0 -80.91    162.9   161.8
Random effects:
 Groups   Name        Variance   Std.Dev.
 Batch    (Intercept) 1.9844e-09 4.4547e-05
 Residual             1.3806e+01 3.7157e+00
Number of obs: 30, groups: Batch, 6
Fixed effects:
            Estimate Std. Error t value
(Intercept)   5.6656     0.6784   8.352
```

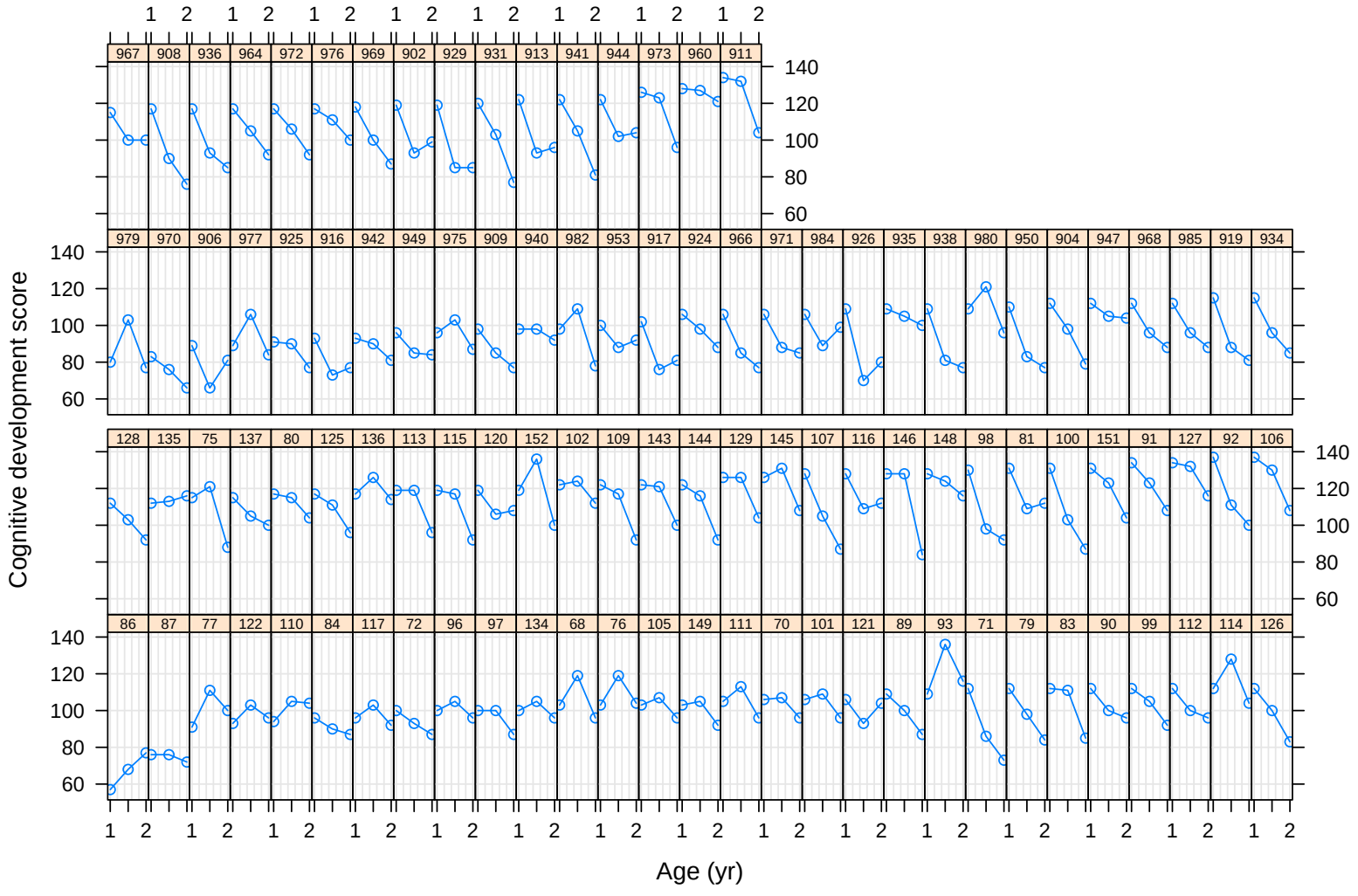
Singular variance-covariance matrices

- With vector-valued random effects we can obtained a singular or near-singular estimate of a variance-covariance matrix that has safely non-zero variances for all components. It is harder to spot such singularities. If there are only two random effects per level of the grouping factor we can recognize this situation as generating highly correlated (sometimes perfectly correlated) random effects. For dimensions greater than 2, even this signal may fail to be present.
- I will likely add a diagnostic method with a name like `rcond` (reciprocal of the condition number) to `lme4` to make this situation easier to spot.

The early childhood cognitive study data

- In an experiment reported in Burchinal et al. (*Cognitive Development, 1997*) young children were randomly assigned to a treatment group and a control group at 6 months of age and their cognitive development was measured at 1, 1.5 and 2 years of age, on an age-normed scale. (The treatment was exposure to an enriched environment.)
- Because the treatment began at 6 months of age, we will use `tos = age - 0.5` (time on study), as the time variable. Because the treatment groups are assumed to be homogeneous at the beginning of the study, we do not expect to see a `trt` main effect. The effect of the treatment, if any, will show up in the `tos:trt` interaction term.
- These data are the initial example in Singer and Willett, *Longitudinal Data Analysis* (Oxford, 2003). Models are fit using both SAS and MLWin but the singularity in the estimated variance-covariance matrix is never noticed.

Early childhood cognitive data plot

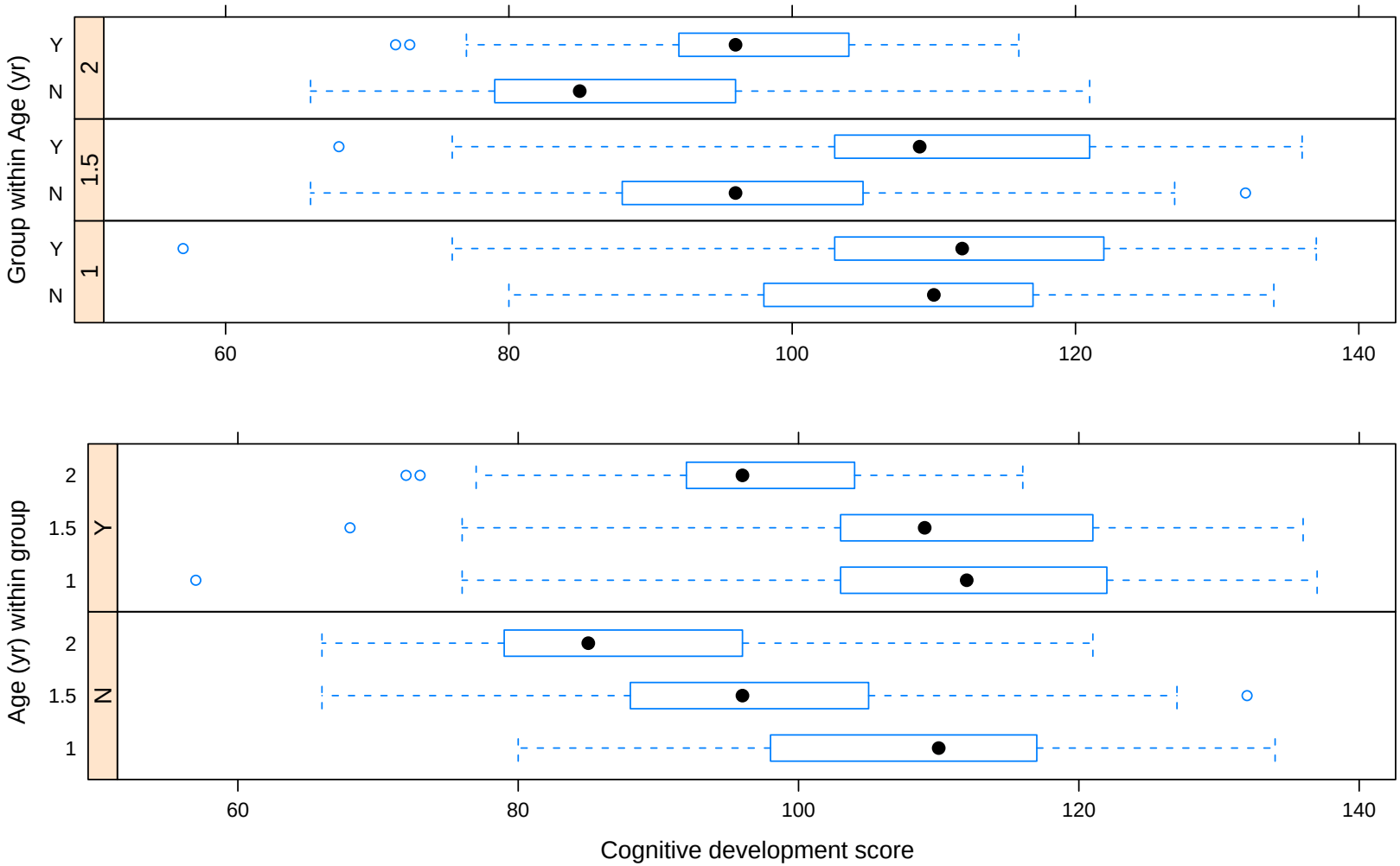


Some comments on the data plot

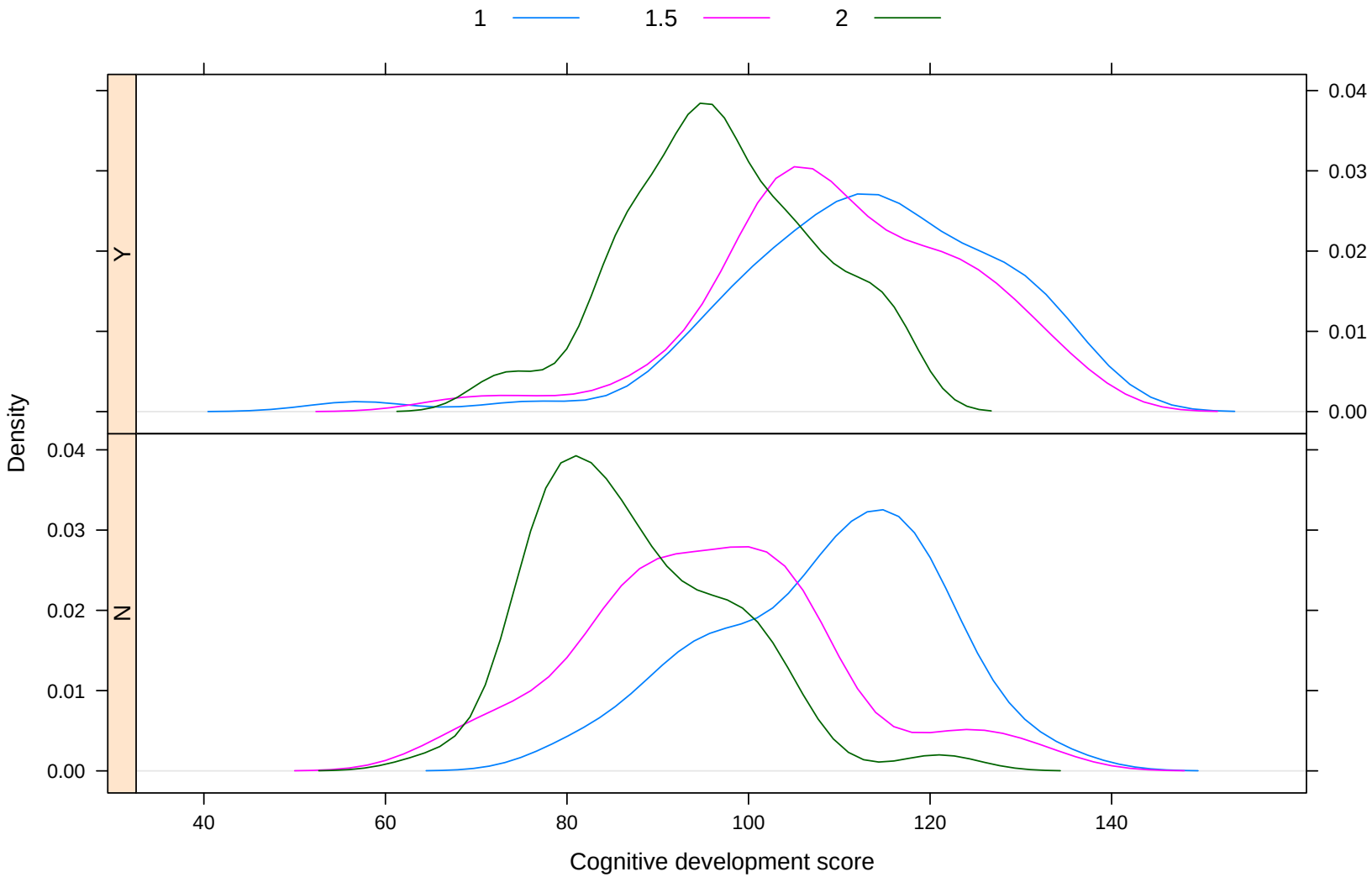
- Notice that, for both groups, the slopes are almost always negative. This is curious if the scale has been age-normed.
- Also, the initial measurement at age 1 yr seems high. If the scale is age-normed then we would expect the observations in (at least) the control group to have a mean near 100.

