Linear mixed effects models in R

Or… Hierarchical statistical models in R

Or… Multilevel statistical models in R

A workshop/discussion group for the Center for Population Biology (CPB) UC Davis

Scott Burgess
(scburgess@ucdavis.edu)
Center for Population Biology Postdoctoral Fellow

April 24 – 25, 2014

A few things first…

- I am *not* an expert on mixed models!
- I am a field biologist who uses mixed models as a tool to make quantitative inferences about natural systems

"Far better an approximate answer to the right question,… than the exact answer to the wrong question, …." - Tukey

A few things first…

- LMM's, and especially GLMM's, are hard! There's *way* more to them than we can cover here.
- Mixed models are still an active area of research and there is often no one right answer; requires your judgment

"The greatest challenge of hierarchical models is pedagogical: how can ecologists learn to use the power of hierarchical models without getting themselves into trouble" (Bolker, 2009, Ecol Apps 19: p589)

"Despite many ecologists' desire for a completely objective way to make inferences from data, the new world of hierarchical models will require a lot of judgment calls" (Bolker 2009 Ecol Apps 19: p591)

From here

A few things first…

- Often used in the field of social and environmental science

➔ large datasets, observational studies
(thousand of data points to estimate parameters and variances)

- "There are no free lunches": small, noisy datasets (e.g., ecological experiments that we are used to) are not particularly well suited to the estimation procedures used in mixed models

- Some classic rules of thumb (that we often bend):
  - 10 data points per experimental unit
  - 10-20 data points per estimated parameter
  - 5-6 groups (random effect groups) to estimate a variance

(from Bolker 2009 Ecol Apps 19: 571-574 and references therein)

Goal of workshop

- Intuitive understanding of mixed models
- Awareness of the issues and current state of the field
- Produce good graphical summaries and quantitative statements of effects
- Pragmatically linking data analysis in R to writing results up for publication

- Aimed at a beginner to intermediate level of understanding
- Focused on making inferences from the fixed effects, rather than the random effects
- Focused on using `lme` or `lmer`

- What this workshop will *not* cover:
  - The underlying mathematics of estimating parameters (e.g., Likelihood vs MCMC; Frequentist vs Bayesian inference)
  - More advanced topics like MCMCglmm, overdispersion, the animal model
  - Multivariate LMM's, overdispersion in GLMM's, modeling heteroscadisticity, autocorrelation, weights.

Resources

<u>Books</u>

Venables, W.N. and Ripley, B.D. (2002) Modern applied statistics with S. Springer. 4th Edition.

Pinheiro, J.C. and Bates, D.M. (2000) Mixed-Effects Models in S and S-Plus. Springer

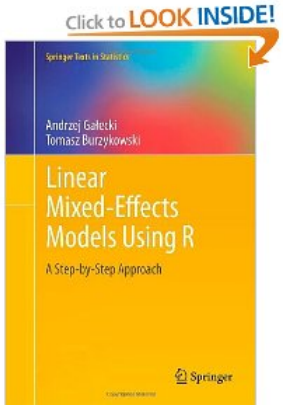
Gelman, A. and Hill, J. (2007) Data analysis using regression and multilevel/ Hierarchical Models. Cambridge University Press

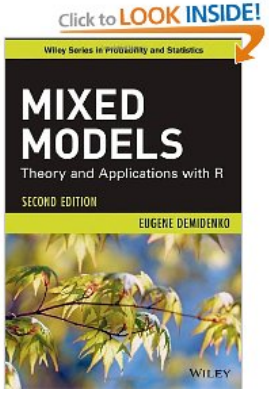
Zurr, A. *et al.* (2009) Mixed effects models and extensions in ecology with R. Springer

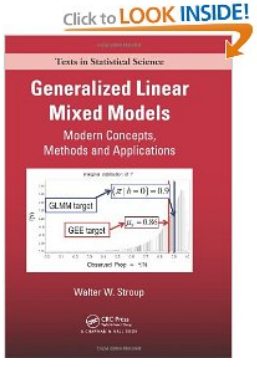
<u>Papers</u>

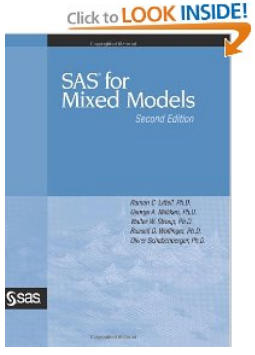
Forum piece in Ecological Applications April 2009, Vol 19(3)

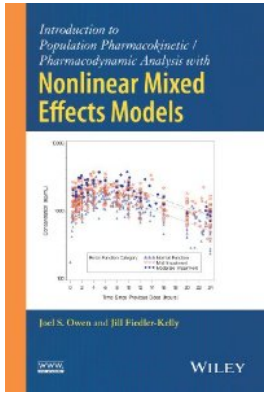
Bolker *et al.* (2008) Generalized linear mixed models: a practical guide for ecology and evolution. TREE 24: 127 – 135.

