Applied Biostatistics

https://moodle.epfl.ch/course/view.php?id=15590

- Introduction to mixed models
- Corn dataset and model
- 3 Definition of linear mixed effects models
- Parameter estimation
- Crossed random effects grouping : Penicillin
- Nested random effects grouping : Rat liver data

Mixed models – why?

- Mixed-effects models provide a flexible and powerful tool for the analysis of grouped data, including:
  - blocked designs
  - repeated measures (each subject measured for each condition; individuals are ‘blocks’)
  - Longitudinal data (measures repeated over time)
  - multilevel data
- Offer flexibility in modeling within-group correlation often present in grouped data
- Handle balanced and unbalanced data in a unified framework
- There is reliable, efficient software for fitting

Books on mixed models

- José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*.
- Brady West, Kathleen B. Welch, Andrzej T. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Available as e-book: `http://www.crcnetbase.com/isbn/9781420010435`
- A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R*.
- Julian J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*

Useful resources

- Douglas Bates, developer of *R* packages `nlme` and `lme4`, gave a 3 day course at UniL on mixed model analysis
- `http://www.unil.ch/ee/page64467.html`
- (We use some of his examples here)
- R-forge site for `lme4`: `http://lme4.r-forge.r-project.org/`
- (Includes links to draft `lmer` book, slides, *R* code)

Effects – fixed and random

- Mixed-effects models describe the relationship between a *response* variable and one or more *covariates* recorded with it
- Consider models based on a *linear predictor* incorporating *coefficients* estimated from observed data
- When levels of a covariate are fixed and reproducible (*e.g.* a covariate `sex` that has levels `male` and `female`) ⇒ *fixed effects* parameters
- When levels of a covariate correspond to the particular experimental units ⇒ *random effects*

y=B0 +B1sex (with sex {0F,1M}) --- Fixed

y=B0+B1sex+B2hospital (hospital {0,1,2})

We only care about these 3 hosps. -> fixed effect

otherwise -> random effect

Fixed effects

- Generally speaking, a factor is *fixed* if the levels of the factor were selected by the investigator to compare the effects of the levels to one another
- Fixed effects influence only the *mean* of the response Y
- Fixed effects are represented by constant parameters, we are interested in estimating them

Random effects

- A factor is *random* if the effects associated with the levels of the factor can be viewed as being like a random sample from a population of effects
- Random effects are represented by (unobserved) random variables, usually assumed to follow a normal distribution
- Random effects influence only the *variance* of the response Y
- For random effects, we can make statements about *variation* in the population of random effects
- Depending on the *goals* of the study, the same factor may be considered either as fixed or random

The Corn dataset

- Here we will consider a subset of data on corn yields from the Caribbean island of Antigua, available as the dataset `ant111b` from the `DAAG` package
- Data are yields from 4 parcels at eight sites
- The `ant111b` data are a balanced one-way classification of the `harvwt` of corn produced at eight `sites`
- Let's have a look:

```
> str(ant111b)

'data.frame': 32 obs. of  9 variables:
 $ site   : Factor w/ 8 levels "DBAN","LFAN",..: 1 2 3 4 5 6 7 8 1 2 ...
 $ parcel: Factor w/ 4 levels "I","II","III",..: 1 1 1 1 1 1 1 1 2 2 ..
 $ code   : num  58 58 58 58 58 58 58 58 58 58 ...
 $ island: num  1 1 1 1 1 1 1 1 1 1 ...
 $ id     : num  3 40 186 256 220 ...
 $ plot   : num  3 4 5.5 4.5 3.5 5 7 7 15.5 15 ...
 $ trt    : num  111 111 111 111 111 111 111 111 111 111 ...
 $ ears   : num  43.5 40.5 20 42.5 31.5 32.5 43.5 50 46 46.5 ...
 $ harvwt: num  5.16 2.93 1.73 6.79 3.25 ...
```