```
> with(subset(Early, age == 1 & trt == "N"), summary(cog))

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   80.0    98.0   110.0   108.5   117.0   134.0

> t.test(subset(Early, age == 1 & trt == "N")$cog,
+     mu = 100)$p.value

[1] 3.004468e-05
```

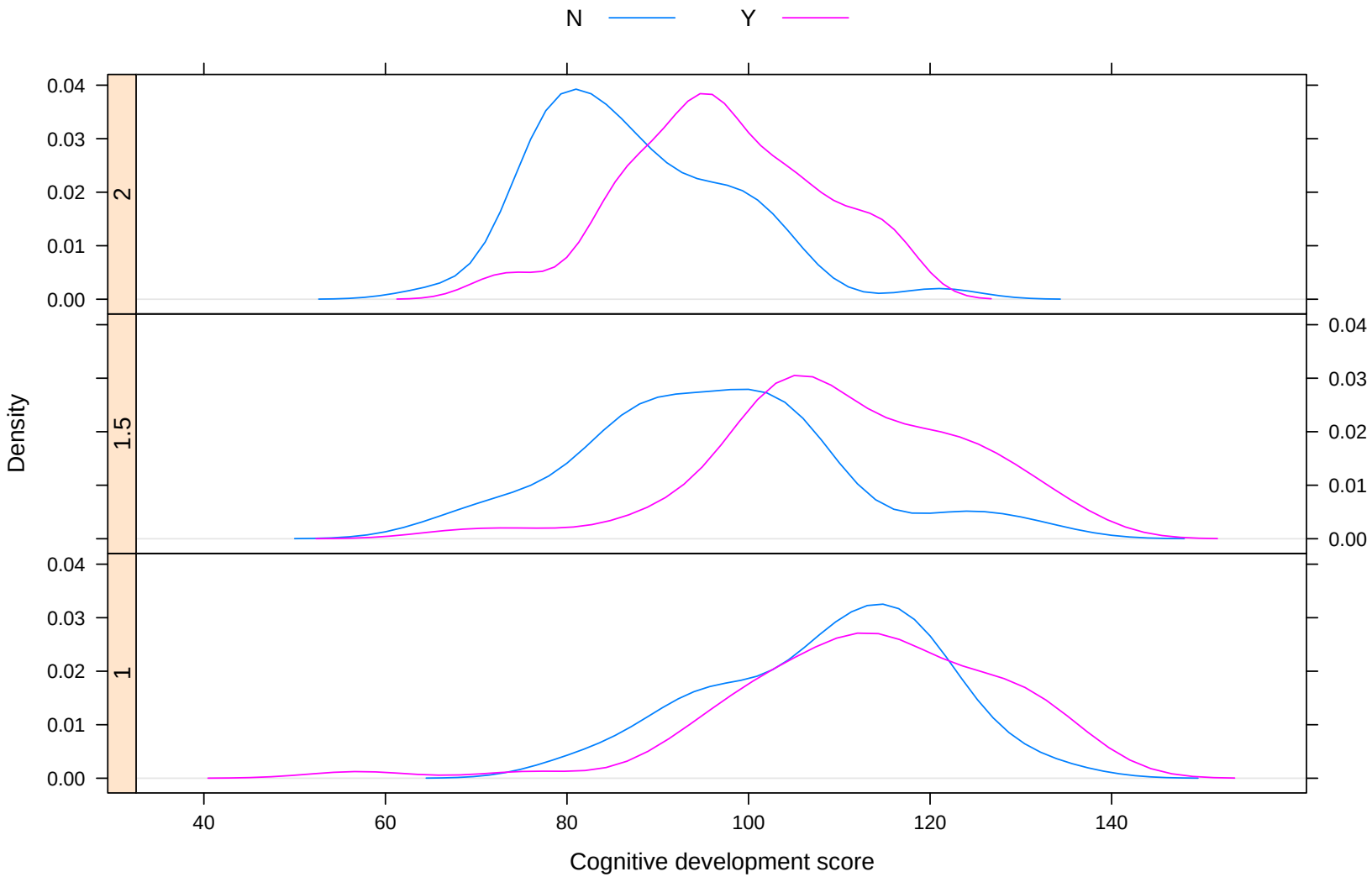
Comparative box-and-whisker plots



Density by age within group



Density by group within age



Initial model fit

```
> Early <- within(Early, tos <- age - 0.5)
> fm1 <- lmer(cog ~ tos * trt + (tos | id), Early,
+     verbose = TRUE)

  0:     2390.0424: 0.942809 0.872872  0.00000
  1:     2375.7837:  1.18238  0.00000 -0.409227
  2:     2364.4300:  1.30850 8.72884e-06 -0.0406037
  3:     2359.1951:  1.62268  0.00000 -0.269248
  4:     2358.8599:  1.49156 0.0430307 -0.272762
  5:     2358.8282:  1.47650 0.0652259 -0.233211
  6:     2358.7448:  1.47409 0.0193923 -0.246519
  7:     2358.7429:  1.48001 0.0127276 -0.248899
  8:     2358.7426:  1.47952 0.00355024 -0.249716
  9:     2358.7425:  1.48080  0.00000 -0.249545
 10:     2358.7425:  1.48070  0.00000 -0.249736
 11:     2358.7425:  1.48069  0.00000 -0.249728
```

The second parameter being exactly zero is a danger sign.

Summary of initial model