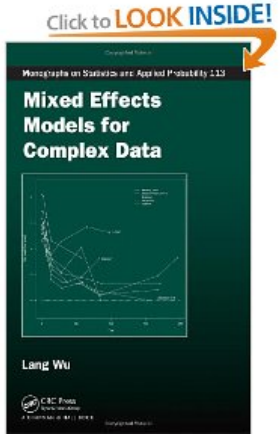








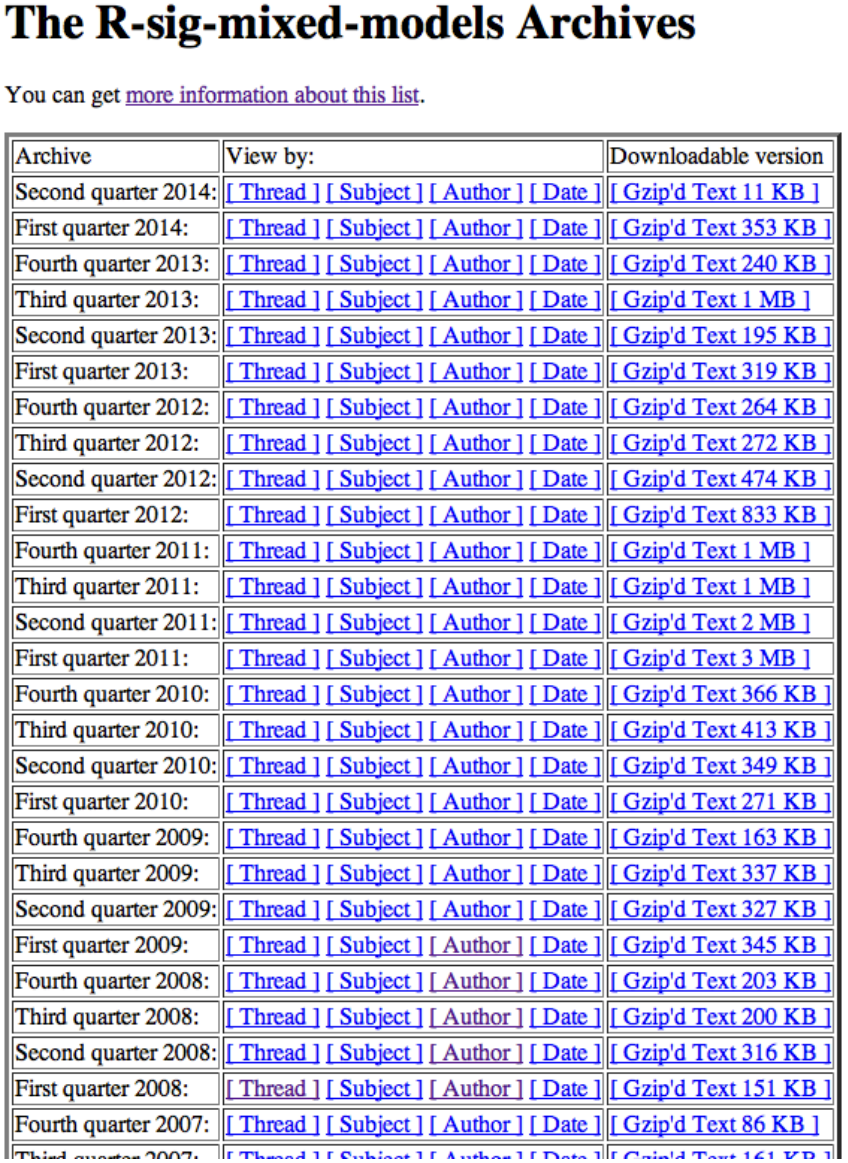




Resources

Websites

- http://glmm.wikidot.com
- R-sig-mixed-mixed-models mailing list
- Richard McElreath's website
- Bates' lme4 book: http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf



The R-sig-mixed-models Archives

You can get more information about this list.

Archive	View by:	Downloadable version
Second quarter 2014:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 11 KB]
First quarter 2014:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 353 KB]
Fourth quarter 2013:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 240 KB]
Third quarter 2013:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 1 MB]
Second quarter 2013:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 195 KB]
First quarter 2013:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 319 KB]
Fourth quarter 2012:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 264 KB]
Third quarter 2012:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 272 KB]
Second quarter 2012:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 474 KB]
First quarter 2012:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 833 KB]
Fourth quarter 2011:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 1 MB]
Third quarter 2011:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 1 MB]
Second quarter 2011:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 2 MB]
First quarter 2011:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 3 MB]
Fourth quarter 2010:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 366 KB]
Third quarter 2010:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 413 KB]
Second quarter 2010:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 349 KB]
First quarter 2010:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 271 KB]
Fourth quarter 2009:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 163 KB]
Third quarter 2009:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 337 KB]
Second quarter 2009:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 327 KB]
First quarter 2009:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 345 KB]
Fourth quarter 2008:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 203 KB]
Third quarter 2008:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 200 KB]
Second quarter 2008:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 316 KB]
First quarter 2008:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 151 KB]
Fourth quarter 2007:	[Thread] [Subject] [Author] [Date]	[Gzip'd Text 86 KB]

Package comparison

There are many more packages, check here: http://glmm.wikidot.com/pkg-comparison

Package >install.packages("pkgs")	Function >library(package)	Estimation	Poisson and Binomial data (i.e., GLMM)?	Random effects structure	Residual structure
nlme	`lme`	ML, REML	No	Good for nested, not so good for crossed	Spatial and temporal correlations (`corStruct`); heteroscedasticity (`varStruct`)
lme4	`lmer`	ML, REML	Yes	Good for nested and crossed	No
MCMCglmm	`MCMCglmm`	Markov chain Monte Carlo (Bayesian)	Yes	Multiple, complex, more flexibility	Temporal autocorrelation of residuals not that flexible

What are mixed-effects models?

- A statistical model containing fixed and random effects
- A generalized statistical model where parameters (e.g., intercepts and slopes) are allowed to vary by group

- Fixed effects influence the *mean* of *y* (constants or parameters).
- Random effects influence the *variance* of *y*. Factor levels come from a population of levels (not interested in parameter values, only in the variance around parameter values explained by the random effect groups)

- Are particularly useful where there is temporal 'pseudoreplication' (repeated measurements) and/or spatial 'pseudoreplication' (nested designs or split plot)

- Main advantage is that they economize on the number of degrees of freedom used up by the factor levels

```
dat[with(dat,order(groups)),]
   groups         x           y
        a  0.000000 -100.370141
        a  3.571429  -52.978398
        a  7.142857  141.927483
        a 10.714286   37.076238
        a 14.285714  118.122216
        a 17.857143  114.706444
        a 21.428571  183.959509
6       a 25.000000  200.190194
1       a 28.571429  375.440097
6       a 32.142857  488.173032
1       a 35.714286  375.296786
6       a 39.285714  367.874960
1       a 42.857143  502.419590
6       a 46.428571  723.041258
1       a 50.000000  589.825270
        b  0.000000   -9.579192
        b  3.571429  258.615224
        b  7.142857  100.105058
        b 10.714286  183.136116
        b 14.285714  134.839112
        b 17.857143  439.993519
        b 21.428571  273.093076
7       b 25.000000  446.706473
2       b 28.571429  442.785043
7       b 32.142857  676.799413
2       b 35.714286  539.771541
7       b 39.285714  677.579336
2       b 42.857143  829.718075
7       b 46.428571  792.869764
2       b 50.000000  842.998033
        c  0.000000   38.302208
        c  3.571429   26.995049
        c  7.142857  226.266013
        c 10.714286  110.042512
        c 14.285714  228.556246
        c 17.857143  165.777985
```

How could we analyze data with grouping?

<u>Complete pooling</u>

- Ignore grouping structure and fit regression: $y_{all} = \beta_0 + \beta_1 x + \varepsilon$ ("pseudoreplication"; plus, how to interpret residuals?)

<u>No pooling</u>

- Estimate separate models within each group, essentially including groups as a fixed effect:

$y_{ij} = \beta_{0j=a} + \beta_{1j=a}x_{j=a} + \beta_{3j=b} + \beta_{4j=b}x_{j=b} + \ldots + \varepsilon$

- Many more parameters, lower power
- Overstates variation among groups (i.e., overfits the data within each group). Makes groups look more different than they really are

```
dat[with(dat,order(groups)),]
   groups         x           y
        a  0.000000 -100.370141
        a  3.571429  -52.978398
        a  7.142857  141.927483
        a 10.714286   37.076238
        a 14.285714  118.122216
        a 17.857143  114.706444
        a 21.428571  183.959509
6       a 25.000000  200.190194
1       a 28.571429  375.440097
6       a 32.142857  488.173032
1       a 35.714286  375.296786
6       a 39.285714  367.874960
1       a 42.857143  502.419590
6       a 46.428571  723.041258
1       a 50.000000  589.825270
        b  0.000000   -9.579192
        b  3.571429  258.615224
        b  7.142857  100.105058
        b 10.714286  183.136116
        b 14.285714  134.839112
        b 17.857143  439.993519
        b 21.428571  273.093076
7       b 25.000000  446.706473
2       b 28.571429  442.785043
7       b 32.142857  676.799413
2       b 35.714286  539.771541
7       b 39.285714  677.579336
2       b 42.857143  829.718075
7       b 46.428571  792.869764
2       b 50.000000  842.998033
        c  0.000000   38.302208
        c  3.571429   26.995049
        c  7.142857  226.266013
        c 10.714286  110.042512
        c 14.285714  228.556246
        c 17.857143  165.777985
```

How could we analyze data with grouping?