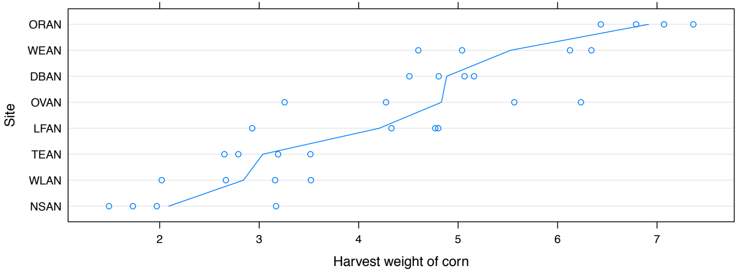
Corn summary

```
> summary(ant111b)
      site   parcel       code        island        id
 DBAN   :4   I  :8   Min.   :58   Min.   :1   Min.   :  3.00
 LFAN   :4   II :8   1st Qu.:58   1st Qu.:1   1st Qu.: 74.62
 NSAN   :4   III:8   Median :58   Median :1   Median :145.75
 ORAN   :4   IV :8   Mean   :58   Mean   :1   Mean   :144.47
 OVAN   :4           3rd Qu.:58   3rd Qu.:1   3rd Qu.:214.25
 TEAN   :4           Max.   :58   Max.   :1   Max.   :283.50
 (Other):8
      plot            trt           ears           harvwt
 Min.   : 3.00   Min.   :111   Min.   :20.00   Min.   :1.490
 1st Qu.:10.38   1st Qu.:111   1st Qu.:40.12   1st Qu.:3.103
 Median :18.75   Median :111   Median :43.00   Median :4.420
 Mean   :18.47   Mean   :111   Mean   :41.22   Mean   :4.292
 3rd Qu.:26.00   3rd Qu.:111   3rd Qu.:45.62   3rd Qu.:5.261
 Max.   :33.50   Max.   :111   Max.   :56.00   Max.   :7.365
```

The site effect

- There is no inherent ordering of the levels of the `site` factor, we can reorder them for our convenience
- The particular sites observed are just a selection of the possible sites on the island
- We want to focus on estimating the *variability in yields* due to site-to-site variability
- The `site` factor will be used in *random effects* terms in our models

Corn data plot



- The line joins the means of the harvest weight of the individual sites, which have been reordered by increasing mean `harvwt`
- The vertical positions can be `jitter`ed slightly to reduce overplotting

A mixed effects model for corn yield

```
> (ant111b.lmer <- lmer(harvwt ~ 1 + (1 | site), data=ant111b) )
```

```
Linear mixed model fit by REML
Formula: harvwt ~ 1 + (1 | site)
   Data: ant111b
   AIC   BIC logLik deviance REMLdev
 100.4 104.8 -47.21    95.08   94.42
Random effects:
 Groups   Name        Variance Std.Dev.
 site     (Intercept) 2.36773  1.53874
 Residual             0.57754  0.75996
Number of obs: 32, groups: site, 8
Fixed effects:
            Estimate Std. Error t value
(Intercept)   4.2917     0.5603   7.659
```

- Our model `ant111b.lmer` has one *fixed effect* parameter (the first `1`), the mean harvest weight, and one *random effect* term (`(1 | site)`), generating a *simple, scalar random effect* for each level of `site`

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 enclosed in parentheses
- The expression on the right of the '$|$' is evaluated as a *factor*, which we call the *grouping factor* for that term

Interpreting the output

- There are two sources of random variation, one for site and one for parcel within site
- The estimated variance components are $\sigma^2_{site} = 2.36773$ and $\sigma^2_{Residual} = 0.57754$
- The proportion of variation due to site is $\dfrac{\sigma^2_{site}}{\sigma^2_{site} + \sigma^2_{Residual}}$ $= 2.36773$ / ($2.36773 + 0.57754$) $\approx 80\%$

Extracting information from the fitted model

- `ant111b.lmer` is an object of class `"mer"` (*mixed effects representation*).
- There are many *extractor* functions that can be applied

```
> fixef(ant111b.lmer)

(Intercept)
     4.2917

> ranef(ant111b.lmer, drop = TRUE)

$site
     DBAN      LFAN      NSAN      ORAN      OVAN      TEAN
 0.559205 -0.079381 -2.075257  2.472606  0.509720 -1.183358
     WEAN      WLAN
 1.163623 -1.367157

> fitted(ant111b.lmer)

 [1] 4.8509 4.2123 2.2165 6.7643 4.8014 3.1084 5.4553 2.9246 4.8509
[10] 4.2123 2.2165 6.7643 4.8014 3.1084 5.4553 2.9246 4.8509 4.2123
[19] 2.2165 6.7643 4.8014 3.1084 5.4553 2.9246 4.8509 4.2123 2.2165
[28] 6.7643 4.8014 3.1084 5.4553 2.9246
```