```
> print(fm1, corr = FALSE)

Linear mixed model fit by REML
Formula: cog ~ tos * trt + (tos | id)
   Data: Early
  AIC  BIC logLik deviance REMLdev
 2375 2405  -1179     2370    2359
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 165.518  12.8654
          tos          10.326   3.2133  -1.000
 Residual              75.495   8.6888
Number of obs: 309, groups: id, 103
Fixed effects:
            Estimate Std. Error t value
(Intercept)  118.407      2.755   42.97
tos          -21.133      1.893  -11.16
trtY           4.219      3.672    1.15
tos:trtY       5.271      2.523    2.09
```

Comments on the fitted model

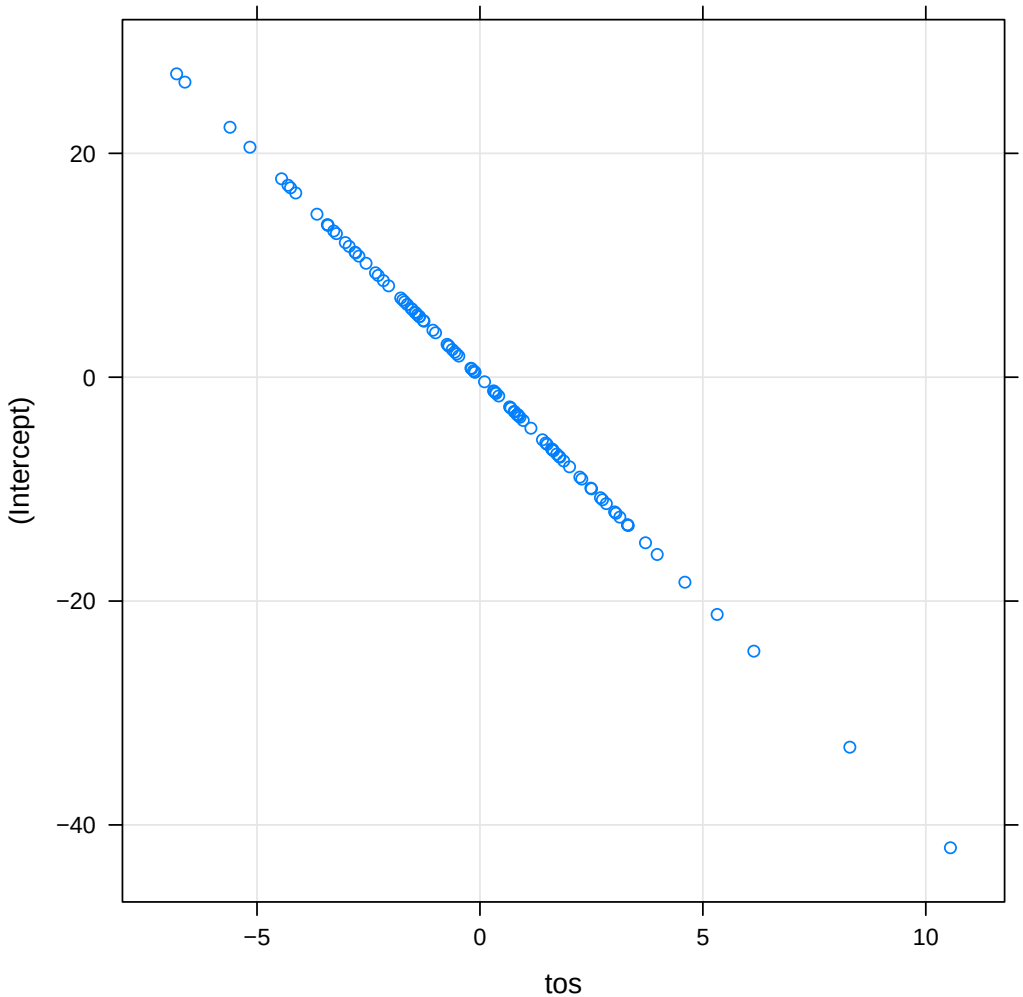
- As expected, the main effect for treatment group is not significant. Recall that we changed the time scale so zero corresponds to the beginning of the intervention.
- The only danger sign in the output about singularity of the variance-covariance of the random effects is the fact that the estimated correlation is exactly -1.
- Even after refitting without the main effect for treatment, this high negative correlation is present.
- If you display the covariance instead of the correlation, it is very difficult to see that something is amiss.

Re-fit without the main effect for treatment

```
> print(fm2 <- update(fm1, . ~ . - trt, verbose = FALSE),
+       corr = FALSE)
```

```
Linear mixed model fit by REML
Formula: cog ~ tos + (tos | id) + tos:trt
   Data: Early
  AIC  BIC logLik deviance REMLdev
 2379 2405  -1182     2371    2365
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 166.403  12.8997
          tos          10.484   3.2379  -1.000
 Residual              75.540   8.6914
Number of obs: 309, groups: id, 103
Fixed effects:
            Estimate Std. Error t value
(Intercept)  120.783      1.824   66.22
tos          -22.470      1.494  -15.04
tos:trtY       7.646      1.447    5.29
```

Conditional modes of the random effects



Handling singularity

- We can try to redefine the model with fewer random effects or with uncorrelated random effects, if such a model would make sense and if the model fits are adequate.
- In this case, the reduced models with independent random effects or with random effects for the intercept only are not adequate.
- There is little that can be done here to handle the singularity in the model fit, because the singularity is related to anomalies in the data.

```
> fm2M <- update(fm2, REML = FALSE)
> fm3M <- lmer(cog ~ tos + tos:trt + (1 | id) + (0 +
+     tos | id), Early, REML = FALSE)
> fm4M <- lmer(cog ~ tos + tos:trt + (1 | id), Early,
+     REML = FALSE)
```

Checking possible reduced models

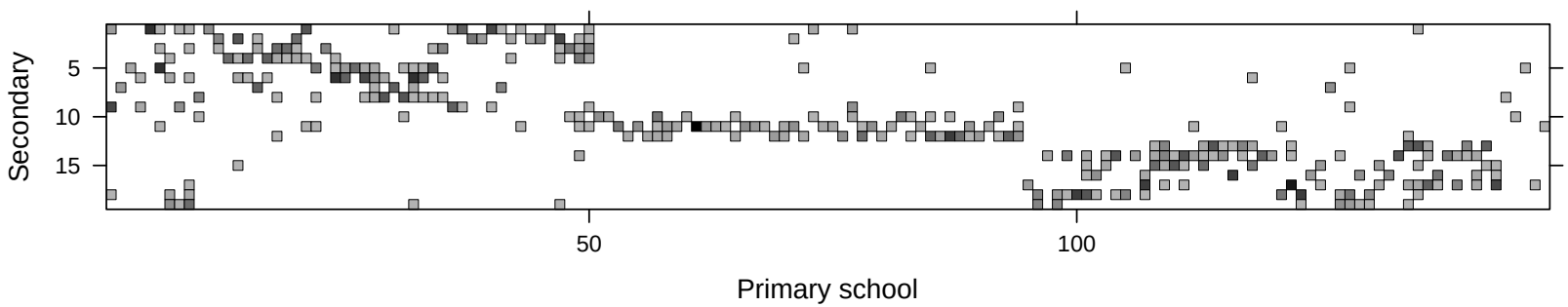
```
> anova(fm3M, fm2M)

Data: Early
Models:
fm3M: cog ~ tos + tos:trt + (1 | id) + (0 + tos | id)
fm2M: cog ~ tos + (tos | id) + tos:trt
     Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm3M  6  2389.0  2411.4 -1188.5
fm2M  7  2385.3  2411.4 -1185.6 5.7206      1    0.01677

> anova(fm4M, fm2M)

Data: Early
Models:
fm4M: cog ~ tos + tos:trt + (1 | id)
fm2M: cog ~ tos + (tos | id) + tos:trt
     Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm4M  5  2387.0  2405.7 -1188.5
fm2M  7  2385.3  2411.4 -1185.6 5.7206      2    0.05725
```

Outline

A smaller, non-nested unbalanced example

To examine the structure of non-nested, unbalanced observational data we need a smaller example than the 1.68 million observations in the 10 years of grade point scores.