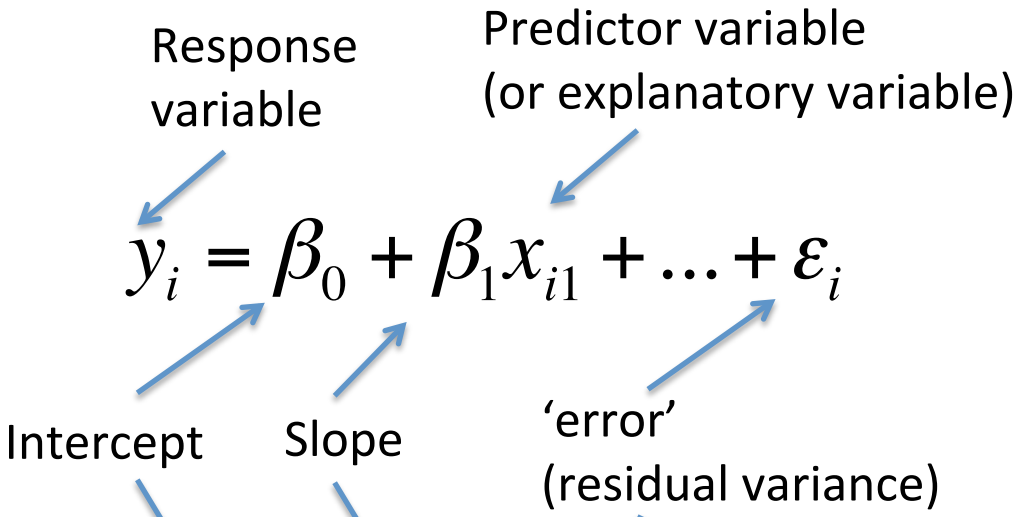
Partial pooling

- ✓ Mixed-effects models
- ✓ A compromise between these extremes
- ✓ Estimates for each group are a weighted average of the mean of the observations within each group (unpooled estimate) and the mean over all groups (pooled estimate)
- ✓ Averages from groups with smaller (larger) sample sizes carry less (more) information, and the weighting 'pulls' the multilevel estimate closer to the overall mean

$$\hat{\beta}_j \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\beta^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\beta^2}}$$

Allowing regression coefficients to vary across groups

We can express a mixed model by starting with a classic linear model:

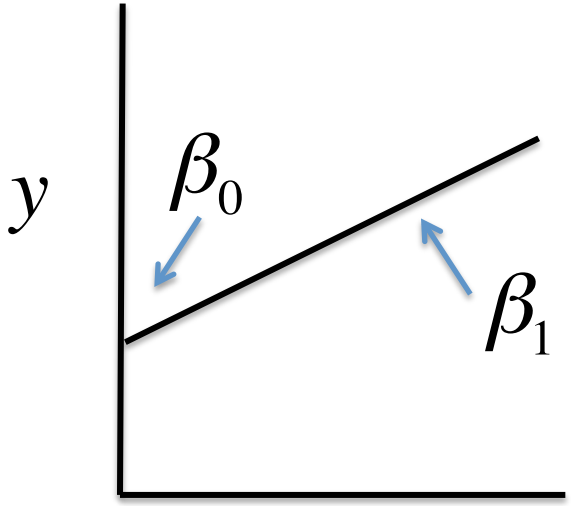


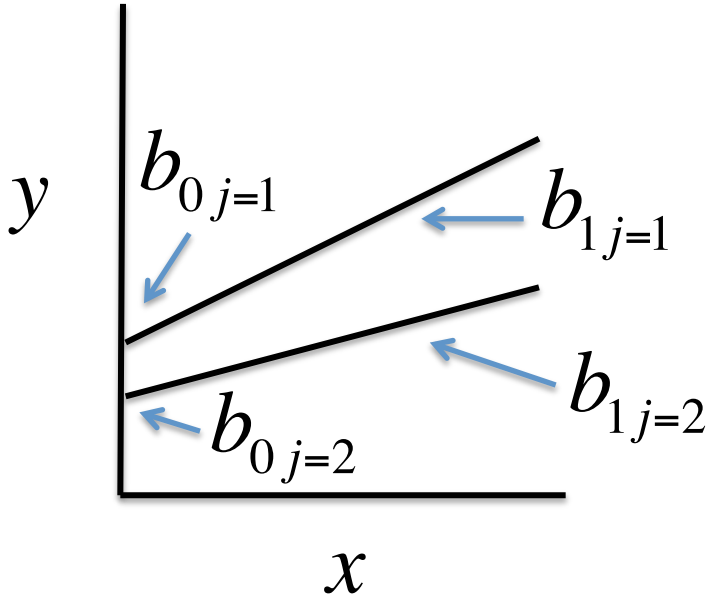
By generalizing to allow coefficients to vary across groups, a linear mixed model is then:

$$y_i = \beta_{0\,j[i]} + \beta_{1\,j[i]} x_{i1} + \ldots + \varepsilon_i$$

*i*th observation, or individual, in the *j*th group

Allowing regression coefficients to vary across groups





$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \varepsilon_i$$

Intercept

Slope

'error'
(residual variance)

$$y_i = \beta_{0\,j[i]} + \beta_{1\,j[i]} x_{i1} + \ldots + \varepsilon_i$$

*i*th observation, or individual,
in the *j*th group

Varying intercepts model

Let's start with a model that has intercepts that vary among the groups, but all groups have the same slope. We can express such a model in multiple ways:

$$y_i \sim \mathrm{N}\left(X_i\beta + \eta_{j[i]}, \sigma_y^2\right)$$

$$\eta_j \sim \mathrm{N}\left(0, \sigma_{\beta_0}^2\right)$$

```
> m1 <- lmer(y ~ x + (1|groups), data=dat)
> summary(m1)
Linear mixed model fit by REML
Formula: y ~ x + (1 | groups)
   Data: dat
   AIC BIC logLik deviance REMLdev
 117.8 119  -54.9    123.8   109.8
Random effects:
 Groups   Name        Variance Std.Dev.
 groups   (Intercept) 18609    136.42
 Residual             11089    105.30
Number of obs: 10, groups: groups, 2

Fixed effects:
            Estimate Std. Error t value
(Intercept)  353.982    112.380   3.150
x             15.049      1.884   7.989

Correlation of Fixed Effects:
  (Intr)
x -0.419
```

Another way to write the same model:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

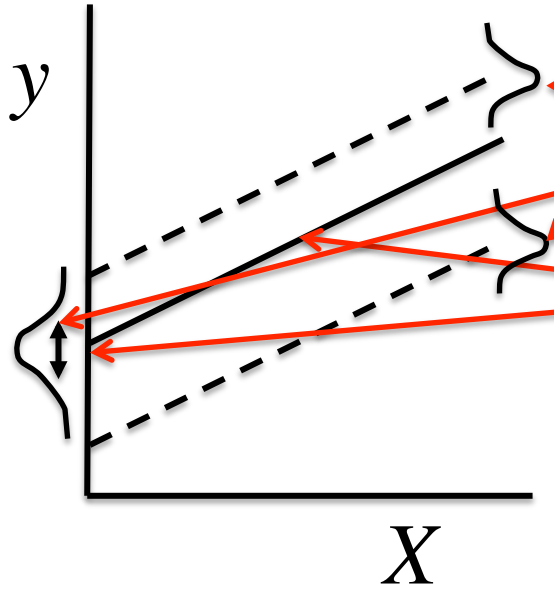
$$b_i \sim N\left(0, \sigma_b^2\right), \quad \varepsilon \sim N\left(0, \sigma^2 I\right)$$