Definition of mixed effects models

- Models with random effects are often written as

$$y_{ij} = \mu + b_i + \epsilon_{ij}, \quad b_i \sim \mathcal{N}(0, \sigma_b^2),$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, I; \quad j = 1, \ldots, J_i$$

- To avoid too many subscripts use vector/matrix notation
- A mixed-effects model incorporates two vector-valued random variables: the response vector, $\boldsymbol{\mathcal{Y}}$, and the random effects vector, $\boldsymbol{\mathcal{B}}$
- We observe the value, $\boldsymbol{y}$, of $\boldsymbol{\mathcal{Y}}$; we do not observe the value of $\boldsymbol{\mathcal{B}}$
- Random effects usually modeled as a multivariate Gaussian (or "normal") random variable, $\boldsymbol{\mathcal{B}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\boldsymbol{\theta}$ is a vector of *variance component parameters*.

Linear mixed models

- The conditional distribution, $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b})$, depends on $\boldsymbol{b}$ only through its mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}=\boldsymbol{b}}$
- The conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}=\boldsymbol{b}}$, depends on $\boldsymbol{b}$ and on the fixed effects parameter vector, $\boldsymbol{\beta}$, through a *linear predictor* expression, $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta}$
- *Model matrices* $\boldsymbol{Z}$ (random) and $\boldsymbol{X}$ (fixed) are determined from the form of the model and the values of the covariates.
- In a *linear mixed model* the conditional distribution is a "spherical" multivariate Gaussian

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

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

Simple, scalar random effects terms

- A term like `(1|site)` in an `lmer` formula is called a *simple, scalar random effects term*
- The expression on the right of the `"|"` operator (usually just the name of a variable) is evaluated as a factor, called the *grouping factor* for the term
- Suppose we have k such terms with $n_i, i = 1, \dots, k$ levels in the ith term's grouping factor. A scalar random effects term generates one random effect for each level of the grouping factor. If all the random effects terms are scalar terms then $q = \sum_{i=1}^{k} n_i$.
- The model matrix $\boldsymbol{Z}$ is the horizontal concatenation of k matrices. For a simple, scalar term, the ith vertical slice, which has n_i columns, is the indicator columns for the n_i levels of the ith grouping factor.

Conditional means of the random effects

- Technically speaking, we do not provide "estimates" of the random effects because they are not parameters
- So if the numbers provided by `ranef` aren't estimates, what are they?
- They are called BLUPs (Best Linear Unbiased Predictors) of the random effects
- Those values are the conditional means, $\boldsymbol{\mu}_{\mathcal{B}|\mathcal{Y}=\boldsymbol{y}}$, evaluated at the estimated parameter values

Fitted values

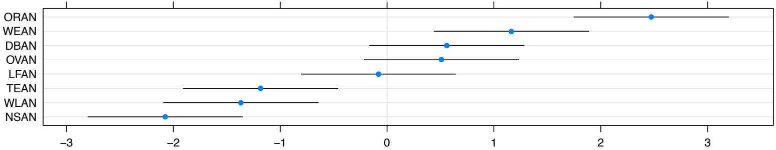
```
> means <- with(ant111b, sapply(split(harvwt, site), mean))
> siteFit <- with(ant111b, sapply(split(fitted(ant111b.lmer),
+ site), mean))
> print(data.frame(mean = means, fitted = siteFit))
```

```
        mean   fitted
DBAN 4.88500 4.850923
LFAN 4.20750 4.212337
NSAN 2.09000 2.216461
ORAN 6.91500 6.764325
OVAN 4.83250 4.801439
TEAN 3.03625 3.108361
WEAN 5.52625 5.455341
WLAN 2.84125 2.924561
```