```
> str(ScotsSec)

'data.frame': 3435 obs. of  6 variables:
 $ verbal : num  11 0 -14 -6 -30 -17 -17 -11 -9 -19 ...
 $ attain : num  10 3 2 3 2 2 4 6 4 2 ...
 $ primary: Factor w/ 148 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1
 $ sex    : Factor w/ 2 levels "M","F": 1 2 1 1 2 2 2 1 1 1 ...
 $ social : num  0 0 0 20 0 0 0 0 0 0 ...
 $ second : Factor w/ 19 levels "1","2","3","4",..: 9 9 9 9 9 9 1 1 9 9

> stab <- xtabs(~second + primary, ScotsSec, sparse = TRUE)
```



Mean attainment by school

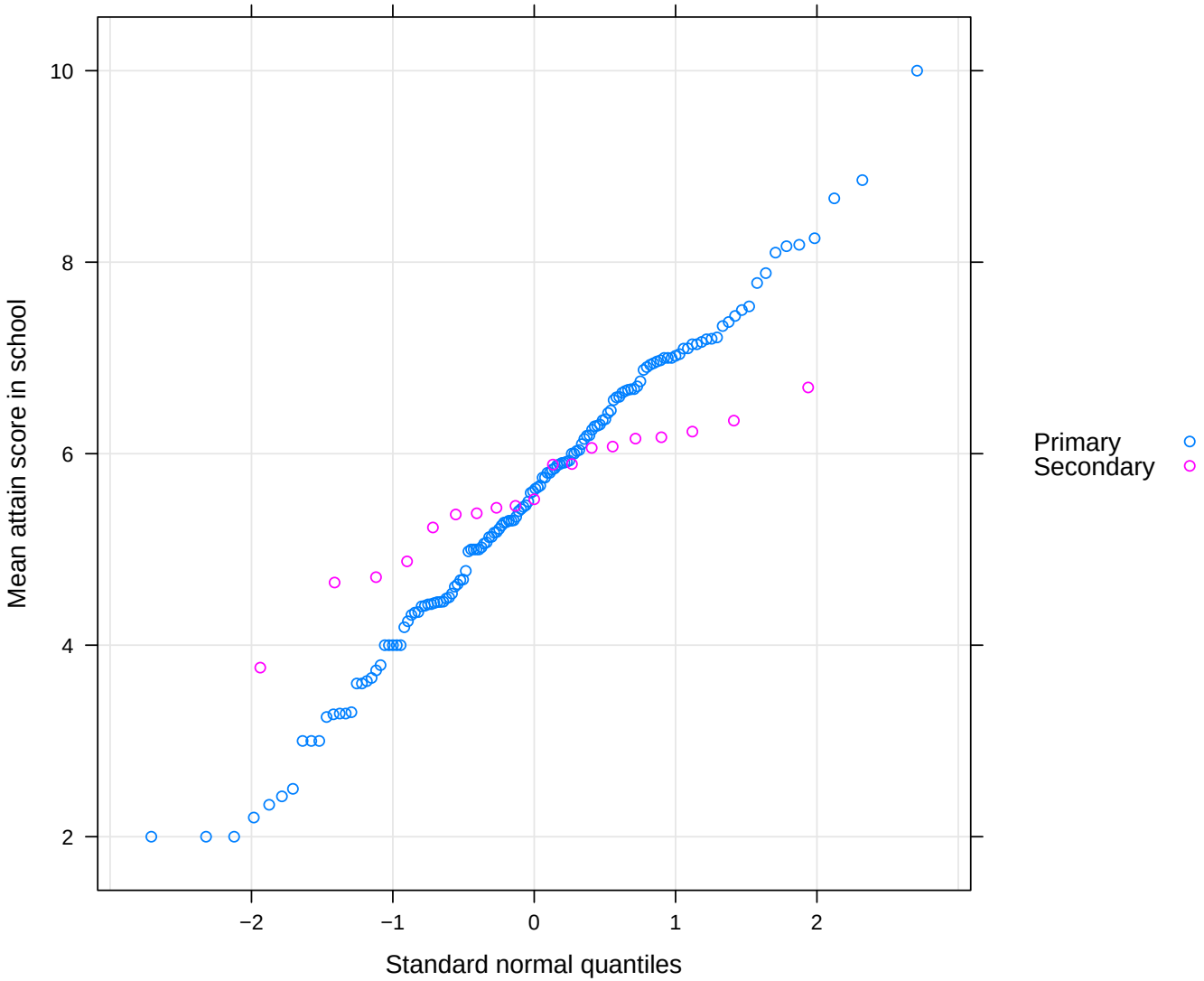
```
> head(patt)
```

	mattain	n	type
1P	4.425926	54	Primary
2P	5.285714	7	Primary
3P	8.666667	3	Primary
4P	6.285714	7	Primary
5P	4.679245	53	Primary
6P	5.927273	55	Primary

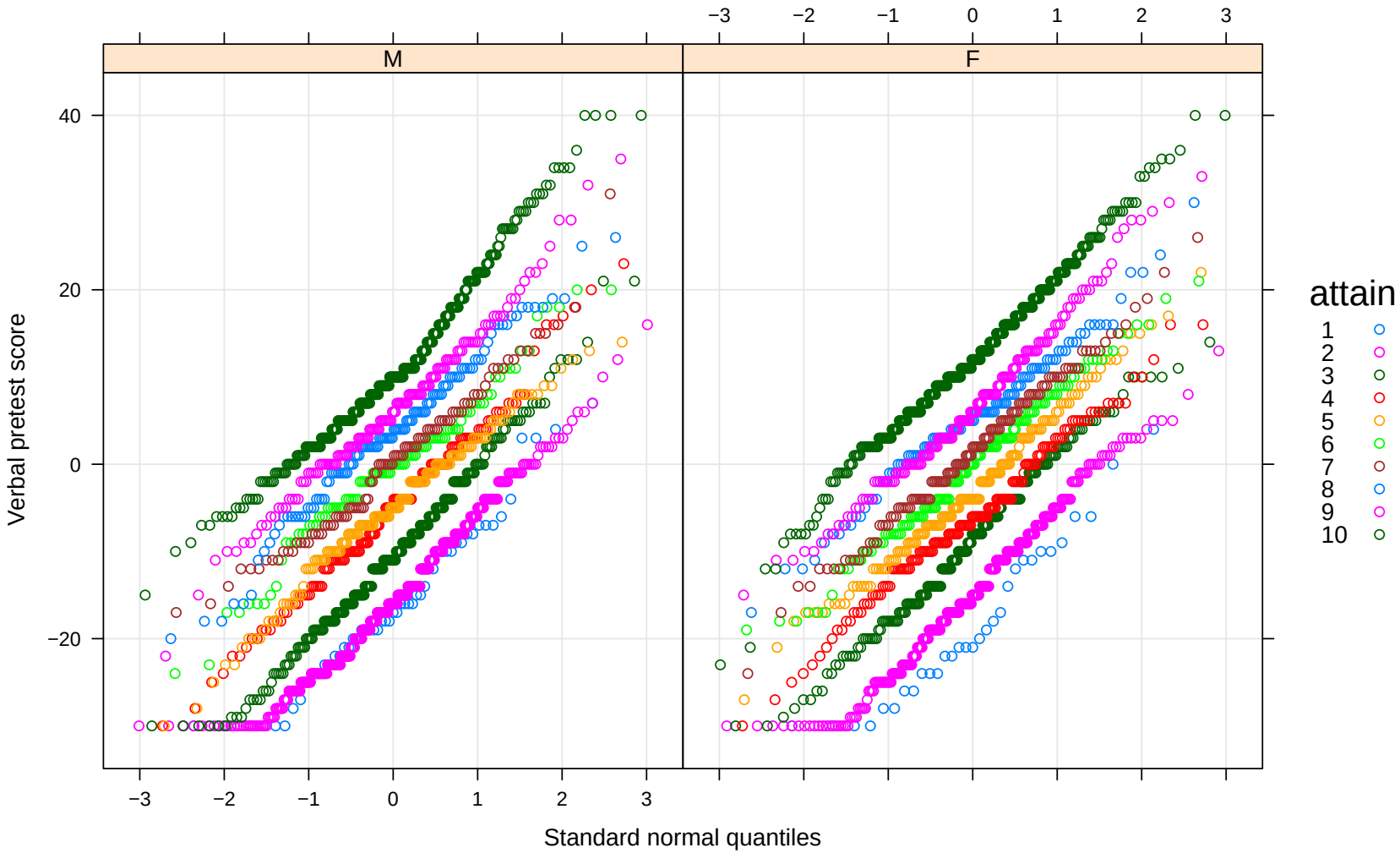
```
> head(satt)
```

	mattain	n	type
1S	5.365297	219	Secondary
2S	6.060302	199	Secondary
3S	5.455128	156	Secondary
4S	6.345324	139	Secondary
5S	6.074286	175	Secondary
6S	5.892000	250	Secondary

Normal probability plot of mean attainment by school



Probability plot of pretest by posttest and sex



An LMM for the secondary school data