Varying intercepts model

Let's start with a model that has intercepts that vary among the groups, but all groups have the same slope. We can express such a model in multiple ways:

$$y_i \sim \mathrm{N}\left(X_i\beta + \eta_{j[i]}, \sigma_y^2\right)$$

$$\eta_j \sim \mathrm{N}\left(0, \sigma_{\beta_0}^2\right)$$




```
> m1 <- lmer(y ~ x + (1|groups), data=dat)
> summary(m1)
Linear mixed model fit by REML
Formula: y ~ x + (1 | groups)
   Data: dat
   AIC BIC logLik deviance REMLdev
 117.8 119  -54.9    123.8   109.8
Random effects:
 Groups   Name        Variance Std.Dev.
 groups   (Intercept) 18609    136.42
 Residual             11089    105.30
Number of obs: 10, groups: groups, 2

Fixed effects:
            Estimate Std. Error t value
(Intercept) 353.982    112.380   3.150
x            15.049      1.884   7.989

Correlation of Fixed Effects:
  (Intr)
x -0.419
```

Varying intercepts model

`lme` vs. `lmer`

(different printout)

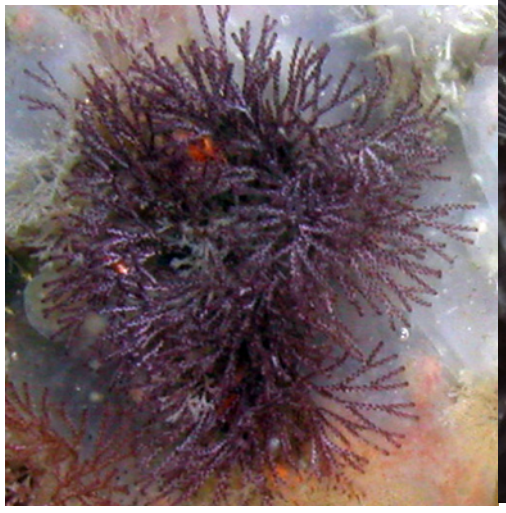
```
> m1 <- lme(y ~ x, random=~1|groups,data=dat)
> summary(m1)
Linear mixed-effects model fit by REML
 Data: dat
       AIC      BIC    logLik
  118.4518 118.7696 -55.2259

Random effects:
 Formula: ~1 | groups
        (Intercept) Residual
StdDev:    692.5044 88.14469

Fixed effects: y ~ x
               Value Std.Error DF   t-value p-value
(Intercept) -359.6502  492.0488  7 -0.730924  0.4886
x             13.6787    1.5768  7  8.675090  0.0001
 Correlation:
  (Intr)
x -0.08

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.0371717 -0.6038736 -0.2863372  0.5221394  1.4038894

Number of Observations: 10
Number of Groups: 2
```

```
> m1 <- lmer(y ~ x + (1|groups), data=dat)
> summary(m1)
Linear mixed model fit by REML
Formula: y ~ x + (1 | groups)
   Data: dat
   AIC   BIC logLik deviance REMLdev
 118.5 119.7 -55.23    127.2   110.5
Random effects:
 Groups   Name        Variance Std.Dev.
 groups   (Intercept) 479561.7 692.504
 Residual               7769.5  88.145
Number of obs: 10, groups: groups, 2

Fixed effects:
            Estimate Std. Error t value
(Intercept) -359.650    492.004  -0.731
x             13.679      1.577   8.675

Correlation of Fixed Effects:
  (Intr)
x -0.080
```

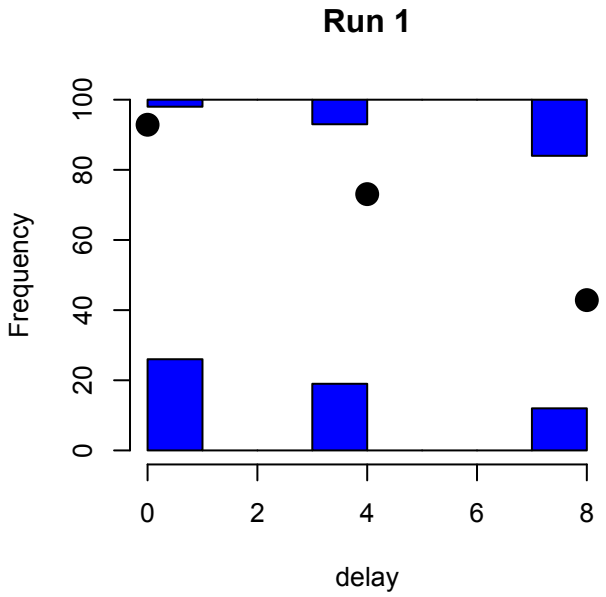
Example 1

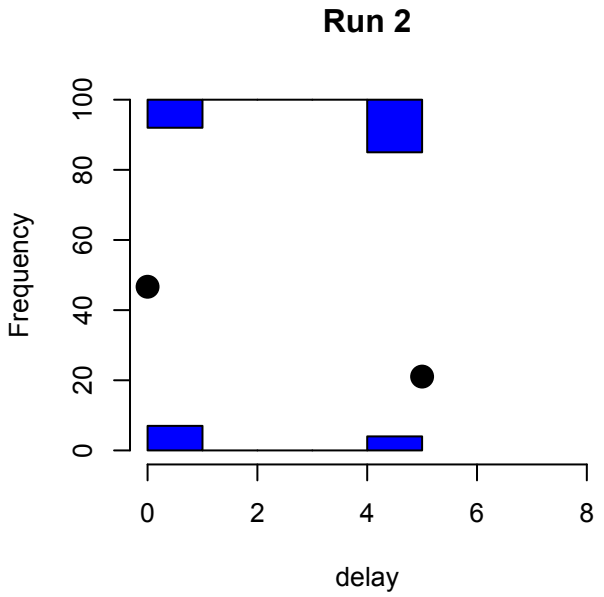
- Delayed the metamorphosis of bryozoan larvae in the lab (explanatory variable, fixed, treated as continuous)
- Measured habitat selection in the laboratory (response variable, binomial 0=bottom or 1=top)
- Repeated the experiment 3 times (random effect)

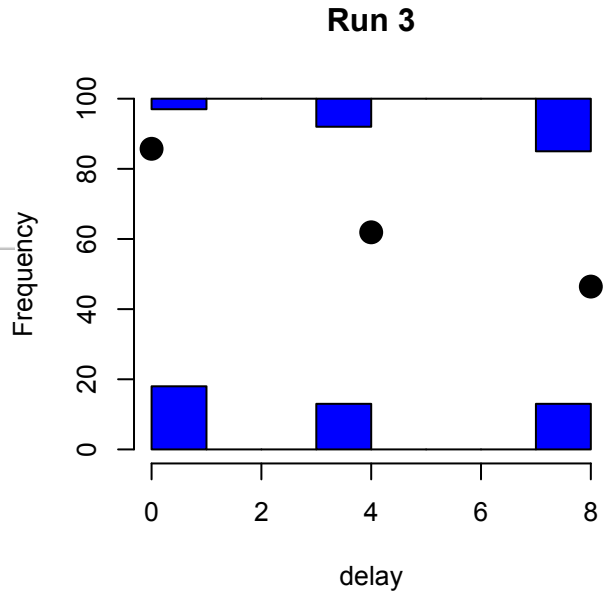
```
> with(habsel,table(run,delay))
    delay
run  0  4  5  8
  1 28 26  0 28
  2 15  0 19  0
  3 21 21  0 28
> with(habsel,table(delay))
delay
 0  4  5  8
64 47 19 56
>
```