- The fitted values are *not* just the sample means
- They are *shrinkage estimates* that are between the grand (overall) mean and the individual sample means

Caterpillar plot for ant111b.lmer

- For linear mixed models the conditional distribution of the random effects, given the data, written $(\boldsymbol{\mathcal{B}}|\boldsymbol{\mathcal{Y}} = \boldsymbol{y})$, is again a multivariate Gaussian distribution
- We can evaluate the means and standard deviations of the individual conditional distributions, $(\boldsymbol{\mathcal{B}}_j|\boldsymbol{\mathcal{Y}} = \boldsymbol{y}), j = 1, \dots, q$
- 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"



Parameter estimation

- We are familiar with *least squares estimation*, as we have done for linear models
- The idea is to estimate the unknown parameter values by minimizing the total of the squared errors
- ANOVA techniques can used in random effect estimation when the data are "pretty", but do not extend more generally and can be problematic (especially for unbalanced data)
- An alternative is provided by *maximum likelihood estimation* – here, we use distributional assumptions to write the *likelihood*, and maximize this quantity (ML estimation)
- This method has the appealing property that the estimates are the values that make the observed data most likely

Example: Binomial distribution

- The distribution of the number of successes X in a (1) fixed number n of (2) independent (3) Bernoulli (yes/no) trials, each with (4) constant success probability p, is called Binomial(n, p)
- For $X \sim Bin(n, p)$,

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- For a given p, we can write the probability of any possible data
- We can instead *consider the data as given* and look at the probability as a function of the unknown parameter p
- The probability function viewed in this way is referred to as the *likelihood function*

Maximum likelihood estimation

- One very intuitive way to estimate the parameter p is by the *method of maximum likelihood*
- For example, the obvious way to estimate p $(= X/n)$ turns out to be the *maximum likelihood estimator* *(MLE)* NA NA
- This method does not work in every case – use numerical optimization

Some properties of MLEs

- *Consistency*: *i.e.*, $\lim_{n\to\infty} P(\mid \hat{\theta}_n - \theta \mid < \epsilon) = 1, \forall\epsilon > 0$
- *Invariance*: if $\hat{\theta}$ is the MLE for the parameter θ, then $h(\hat{\theta})$ is the MLE for parameter $h(\theta)$
- *Asymptotically unbiased*, that is the bias goes to 0 as $n \to \infty$ (but may be biased in finite samples)
- *Asymptotic efficiency*, *i.e.* no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE
- *Asymptotically Normal*: *i.e.*, the distribution of $\hat{\theta}_n$ as $n \to \infty$ tends to a normal distribution; this provides a framework and justification for making *inferences* with MLEs (*e.g.* making a confidence interval)

REML estimates vs. ML estimates

- The default parameter estimation for linear mixed models is *restricted (or “residual”) maximum likelihood (REML)*
- Likelihood partitioned into two parts, one of which is free of the fixed effects – maximizing this produces REML estimates
- Maximum likelihood (ML) estimates can be requested by specifying `REML = FALSE` in the call to `lmer`
- Generally REML estimates of variance components are preferred – unbiased in some situations and 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, REML and ML estimates of the fixed effects are the same

Re-fitting the model for ML estimates

```
> (ant111b.lmer1 <- update(ant111b.lmer, REML = FALSE))
Linear mixed model fit by maximum likelihood
Formula: harvwt ~ 1 + (1 | site)
   Data: ant111b
 AIC   BIC logLik deviance REMLdev
 101 105.4 -47.51    95.03   94.47
Random effects:
 Groups   Name        Variance Std.Dev.
 site     (Intercept) 2.05372  1.43308
 Residual             0.57754  0.75996
Number of obs: 32, groups: site, 8
Fixed effects:
            Estimate Std. Error t value
(Intercept)   4.2917     0.5242   8.188
```

Estimates of variance components can be zero

- We know that the variance of the random effects is ≥ 0
- For some data sets the ML or REML estimate $\widehat{\sigma_b^2}$ is zero
- For example: when variability between groups is not large compared to the within-batch variability
- The mixed model with an estimated variance $\widehat{\sigma_b^2} = 0$ is equivalent to a model with only fixed effects terms