```
Linear mixed model fit by REML
Formula: attain ~ verbal + sex + (1 | primary) + (verbal | second)
   Data: ScotsSec
   AIC   BIC logLik deviance REMLdev
 14875 14924  -7429    14842   14859
Random effects:
 Groups   Name        Variance   Std.Dev.  Corr
 primary  (Intercept) 2.7825e-01 0.5274908
 second   (Intercept) 2.4862e-02 0.1576754
          verbal      4.0435e-05 0.0063588 0.799
 Residual             4.2434e+00 2.0599441
Number of obs: 3435, groups: primary, 148; second, 19
Fixed effects:
            Estimate Std. Error t value
(Intercept) 5.912760   0.080172   73.75
verbal      0.159249   0.003174   50.17
sexF        0.113791   0.071440    1.59
```

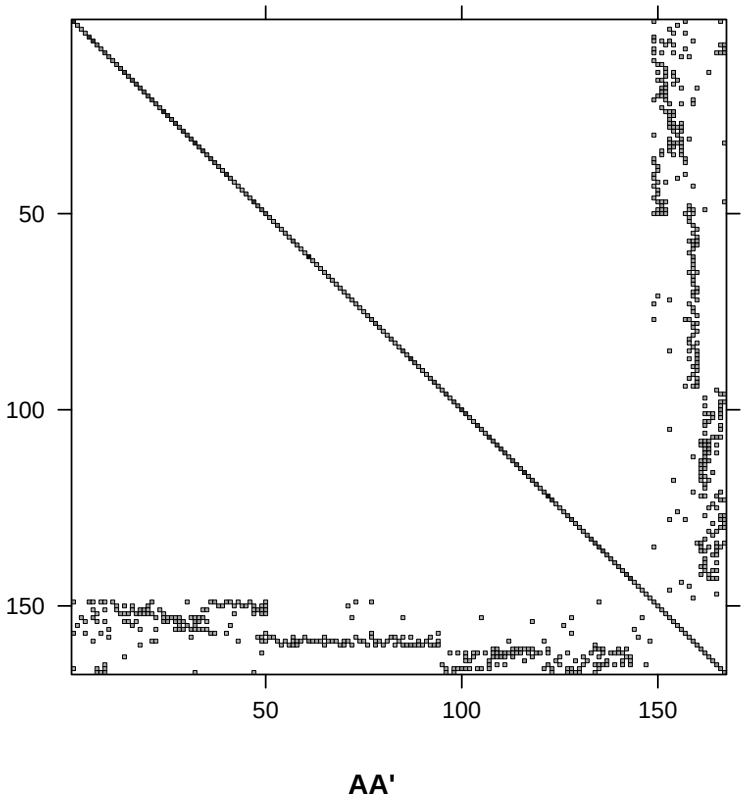
$n = 3435$, $p = 3$, $k = 2$, $n_1 = 148$, $n_2 = 19$, $q_1 = 1$, $q_2 = 2$, $q = 186$

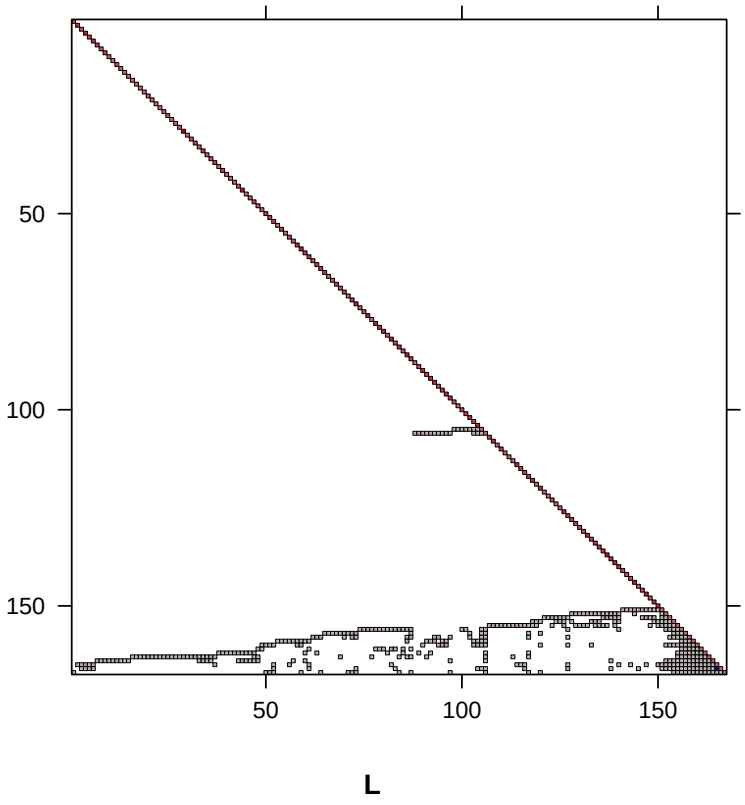
Reduced LMM for the secondary school data

```
Linear mixed model fit by REML
Formula: attain ~ verbal + sex + (1 | primary) + (1 | second)
   Data: ScotsSec
   AIC   BIC logLik deviance REMLdev
 14872 14909  -7430    14843   14860
Random effects:
 Groups   Name        Variance Std.Dev.
 primary  (Intercept) 0.276261 0.52561
 second   (Intercept) 0.014455 0.12023
 Residual             4.251960 2.06203
Number of obs: 3435, groups: primary, 148; second, 19
Fixed effects:
            Estimate Std. Error t value
(Intercept) 5.919273   0.076151   77.73
verbal      0.159593   0.002778   57.46
sexF        0.115966   0.071463    1.62
```

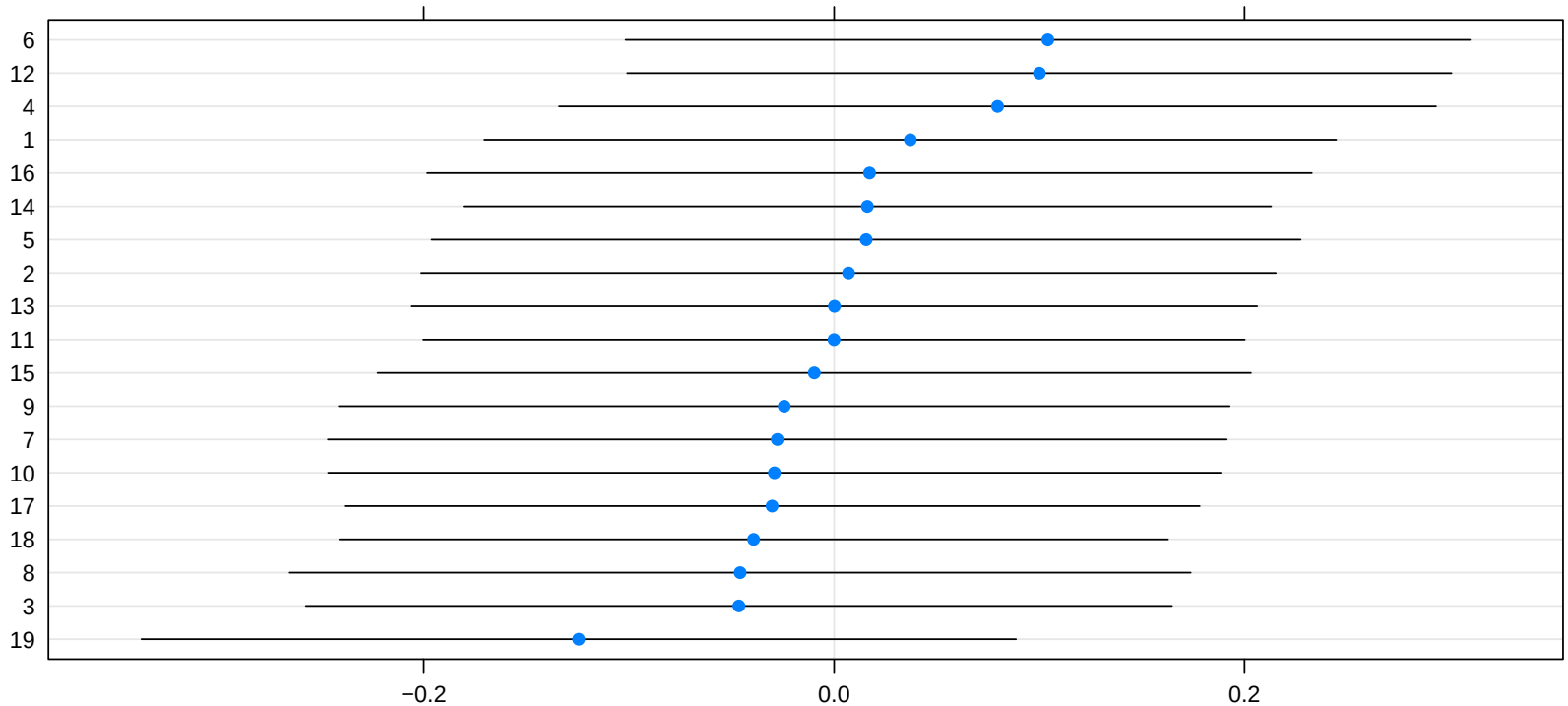
$n = 3435$, $p = 3$, $k = 2$, $n_1 = 148$, $n_2 = 19$, $q_1 = 1$, $q_2 = 1$, $q = 167$

Reordering unbalanced, non-nested data





Secondary school random effects are poorly defined

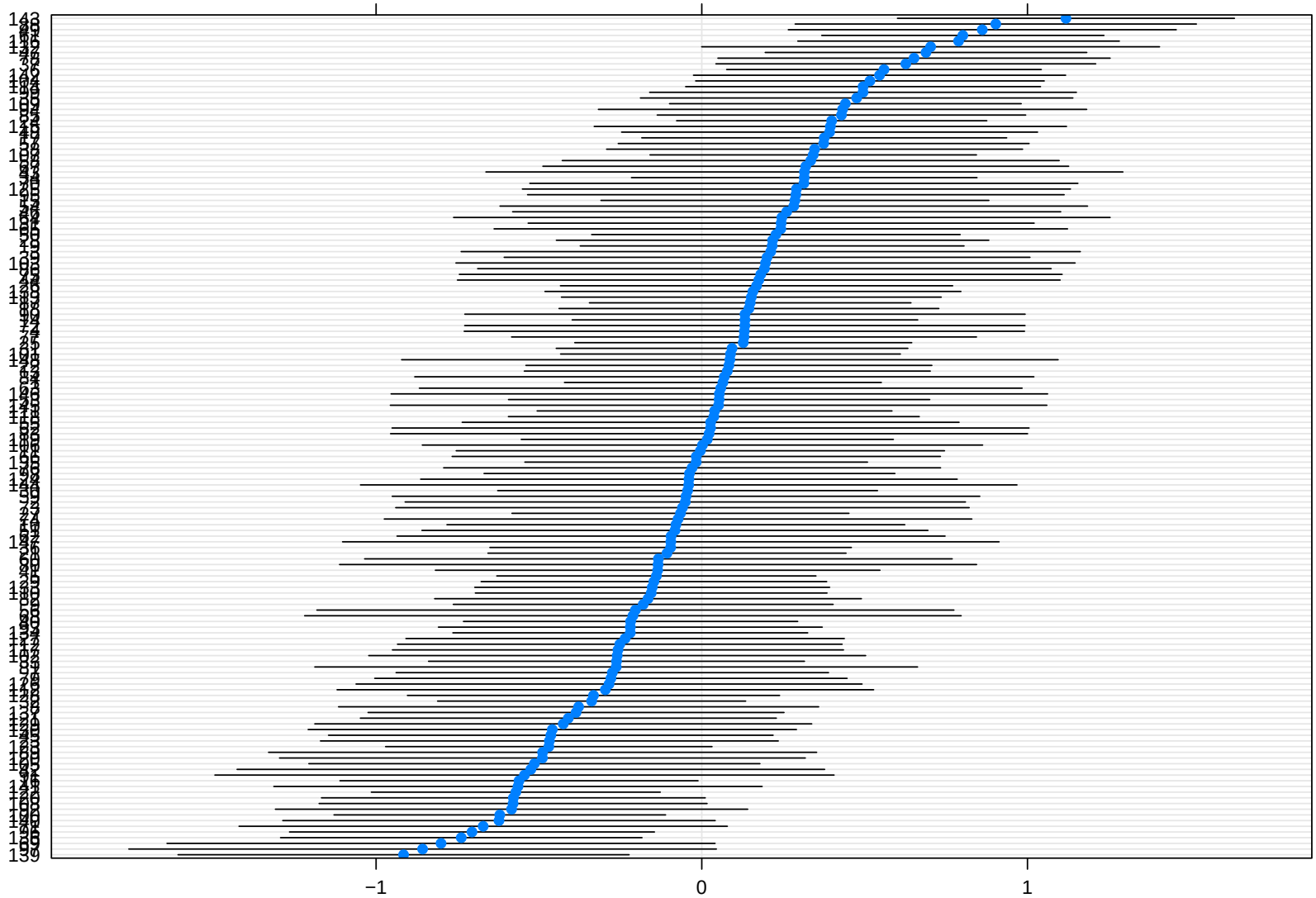


These indicate we should re-fit the model.

Model without random effects for second