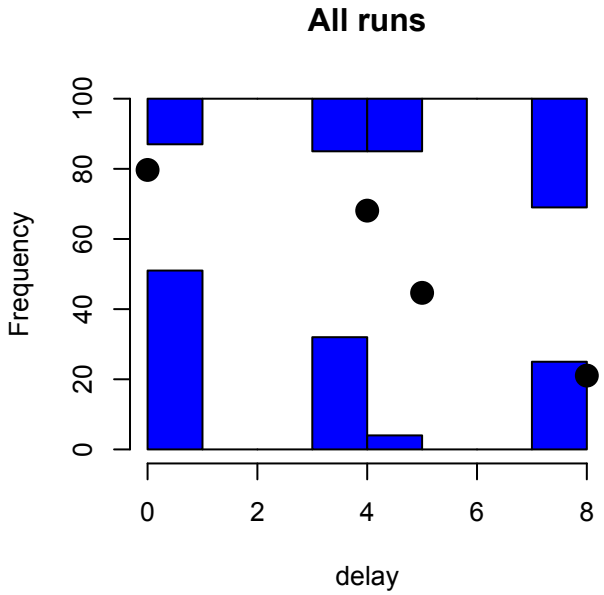


```
> with(habsel,table(run,delay))
   delay
run  0  4  5  8
  1 28 26  0 28
  2 15  0 19  0
  3 21 21  0 28
> with(habsel,table(delay))
delay
 0  4  5  8
64 47 19 56
>
```



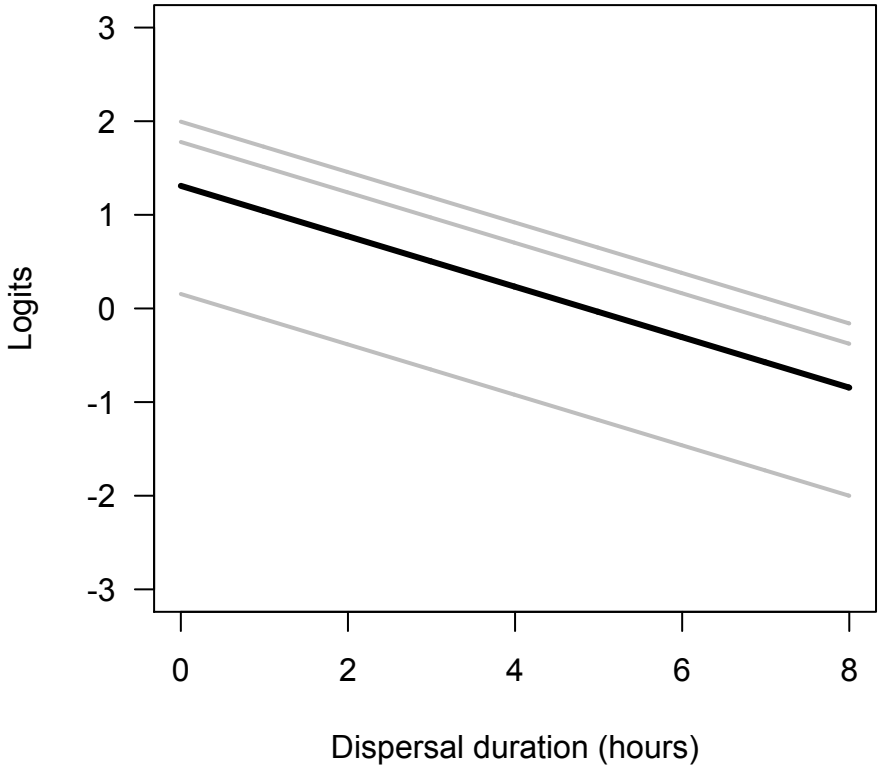






```
> m2 <- lmer(surface ~ delay + (1|run), data=habsel, family=binomial)
> summary(m2)
Generalized linear mixed model fit by the Laplace approximation
Formula: surface ~ delay + (1 | run)
   Data: habsel
   AIC BIC logLik deviance
 224.4 234 -109.2    218.4
Random effects:
 Groups Name        Variance Std.Dev.
 run    (Intercept) 0.76912  0.87699
Number of obs: 186, groups: run, 3

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.30954    0.58578   2.236   0.0254
delay       -0.26938    0.05576  -4.831 1.36e-06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05

Correlation of Fixed Effects:
      (Intr)
delay -0.405
> fixef(m2)
(Intercept)       delay
  1.3095406  -0.2693818
> coef(m2)
$run
  (Intercept)      delay
1   1.9950896 -0.2693818
2   0.1544771 -0.2693818
3   1.7780572 -0.2693818
```