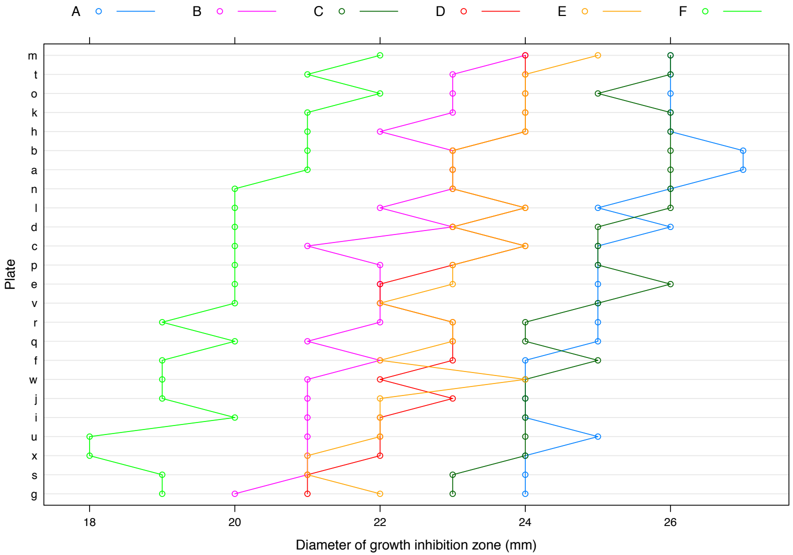
Penicillin dataset

```
> 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
```

- Six samples of penicillin were tested on each of 24 plates
- The response is diameter (mm) of the growth inhibition zone, providing a measurement of sample potency
- *Balanced*, *unreplicated* two-way *crossed* classification

Penicillin data plot



Model with crossed simple random effects for Penicillin

```
> (pen.lmer <- 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  0.84670
 sample   (Intercept) 3.73092  1.93156
 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 pen.lmer

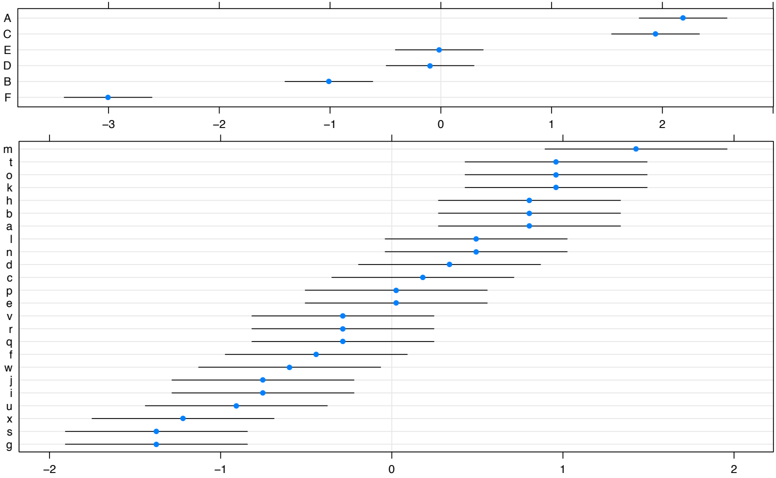
- The model for the $n = 144$ observations has $p = 1$ fixed effects parameter and $q = 30$ random effects from $k = 2$ random effects terms in the formula

```
> fixef(pen.lmer)

(Intercept)
     22.972

> ranef(pen.lmer, 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
 2.187245 -1.010563  1.938065 -0.096903 -0.013843 -3.004001
```

Prediction intervals for random effects

- The values returned by the `ranef` extractor are the conditional means (for a linear mixed model) $\boldsymbol{\mu}(\boldsymbol{\mathcal{B}}_j|\boldsymbol{\mathcal{Y}} = \boldsymbol{y})$ of the random effects, evaluated at the parameter estimates
- Can also evaluate the condtional variance-covariance of $\boldsymbol{\mathcal{B}}_j|\boldsymbol{\mathcal{Y}} = \boldsymbol{y}$ and use it to obtain a prediction interval
- These are returned by `ranef` when the optional argument `postVar` is TRUE
- We can visualize these prediction intervals for each set of random effects in a caterpillar plot

Prediction intervals for Penicillin random effects



Rat liver data

- In this experiments 3 treatments have been administered to 2 rats each
- From each of these 6 rats, three pieces of liver were taken
- Glycogen content was measured twice for each of the 18 pieces
- ⇒ In total, 36 observations