```
> Sm3 <- update(Sm2, . ~ . - (1 | second))
> anova(Sm3, Sm2)

Data: ScotsSec
Models:
Sm3: attain ~ verbal + sex + (1 | primary)
Sm2: attain ~ verbal + sex + (1 | primary) + (1 | second)
    Df     AIC     BIC  logLik Chisq Chi Df Pr(>Chisq)
Sm3  5 14853.3 14884.0 -7421.6
Sm2  6 14855.0 14891.8 -7421.5 0.295      1      0.587
```

Random effects for primary are not that remarkable



Outline

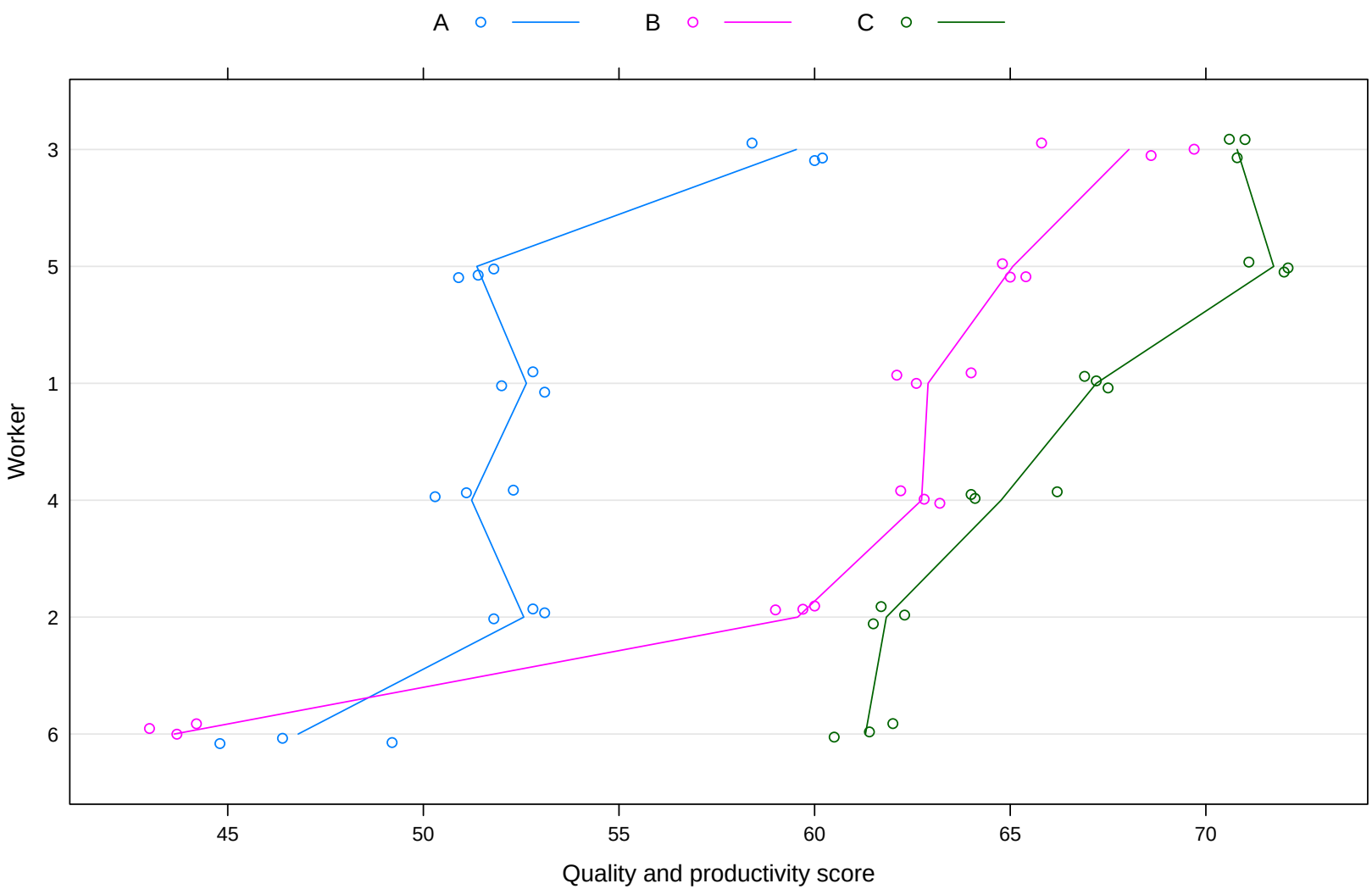
Interactions of covariates and grouping factors

- For longitudinal data, having a random effect for the slope w.r.t. time by subject is reasonably easy to understand.
- Although not generally presented in this way, these random effects are an interaction term between the grouping factor for the random effect (subject) and the time covariate.
- We can also define interactions between discrete covariates in the fixed-effects terms and a random-effects grouping factor. However, there is more than one way to define such an interaction.
- Different ways of expressing such interactions lead to different numbers of random effects.
- Models with interactions defined in different ways have levels of complexity, affecting both their expressive power and the ability to estimate all the parameters in the model.

Machines data

- Milliken and Johnson (1989) provide (probably artificial) data on an experiment to measure productivity according to the machine being used for a particular operation.
- In the experiment, a sample of six different operators used each of the three machines on three occasions — a total of nine runs per operator.
- These three machines were the specific machines of interest and we model their effect as a fixed-effect term.
- The operators represented a sample from the population of potential operators. We model this factor, (`Worker`), as a random effect.
- This is a replicated "subject/stimulus" design with a fixed set of stimuli that are themselves of interest. (In other situations the stimuli may be a sample from a population of stimuli.)

Machines data plot



Comments on the data plot

- There are obvious differences between the scores on different machines.
- It seems likely that `Worker` will be a significant random effect, especially when considering the low variation within replicates.
- There also appears to be a significant `Worker:Machine` interaction. `Worker` 6 has a very different pattern w.r.t. machines than do the others.
- We can approach the interaction in one of two ways: define simple, scalar random effects for `Worker` and for the `Worker:Machine` interaction or define vector-valued random effects for `Worker`

Random effects for subject and subject/stimulus

```
> print(fm1 <- lmer(score ~ Machine + (1 | Worker) +
+     (1 | Worker:Machine), Machines), corr = FALSE)
Linear mixed model fit by REML
Formula: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
   Data: Machines
   AIC   BIC logLik deviance REMLdev
 227.7 239.6 -107.8    225.5   215.7
Random effects:
 Groups         Name        Variance Std.Dev.
 Worker:Machine (Intercept) 13.90963 3.72956
 Worker         (Intercept) 22.85529 4.78072
 Residual                    0.92464 0.96158
Number of obs: 54, groups: Worker:Machine, 18; Worker, 6
Fixed effects:
            Estimate Std. Error t value
(Intercept)   52.356      2.486  21.062
MachineB       7.967      2.177   3.659
MachineC      13.917      2.177   6.393
```

Vector-valued random effects by subject

```
> print(fm2 <- lmer(score ~ Machine + (0 + Machine |
+     Worker), Machines), corr = FALSE)
```

```
Linear mixed model fit by REML
Formula: score ~ Machine + (0 + Machine | Worker)
   Data: Machines
   AIC   BIC logLik deviance REMLdev
 228.3 248.2 -104.2    216.6   208.3
Random effects:
 Groups   Name     Variance Std.Dev. Corr
 Worker   MachineA 16.64098 4.07934
          MachineB 74.39558 8.62529  0.803
          MachineC 19.26646 4.38936  0.623 0.771
 Residual           0.92463 0.96158
Number of obs: 54, groups: Worker, 6
Fixed effects:
            Estimate Std. Error t value
(Intercept)   52.356      1.681  31.150
MachineB       7.967      2.421   3.291
MachineC      13.917      1.540   9.037
```

Comparing the model fits

- Although not obvious from the specifications, the model fits are nested. If the variance-covariance matrix for the vector-valued random effects has a special form, called *compound symmetry*, the model reduces to model `fm1`.
- The p-value from this comparison is borderline significant.