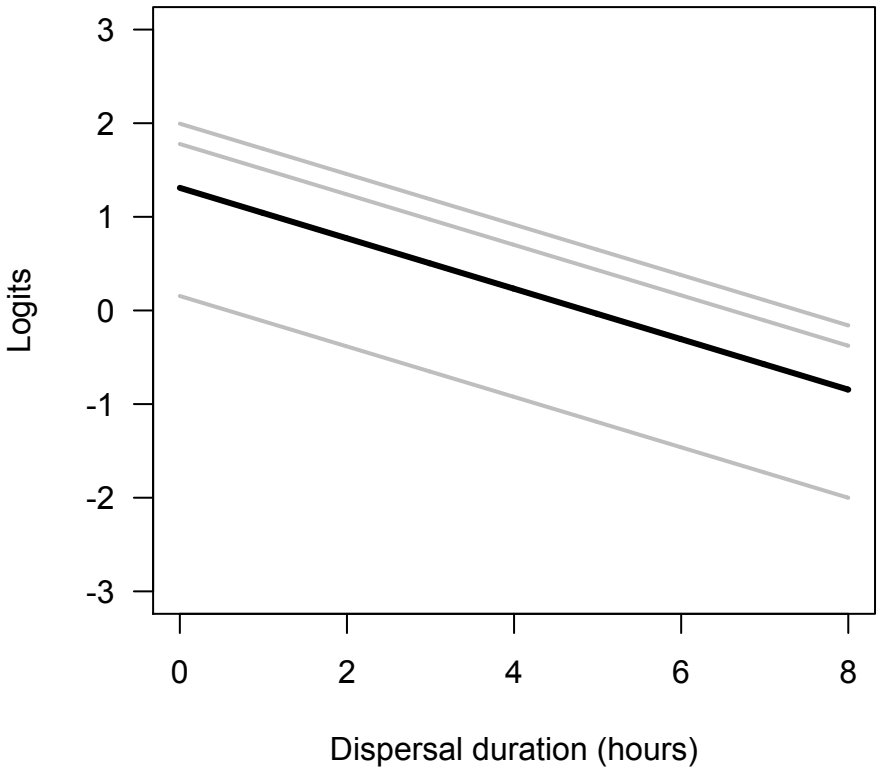
Varying intercepts model

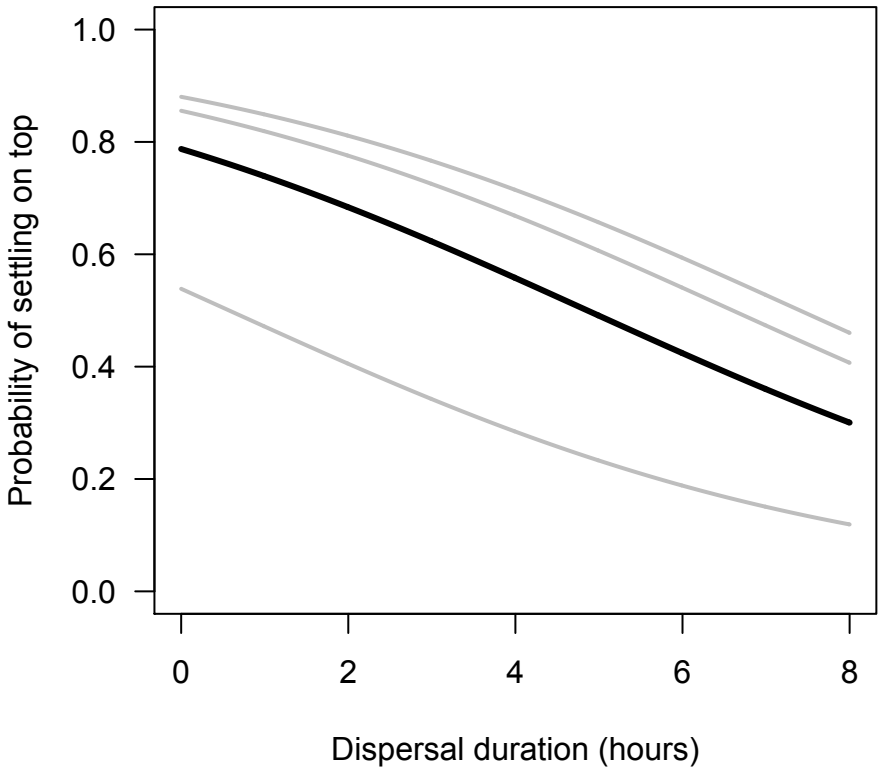


- Use `plogis` to get back to probability scale (between 0 and 1)

```
x = log(p/(1-p))
```

```
p = 1/(1+exp(-x))
```





The results section of a paper

You could say something like:
"For every hour of delay, the *log odds ratio* that larvae settled on the 'top' habitat declined by 0.27 (95% CI = 0.16 – 0.38)."

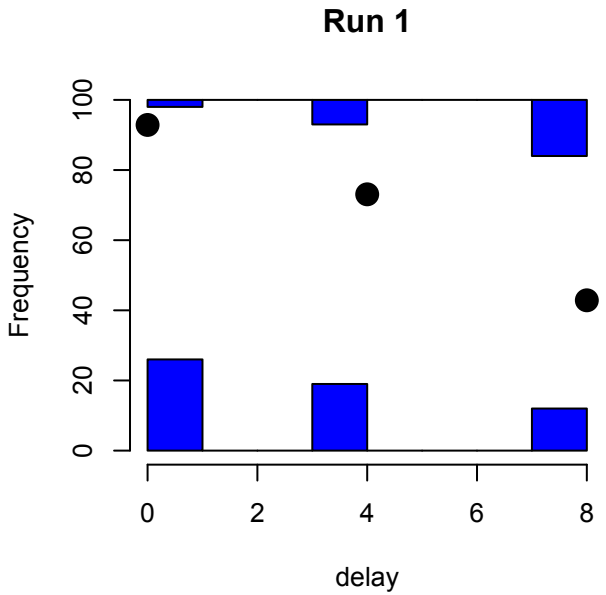
```
> # Estimate of delay parameter for a paper
> fixef(m2)[2] # mean
     delay
-0.2693633
> fixef(m2)[2] + c(-1.96,1.96) * 0.05576 # 95% CI
[1] -0.3786529 -0.1600737
> confint(m2)[3,]
Computing profile confidence intervals ...
     2.5 %     97.5 %
-0.3861103 -0.1612896
```

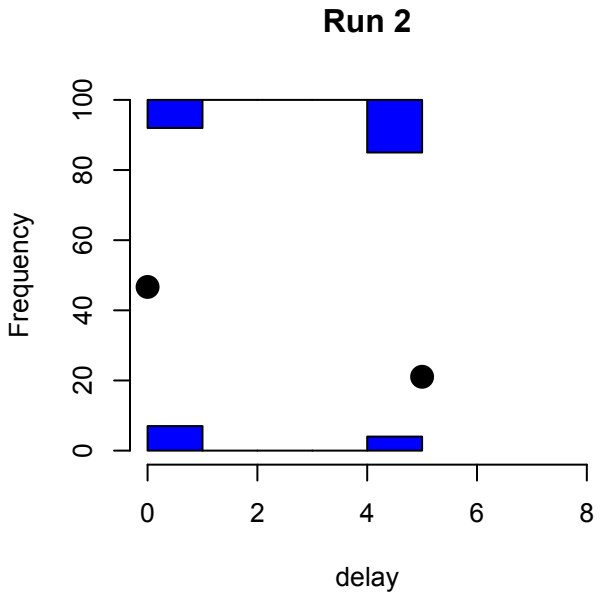
If this was a Gaussian LMM, the above would sound better, because the response variable is less abstract (log odds ratio, versus say, body size)

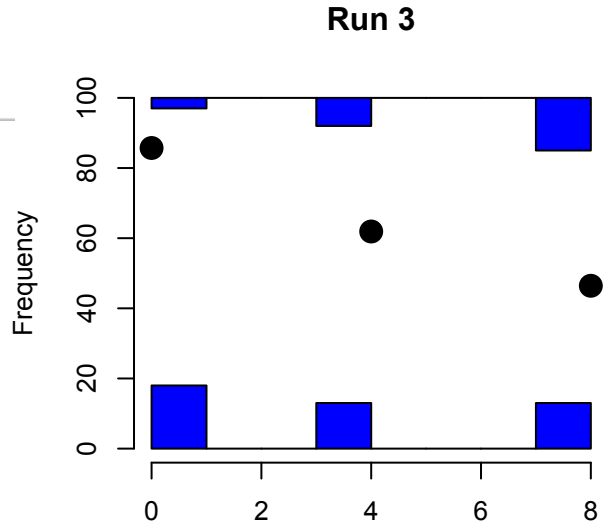
What I actually said in the paper:
"The probability that larvae settled on good surfaces was negatively related to dispersal duration: older larvae were more likely to settle on poor habitats than younger larvae (slope parameter: 0.27 [95% CI = 0.16 – 0.38], X2=25.91, P<0.001; Fig. 2)."