```
> rats <- read.table("rats.txt", header = T)
> head(rats)

  Glycogen Treatment Rat Liver
1      131         1   1     1
2      130         1   1     1
3      131         1   1     2
4      125         1   1     2
5      136         1   1     3
6      142         1   1     3
```

Structure of rat liver data I

```
> # attach(rats,warn.conflicts = FALSE)
> rats$Treatment <- with(rats, factor(Treatment))
> rats$Rat <- with(rats, factor(Rat))
> rats$Liver <- with(rats, factor(Liver))
> str(rats)
```

```
'data.frame': 36 obs. of  4 variables:
 $ Glycogen : int  131 130 131 125 136 142 150 148 140 143 ...
 $ Treatment: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ Rat      : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 2 2 2 2 ...
 $ Liver    : Factor w/ 3 levels "1","2","3": 1 1 2 2 3 3 1 1 2 2 ...
```

- There are 2 levels of `Rat` – but there are 6 rats
- There are 3 levels of `Liver` – but there are 18 liver pieces

Structure of rat liver data II

```
> xtabs(~ Treatment + Rat, rats, sparse=TRUE)

3 x 2 sparse Matrix of class "dgCMatrix"
  1 2
1 6 6
2 6 6
3 6 6

> xtabs(~ Rat + Liver, rats, sparse=TRUE)

2 x 3 sparse Matrix of class "dgCMatrix"
  1 2 3
1 6 6 6
2 6 6 6
```

- These tabulations suggest that the `Treatment` and `Rat` variables, and the `Rat` and `Liver` variables, are *crossed*

Implicit nesting

- Although the variable coding makes it appear that the variables are crossed, this is *NOT* the case
- The labels of the variable `Rat` (‘1’ and ‘2’) are only meaningful *within* a `Treatment`
- Similarly, the labels of `Liver` are only meaningful *within* `Rat`
- `Rat` is *nested* within `Treatment` (and `Liver` within `Rat` within `Treatment`), but that is not reflected in the data coding
- This is an example of an *implicitly nested* representation

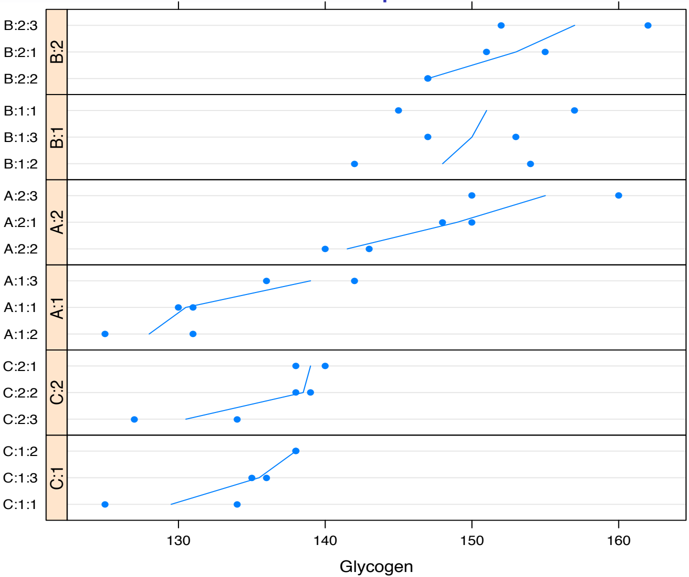
Avoid implicitly nested representations

- It used to be that nesting was nearly always coded implicitly (often due to software requirements that assumed a *hierarchy* of random effects)
- This practice is error prone and confusing, and not required by `lme4`, which allows for very general model specifications
- 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
- `Liver` samples were drawn from 6, not 2, distinct rats, so should be a factor with 18 levels (not 3); similarly for `Rat` within `Treatment` (6 not 2 levels)

Explicit nesting coding