```
> fm2M <- update(fm2, REML = FALSE)
> fm1M <- update(fm1, REML = FALSE)
> anova(fm2M, fm1M)
```

```
Data: Machines
Models:
fm1M: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
fm2M: score ~ Machine + (0 + Machine | Worker)
     Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm1M  6  237.27  249.20 -112.64
fm2M 10  236.42  256.31 -108.21 8.8516      4    0.06492
```

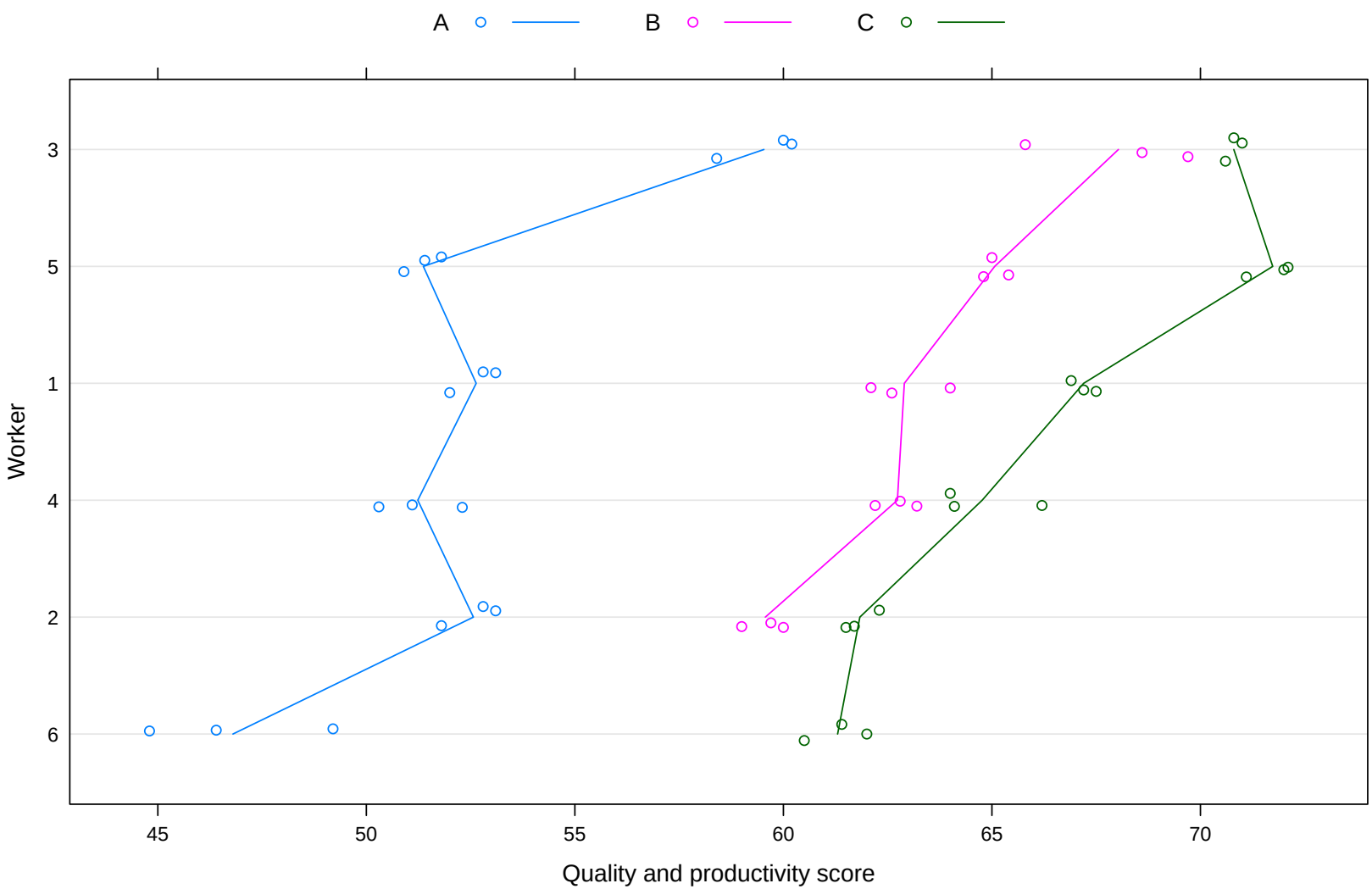
Model comparisons eliminating the unusual combination

- In a case like this we may want to check if a single, unusual combination (`Worker` 6 using `Machine` "B") causes the more complex model to appear necessary. We eliminate that unusual combination.

```
> Machines1 <- subset(Machines, Worker != "6" | Machine !=
+     "B")
> xtabs(~Machine + Worker, Machines1)
```

```
       Worker
Machine 1 2 3 4 5 6
      A 3 3 3 3 3 3
      B 3 3 3 3 3 0
      C 3 3 3 3 3 3
```

Machines data after eliminating the unusual combination



Model comparisons without the unusual combination

```
> fm1aM <- lmer(score ~ Machine + (1 | Worker) + (1 |
+     Worker:Machine), Machines1, REML = FALSE)
> fm2aM <- lmer(score ~ Machine + (0 + Machine | Worker),
+     Machines1, REML = FALSE)
> anova(fm2aM, fm1aM)
```