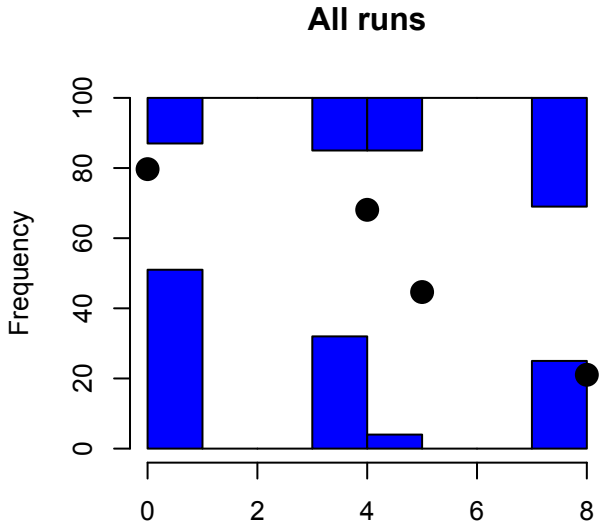
Varying intercepts and slopes model

```
> with(habsel,table(run,delay))
   delay
run  0  4  5  8
  1 28 26  0 28
  2 15  0 19  0
  3 21 21  0 28
> with(habsel,table(delay))
delay
 0  4  5  8
64 47 19 56
>
```



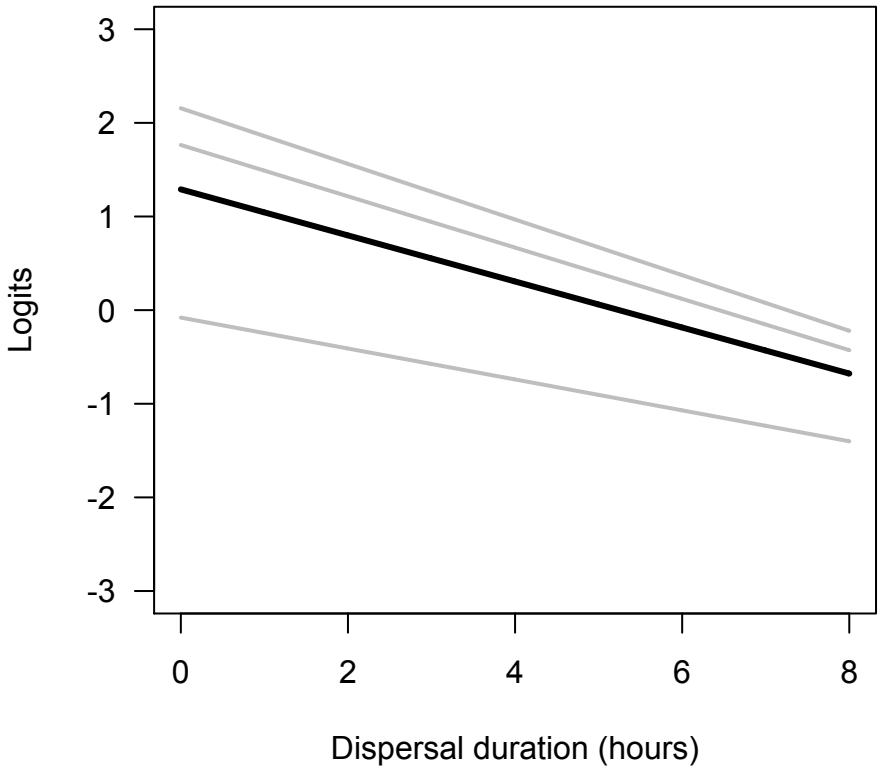






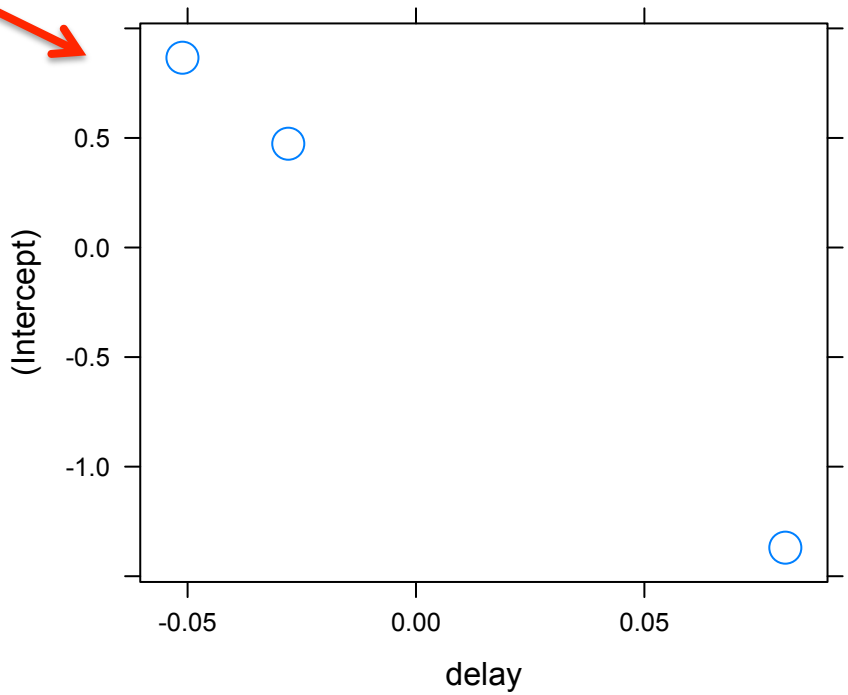
```
> m3 <- lmer(surface ~ delay + (delay|run), data=habsel, family=binomial)
> summary(m3)
Generalized linear mixed model fit by the Laplace approximation
Formula: surface ~ delay + (delay | run)
   Data: habsel
   AIC   BIC logLik deviance
 227.7 243.9 -108.9    217.7
Random effects:
 Groups Name        Variance  Std.Dev. Corr
 run    (Intercept) 1.1064813 1.051894
        delay       0.0038555 0.062093 -1.000
Number of obs: 186, groups: run, 3

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.28992    0.67660   1.906 0.056590 .
delay       -0.24595    0.06552  -3.754 0.000174 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.

Correlation of Fixed Effects:
      (Intr)
delay -0.796
> fixef(m3)
(Intercept)       delay
  1.2899206  -0.2459507
> coef(m3)
$run
  (Intercept)      delay
1  2.15604113 -0.2970774
2 -0.07981339 -0.1650958
3  1.76359778 -0.2739116
```

Varying intercepts and slopes model



```
> m3 <- lmer(surface ~ delay + (delay|run), data=habsel, family=binomial)
> summary(m3)
Generalized linear mixed model fit by the Laplace approximation
Formula: surface ~ delay + (delay | run)
   Data: habsel
   AIC   BIC logLik deviance
 227.7 243.9 -108.9    217.7
Random effects:
 Groups Name        Variance  Std.Dev. Corr
 run    (Intercept) 1.1064813 1.051894
        delay       0.0038555 0.062093 -1.000
Number of obs: 186, groups: run, 3

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.28992    0.67660   1.906 0.056590
delay       -0.24595    0.06552  -3.754 0.000174
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05

Correlation of Fixed Effects:
      (Intr)
delay -0.796
> fixef(m3)
(Intercept)       delay
  1.2899206  -0.2459507
> coef(m3)
$run
  (Intercept)      delay
1  2.15604113 -0.2970774
2 -0.07981339 -0.1650958
3  1.76359778 -0.2739116
```

Varying intercepts and slopes model needs a lot more 'groups' to estimate



Exercise 1

- To load data, type this into R to download it from Dropbox:

```
> install.packages("repmis")
> library(repmis) # then type y when prompted
> FinURL <- paste0("https://dl.dropboxusercontent.com/
u/25547718/Lizards.csv")
> dat <- repmis::source_data(FinURL,sep=",",header=T)
```

Exercise 1

1) Fit the same model using `lmer`, `lme`, and `MCMCglmm` and compare estimates of fixed and random effects. Do they compare?

```
lmer(y ~ x + (1|group), data=dat)
lme(y ~ x, random=~1|group, data=dat)
MCMCglmm(y ~ x, random=~group, data=dat)
```