```
> rats$Treatment <- factor(rats$Treatment, labels=LETTERS[1:3])
> rats$rr <- with(rats, Treatment:factor(Rat))
> rats$ll <- with(rats, Treatment:factor(Rat):factor(Liver))
> str(rats)

'data.frame': 36 obs. of  6 variables:
 $ Glycogen : int  131 130 131 125 136 142 150 148 140 143 ...
 $ Treatment: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
 $ Rat      : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 2 2 2 2 ...
 $ Liver    : Factor w/ 3 levels "1","2","3": 1 1 2 2 3 3 1 1 2 2 ...
 $ rr       : Factor w/ 6 levels "A:1","A:2","B:1",..: 1 1 1 1 1 1 2 2
 $ ll       : Factor w/ 18 levels "A:1:1","A:1:2",..: 1 1 2 2 3 3 4 4 5

> head(rats)

  Glycogen Treatment Rat Liver  rr    ll
1      131         A   1     1 A:1 A:1:1
2      130         A   1     1 A:1 A:1:1
3      131         A   1     2 A:1 A:1:2
4      125         A   1     2 A:1 A:1:2
5      136         A   1     3 A:1 A:1:3
6      142         A   1     3 A:1 A:1:3
```

Rat data plot



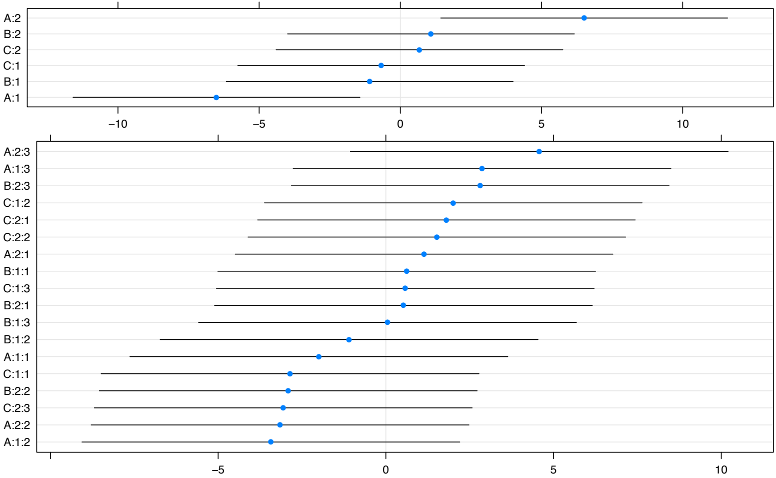
Model with nested random effects

```
> (rats.lmer <- lmer(Glycogen ~ Treatment +(1|rr) +(1|ll), rats)

Linear mixed model fit by REML
Formula: Glycogen ~ Treatment + (1 | rr) + (1 | ll)
   Data: rats
   AIC   BIC logLik deviance REMLdev
 231.6 241.1 -109.8    234.3   219.6
Random effects:
 Groups   Name        Variance Std.Dev.
 ll       (Intercept) 14.167   3.7639
 rr       (Intercept) 36.065   6.0054
 Residual             21.167   4.6007
Number of obs: 36, groups: ll, 18; rr, 6
Fixed effects:
            Estimate Std. Error t value
(Intercept)  140.500      4.707  29.850
TreatmentB    10.500      6.656   1.577
TreatmentC    -5.333      6.656  -0.801

Correlation of Fixed Effects:
           (Intr) TrtmnB
TreatmentB -0.707
```

What about p-values?

- `lmer` does not calculate p-values for the fixed effects coefficients
- For technical reasons, in general computing a p-value for $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ is not always straightforward
- The "t value" in the output does not always have a Student's t distribution under the null
- p-values are "exact" for small, balanced datasets, but not for unbalanced data
- When the number of groups and observations are large, you can consider the "t value" as having a standard normal distribution
- Use the convention that a coefficient is "significant" if $|t| > 2$

Random effects from model rats.lmer



Comments

- There does not seem to be a signficant `Treatment` effect, apparently because the two rats who got treatment A had very different levels of glycogen
- There is also considerable section to section (`Liver`) variability within rat
- Even within the same `Liver` section for the same `Rat` there is variability (especially for rat B:1)