```
Data: Machines1
Models:
fm1aM: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
fm2aM: score ~ Machine + (0 + Machine | Worker)
      Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
fm1aM  6 208.554 220.145 -98.277
fm2aM 10 208.289 227.607 -94.144 8.2655      4    0.08232
```

Trade-offs when defining interactions

- It is important to realize that estimating scale parameters (i.e. variances and covariances) is considerably more difficult than estimating location parameters (i.e. means or fixed-effects coefficients).
- A vector-valued random effect term having q_i random effects per level of the grouping factor requires $q_i(q_i + 1)/2$ variance-covariance parameters to be estimated. A simple, scalar random effect for the interaction of a "random-effects" factor and a "fixed-effects" factor requires only 1 additional variance-covariance parameter.
- Especially when the "fixed-effects" factor has a moderate to large number of levels, the trade-off in model complexity argues against the vector-valued approach.
- One of the major sources of difficulty in using the `lme4` package is the tendency to overspecify the number of random effects per level of a grouping factor.

Outline

Definition of linear mixed models

- As previously stated, we define a linear mixed model in terms of two random variables: the n-dimensional $\boldsymbol{\mathcal{Y}}$ and the q-dimensional $\boldsymbol{\mathcal{B}}$
- The probability model specifies the conditional distribution

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

and the unconditional distribution

$$\boldsymbol{\mathcal{B}} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})\right), \quad (\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}) \perp \boldsymbol{\mathcal{B}}$$

as independent, multivariate Gaussian distributions depending on the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ.

- The relative variance-covariance matrix for $\boldsymbol{\mathcal{B}}$, written $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, can be factored as

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{\theta})' = (\boldsymbol{T}\boldsymbol{S})(\boldsymbol{T}\boldsymbol{S})'.$$

We say that the product $\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})$ is a left square-root factor of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

The conditional distribution, $\mathcal{Y}|\mathcal{B}$

- The mean of the conditional distribution, $\mathcal{Y}|\mathcal{B}$, is a linear function of β and b.

$$\mu_{\mathcal{Y}|\mathcal{B}}(b) = \mathrm{E}[\mathcal{Y}|\mathcal{B} = b] = \eta = X\beta + Zb$$

- For generalized linear models we will distinguish between the conditional mean, $\mu_{\mathcal{Y}|\mathcal{B}}(b)$, which may be bounded, and the linear predictor, η, which is always unbounded. For linear mixed models, $\mu_{\mathcal{Y}|\mathcal{B}}(b) = \eta$.
- Components of $\mathcal{Y}$ are *conditionally independent*, given $\mathcal{B}$. That is, the conditional distribution, $(\mathcal{Y}|\mathcal{B} = b)$, is determined by the (scalar) distribution of each component.
- Hence, the conditional distribution, $(\mathcal{Y}|\mathcal{B} = b)$, is completely determined by the conditional mean, $\mu_{\mathcal{Y}|\mathcal{B}}$, and the common scale parameter, σ.

The unscaled conditional density of $\mathcal{B}|\mathcal{Y} = y$

- Because it is y, not b, that we observe, we are interested in evaluating the other conditional distribution, $(\mathcal{B}|\mathcal{Y} = y)$. We will write its density as $[\mathcal{B}|\mathcal{Y}](b|y)$ (it is always continuous, even when, as in some GLMMs, $\mathcal{Y}$ is discrete).
- Given y, θ, β and, if used, σ, we can evaluate $[\mathcal{B}|\mathcal{Y}](b|y)$, up to a scale factor, as $[\mathcal{Y}|\mathcal{B}](y|b)\,[\mathcal{B}](b)$.
- The inverse of the scale factor,

$$\int_{\mathbb{R}^q} [\mathcal{Y}|\mathcal{B}](y|b)\,[\mathcal{B}](b)\,db,$$

is exactly the *likelihood*, $L(\theta, \beta, \sigma^2|y)$ (or $L(\theta, \beta, |y)$ when σ is not used).

The unscaled conditional density of $\mathcal{U}|\mathcal{Y} = y$

- To simplify the integral defining the likelihood, we change the variable of integration to $\boldsymbol{u}$, where $\boldsymbol{\mathcal{U}}$ is a vector-valued random variable with unconditional distribution $\boldsymbol{\mathcal{U}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_q)$ (or $\boldsymbol{\mathcal{U}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$, when σ is not used), and $\boldsymbol{\mathcal{B}} = \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{\mathcal{U}}$.
- The linear predictor, $\boldsymbol{\eta}$, which determines the conditional density, $[\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}](\boldsymbol{y}|\boldsymbol{u})$, becomes

$$\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{A}(\boldsymbol{\theta})'\boldsymbol{P}'\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{A}(\boldsymbol{\theta})' = \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})$, and likelihood

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) = \int_{\mathbb{R}^q} [\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}](\boldsymbol{y}|\boldsymbol{u})\, [\boldsymbol{\mathcal{U}}](\boldsymbol{u})\, d\boldsymbol{u}.$$

Maximizing the unscaled density $\mathcal{U}|\mathcal{Y}=y$

- In our general strategy for evaluating the likelihood, $L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma^2|\boldsymbol{y})$, we first maximize the unscaled density of $\mathcal{U}|\mathcal{Y}=\boldsymbol{y}$, w.r.t. $\boldsymbol{u}$.
- Both $[\mathcal{Y}|\mathcal{U}](\boldsymbol{y}|\boldsymbol{u})$ and $[\mathcal{U}](\boldsymbol{u})$ are *spherical* normal densities, which means that the components are independent with constant variance, e.g. $\mathrm{Var}(\mathcal{U})=\sigma^2\boldsymbol{I}$, ("spherical" because the contours of constant density are spheres).
- That is, probability density is related to the (squared) lengths, $\|\boldsymbol{y}-\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\|^2$ and $\|\boldsymbol{u}\|^2$, with the same scale factor, σ^2.
- The *conditional mode* of $\boldsymbol{U}|\boldsymbol{Y}$ – the value that maximizes the conditional density (and also the unscaled version) – does not depend on σ^2.

$$\begin{aligned}\tilde{\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\beta}) &= \arg\max_{\boldsymbol{u}}[\mathcal{Y}|\mathcal{U}](\boldsymbol{y}|\boldsymbol{u})\,[\mathcal{U}](\boldsymbol{u})\\ &= \arg\min_{\boldsymbol{u}}\left(\|\boldsymbol{y}-\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\|^2+\|\boldsymbol{u}\|^2\right)\end{aligned}$$

Solving for the conditional mode

- Incorporating the definition of $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}$ provides

$$\|\boldsymbol{y}-\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\|^2=\|\boldsymbol{y}-\boldsymbol{A}'\boldsymbol{P}'\boldsymbol{u}-\boldsymbol{X}\boldsymbol{\beta}\|^2$$

- Recall that $\boldsymbol{P}$ is a permutation matrix. These have the property that $\boldsymbol{P}^{-1}=\boldsymbol{P}'$, allowing us to write

$$\|\boldsymbol{0}-\boldsymbol{P}'\boldsymbol{u}\|^2=\boldsymbol{u}'\boldsymbol{P}\boldsymbol{P}'\boldsymbol{u}=\boldsymbol{u}'\boldsymbol{u}=\|\boldsymbol{u}\|^2$$

- Combining these produces

$$\tilde{\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\beta})=\arg\min_{\boldsymbol{u}}\left\|\begin{bmatrix}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{bmatrix}-\begin{bmatrix}\boldsymbol{A}'\\ \boldsymbol{I}\end{bmatrix}\boldsymbol{P}'\boldsymbol{u}\right\|^2$$

Hence, $\tilde{\boldsymbol{u}}$ satisfies

$$\boldsymbol{P}\left(\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{I}\right)\boldsymbol{P}'\tilde{\boldsymbol{u}}=\boldsymbol{L}\boldsymbol{L}'\tilde{\boldsymbol{u}}=\boldsymbol{P}\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$$

where $\boldsymbol{L}(\boldsymbol{\theta})$ is the sparse left Cholesky factor of $\boldsymbol{P}\left(\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})'+\boldsymbol{I}\right)\boldsymbol{P}'$.

Evaluating the likelihood - linear mixed models

- Because $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}$ depends linearly on both $\boldsymbol{u}$ and $\boldsymbol{\beta}$, the conditional mode $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and the conditional maximum likelihood estimate, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, can be determined simultaneously as the solutions to a penalized least squares problem

$$\begin{bmatrix}\tilde{\boldsymbol{u}}(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \arg\min_{\boldsymbol{u},\boldsymbol{\beta}} \left\| \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{A}'\boldsymbol{P}' & \boldsymbol{X}\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix} \begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\beta}\end{bmatrix} \right\|^2$$

  for which the solution satisfies

$$\begin{bmatrix}\boldsymbol{P}\left(\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{I}\right)\boldsymbol{P}' & \boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\\ \boldsymbol{X}'\boldsymbol{A}'\boldsymbol{P}' & \boldsymbol{X}'\boldsymbol{X}\end{bmatrix} \begin{bmatrix}\tilde{\boldsymbol{u}}(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \begin{bmatrix}\boldsymbol{P}\boldsymbol{A}\boldsymbol{y}\\ \boldsymbol{X}'\boldsymbol{y}\end{bmatrix}$$

- The Cholesky factor of the system matrix for the PLS problem is

$$\begin{bmatrix}\boldsymbol{P}\left(\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{I}\right)\boldsymbol{P}' & \boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\\ \boldsymbol{X}'\boldsymbol{A}'\boldsymbol{P}' & \boldsymbol{X}'\boldsymbol{X}\end{bmatrix} = \begin{bmatrix}\boldsymbol{L} & \boldsymbol{0}\\ \boldsymbol{R}'_{\boldsymbol{ZX}} & \boldsymbol{R}'_{\boldsymbol{X}}\end{bmatrix} \begin{bmatrix}\boldsymbol{L}' & \boldsymbol{R}_{\boldsymbol{ZX}}\\ \boldsymbol{0} & \boldsymbol{R}_{\boldsymbol{X}}\end{bmatrix}$$

- The dense matrices $\boldsymbol{R}_{\boldsymbol{ZX}}$ and $\boldsymbol{R}_{\boldsymbol{X}}$ are stored in the `RZX` and `RX` slots, respectively.

Special case of linear mixed models (cont'd)

- It is not necessary to solve for $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$. All that is needed for evaluation of the profiled log-likelihood is the penalized residual sum of squares, r^2, and the determinant

$$|\boldsymbol{A}\boldsymbol{A}' + \boldsymbol{I}| = |\boldsymbol{L}|^2$$

- Because $\boldsymbol{L}$ is triangular, its determinant is simply the product of its diagonal elements.
- Because $\boldsymbol{A}\boldsymbol{A}' + \boldsymbol{I}$ is positive definite, $|\boldsymbol{L}|^2 > 0$.
- The profiled deviance, as a function of $\boldsymbol{\theta}$ only ($\boldsymbol{\beta}$ and σ^2 at their conditional estimates), is

$$d(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}|^2) + n\left(1 + \log\left(\frac{2\pi r^2}{n}\right)\right)$$

REML results

- Although not often derived in this form, Laird and Ware showed that the REML criterion can be derived as the integral of the likelihood w.r.t. $\boldsymbol{\beta}$.
- The same techniques as used to evaluate the integral w.r.t. $\boldsymbol{b}$ can be used to evaluate the integral for the REML criterion. In this case the integral introduces the factor $|\boldsymbol{R_X}|^2$.
- The profiled REML deviance, as a function of $\boldsymbol{\theta}$ only (σ at its conditional estimate), is

$$d_R(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}|^2|\boldsymbol{R_X}|^2) + (n-p)\left(1 + \log\left(\frac{2\pi r^2}{n-p}\right)\right)$$

Recap

- For a linear mixed model, even one with a huge number of observations and random effects like the model for the grade point scores, evaluation of the ML or REML profiled deviance, given a value of $\boldsymbol{\theta}$, is straightforward. It involves updating $\boldsymbol{T}$ and $\boldsymbol{S}$, then updating $\boldsymbol{A}$, $\boldsymbol{L}$, $\boldsymbol{R_{ZX}}$, $\boldsymbol{R_X}$, calculating the penalized residual sum of squares, r and a couple of determinants of triangular matrices.
- The profiled deviance can be optimized as a function of $\boldsymbol{\theta}$ only. The dimension of $\boldsymbol{\theta}$ is usually very small. For the grade point scores there are only three components to $\boldsymbol{\theta}$.