2) Is there more variation among lizards or within lizards (i.e., between years in each lizard)

3) Write down the equation for the model you fit

4) Fit models with and without random effects and compare estimates of fixed effects

5) Write down a quantitative statement about the results you find

6) Produce a plot of the fitted model

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad i = 1,\ldots,60 \; lizards$$

$$b_i \sim N\left(0, \sigma_b^2\right), \quad \varepsilon \sim N\left(0, \sigma^2 I\right)$$

rbold=1,sex=1
rbold=2,sex=1
rbold=1,sex=2
rbold=2,sex=2

$$X_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Table of means (i.e., $X_i\beta$)

	Sex=1 (Male)	Sex=2 (Female)
rbold=1 (Shy)	β_0	$\beta_0 + \beta_2$
rbold=2 (bold)	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Nested random effects

```
(1|site/block) = (1|site) + (1|site:block)
(1|site/block2) = (1|site) + (1|block2)
```

Where `block2` is a new variable created by:

```
>block2 <- interaction(dat$site,dat$block)
```

- Implicit nesting occurs when the levels of the inner grouping factor (`block`) are reused for different levels of the outer grouping factor (`site`)
- In the example below, block 1 at site A is distinguished from block 1 at site B only if it is assumed that block is nested within site by specifying `(1|site/block)`, or by creating new label `block2`, as above

site	block	block2
A	1	A.1
A	1	A.1
A	2	A.2
A	2	A.2
B	1	B.1
B	1	B.1
B	2	B.2
B	2	B.2

In lmer, each distinct group must correspond to a distinct level in the corresponding grouping factor

Example in R….

Example 2

Load the data…

```
> FinURL <- paste0("https://dl.dropboxusercontent.com/
u/25547718/Nesting_example.csv")
> dat <- repmis::source_data(FinURL,sep=",",header=T)
```

1) Fit four models using `lmer` with different random effects specifications that show the same random effect variances.

2) Fit a model using `lmer` with a random effects specification that you know is incorrect, given the way that site and block are labeled in the data frame.

Small dataset from experiments

- What happens when only have 2 groups?

(some people advise not using mixed models when there is less than 4-6 groups)

➢ An unfortunate 'no man's land' between ANOVA and mixed models

- If design is linear, balanced (same # of observations in each group), and has only a few groups, then use ANOVA (`aov`)?
  - But then restricted to inferences using p-values

- Fit groups (random effects) as fixed effects.
  - Reduced power, but is also likely to be an 'incorrect' model based on the inference you want to make

- If discrete explanatory variables (e.g., Treatment "High" and "Low"), average within each group, and use groups means as 'replicate' in a standard linear model (Example in R)

Hypothesis tests and model selection

…argh!

- One of the main issues in using p-values or Information Criterion (e.g., AIC) in mixed effects models is: *how many degrees of freedom are there? How many parameters are in a mixed model?*

(this was a huge topic on the R-sig-mixed model mailing list around 2006+; for a summary, read the GLMM wiki FAQ)

- Another issue is that the null distributions do not approximate the reference distributions of the test statistic very well when sample sizes are small (i.e., ecological data)

$$AICc = 2k - 2\ln(L) + \frac{2k(k+1)}{n-k-1}$$

k = number of parameters
n = sample size
L = Likelihood

$$\chi^2_{k_2-k_1}$$

$$F_{k_1-1,k_2-1}$$

Hypothesis tests

- How can I test whether a random effect is significant? “Perhaps you shouldn’t” (Bolker, GLMM wiki)

- “It is not appropriate to use statistical significance as a criterion for including particular group indicators” (Gelman and Hill 2007)

- “Using step wise approaches to eliminate non-significant terms is dangerous” (again, this was a huge topic 2006+) (Bolker, GLMM wiki)

- The best way to test hypothesis for fixed effects is not yet completely known (Pinheiro & Bates 2000; Bolker, GLMM wiki)

Hypothesis tests

- P-values from likelihood ratio tests to compare nested models (when models differ in fixed effects, need to use ML, not REML) are approximate and often at least twice as small as they should be (Pinheiro & Bates 2000 p83; Bolker, GLMM wiki)

- For balanced, nested LMM's, conditional F tests, the denominator df are determined by the grouping level at which the term is estimated (Pinheiro & Bates 2000 p91; Bolker, GLMM wiki)

"Because the primary authors of lme4 are not convinced of the utility of the general approach of testing with reference to an approximate null distribution, and because of the overhead of anyone else digging into the code to enable the relevant functionality (as a patch or an add-on), this situation is unlikely to change in the future" (Bolker, GLMM wiki)

Type `?pvalues` into R

http://rwiki.sciviews.org/doku.php?id=guides:lmer-tests

Information Criterion

- AIC's have the same problems as getting sensible p-values (asymptotic test, i.e., large sample approximations)
- How to count the number of parameters depends on the level of focus
- For inference at population level, use marginal AIC, where a single random effect variance counts as 1 degree of freedom
- For inference at the individual level, use conditional AIC, where a the degrees of freedom is a number between 1 and #groups – 1.
- FYI, `lme` and `lmer` count the number of parameters as $q(q+1)/2$, where q is the number of components that have random effect variance

Guidelines

(this is just from the GLMM wiki)

- P-values from balanced, nested LMM designs can be obtained using conditional F-tests i.e., > `anova(m1)`
  - Only approximate for unnalanced, nested LMM
- For glmm's, difference in deviance is only asymptotically *F* or χ^2 distributed. LR test generally not that informative, but more useful for random effects, than fixed effects > `anova(m1,m2)`
- MCMCglmm or parametric bootstrap techniques are the preferred way
- Use confidence intervals on parameter estimates and model predictions (but they also have issues)

Predictions, confidence intervals, and making plots

Often we want to plot the predicted values for the observed responses under the fitted model, or we want to use the fitted model to predict values for new observations

In mixed models, fitted values and predictions may be obtained at different levels, so what level of variation do we show? (No one right answer, and not fully known yet..)

Population level predictions usually of interest for experiments; these predictions estimate the the marginal expected value of the response.

Predicted values at the jth level of nesting estimate the conditional expectation of the response, given the random effects at 'higher' levels of grouping.

Predictions, confidence intervals, and making plots

The `predict` function is not available for `lmer`.

Options are:

1) Analytical (see glmm wiki FAQ; 'work in progress')
2) Simulation (see Gelman and Hill 2007 book, p272-274)
3) Use `MCMCglmm` to get the intervals for 'free' (using `predict.MCMCglmm`, but currently does not support new data)

$$X\beta \pm 1.96 * \sqrt{X\Psi X^T}$$

- R example…

matrix of model parameters; `fixef(model)`

model matrix; `model.matrix(…)`

Variance-covariance matrix of model; `vcov`

One way to deal with overdispersion in a Poisson glmm

- Fit a fully ‘saturated’ model, with a random effect for ‘observation’
- Since a Poisson glmm does not estimate a ‘residual’ variance (because mean = variance in a Poisson distribution), we fit a model that actually estimates the ‘residual’ variance as a random effect, basically modeling the variance explicitly instead of assuming mean=variance
- R example…

Data collected over time

- First thing to ask yourself: Are you interested in the effect of time (i.e., how things vary over time), or just want to account for the effect of sampling multiple times (i.e., repeated measures)?

- R example….