36-463/663: Multilevel & Hierarchical Models

Some Random Effects Configurations

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Random Effect Configurations
  - *<u>Most of our models so far</u>*: Level 1 units (observations) belong to a single set of mutually exclusive Level 2 groups (clusters)
  - *<u>Crossed Random Effects</u>*: Level 1 units can belong to clusters in two (or more) different sets of mutually exclusive Level 2 clusters
  - *<u>Nested Random Effects</u>*: Level 1 units belong to Level 2 clusters, and Level 2 clusters belong to Level 3 clusters (and so on, perhaps…)

Level 1 units belong to a single set of Level 2 clusters

- This is mostly what we've seen with two-level mixed models (only one cluster index *j*)
- For example:

Level 1:

$$y_i = \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \quad , \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Level 2:

$$\begin{aligned} \alpha_{0j} &= \beta_{00} + \beta_{01} u_j + \eta_{0j} \quad && , \ \eta_{0j} \overset{iid}{\sim} N(0, \tau_0^2) \\ \alpha_{1j} &= \beta_{10} + \eta_{1j} && , \ \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2) \end{aligned}$$

```
lmer(y ~ x + u + (1 + x| group))
```


Crossed Random Effects – Level 1 units belong to 2 sets of Level 2 Clusters

Level 1:

$$y_i = \beta_0 + \alpha_{1j[i]} + \alpha_{2k[i]} + \epsilon_i \quad , \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Level 2:

$$\begin{aligned} \alpha_{1j} &= \beta_{10} + \eta_{1j} \quad && , \ \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2) \\ \alpha_{2k} &= \beta_{20} + \eta_{2k} && , \ \eta_{2k} \overset{iid}{\sim} N(0, \tau_2^2) \end{aligned}$$

- Two cluster indices, *j* and *k*
- Can have fixed effects as well
- Note that the intercept is estimated 3 times – the three intercepts are not identifiable! Set $\beta_{10} = \beta_{20} = 0$
- Must think carefully (and some trial and error) about redundant/unidentifiable parameters…

Example: Wheat Yields

- 224 observations of yields from 56 wheat varieties, planted in 4 blocks, in Nebraska*
- Also have latitude and longitude of each plot
- Variables:
  - Block – factor 1, 2, 3, 4
  - Variety – factor ARAPAHOE, BRULE, BUCKSKIN, etc.
  - yield – measured wheat yield in each plot
  - latitude – latitude of each plot
  - longitude – longitude of each plot

*Stroup, W. W., P. S. Baenziger, and D. K. Mulitze. (1994). Removing spatial variation from wheat yield trials: a comparison of methods. *Crop Science* **86**: 62–66.

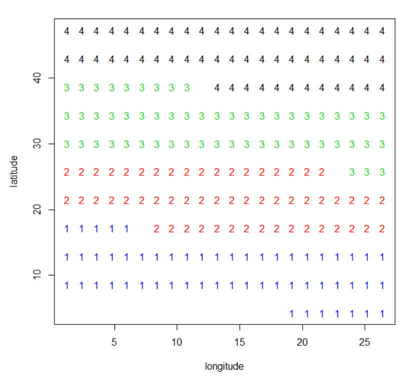
A Little Exploration…

```
library(arm)   # pulls in lme4 too
library(nlme)  # for data sets
data(Wheat2)   # from nlme library

summary(lm(yield ~ Block,
   data=Wheat2))
#           Estimate      SE       t
# (Intcpt) 25.5270   0.4631 55.122
# Block.L    3.2840   0.9262  3.546
# Block.Q   -2.1487   0.9262 -2.320
# Block.C    4.1114   0.9262  4.439

summary(lm(yield ~ variety,
   data=Wheat2))
# only two varieties much different
# from ARAPAHOE (baseline category)
```

```
with(Wheat2, plot(latitude ~
longitude, col=Block,pch=Block))
```




```
# so, Block seems to be trying
# to control for latitude...
```

Fitting and interpreting a crossed random effects model…

```
# try several models with the design
# factors "Block" and "variety" in
# them…
#          yield  ~ ...
# fbv:     Block + variety
# fbvll:   Block + variety +
#             latitude + longitude
# rbv:     (1|Block) + (1|variety)
# rbv.fll: (1|Block) + (1|variety) +
#             latitude + longitude
# rv.fbll: Block + (1|variety) +
#             latitude + longitude
#
#     fbv fbvll  rbv rbv.fll rv.fbll
# AIC 1562 1460 1520    1426    1420
# BIC 1766 1671 1534    1446    1448
#
# AIC likes rv.fbll…
# BIC slightly prefers rbv.fll
```

```
summary(rbv.fll)

# Random effects:
#  Groups    Name     Variance  Std.Dev.
#  variety  (Intcpt)   0.1643    0.4053
#  Block    (Intcpt)   1.3805    1.1750
#  Residual           31.3764    5.6015
# Number of obs: 224,
#        groups:  variety, 56; Block, 4
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept) 24.42114    1.78023  13.718
# latitude    -0.21001    0.05060  -4.151
# longitude    0.48460    0.04966   9.759
#
# latitude and longitude do account for
# a lot of the variation in yield; yet we
# still get a healthy Block random
# effect… lat. & long. aren't everything…
```


The model rbv.fll for the Wheat data:

Level 1:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(latitude)_i + \beta_2(longitude)_i + \\ & \quad \alpha_{1j[i]} + \alpha_{2k[i]} + \epsilon_i \qquad , \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

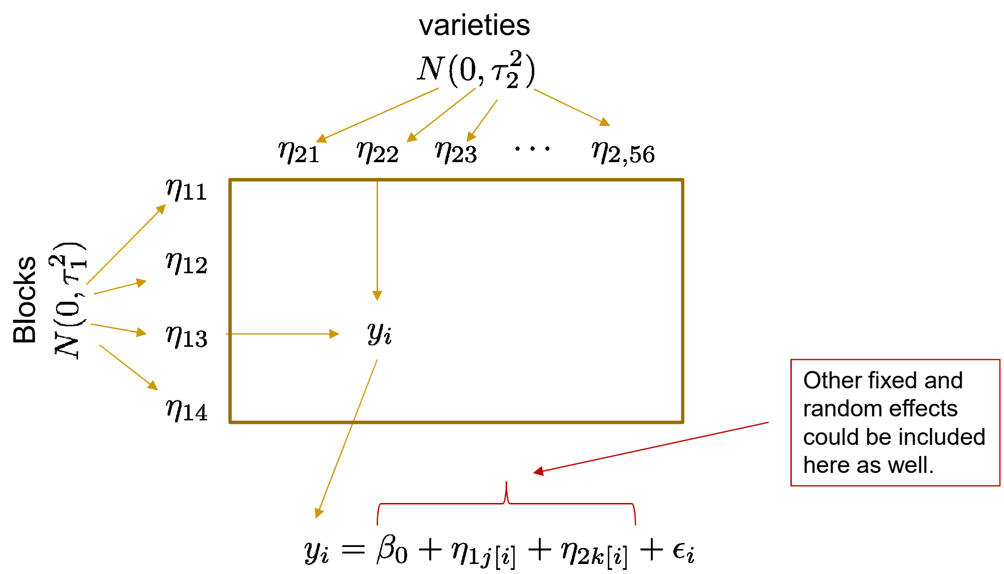
Level 2:

$$\begin{aligned} \alpha_{1j} &= \eta_{1j} \qquad , \ \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2) \\ \alpha_{2k} &= \eta_{2k} \qquad , \ \eta_{2k} \overset{iid}{\sim} N(0, \tau_2^2) \end{aligned}$$

with estimates

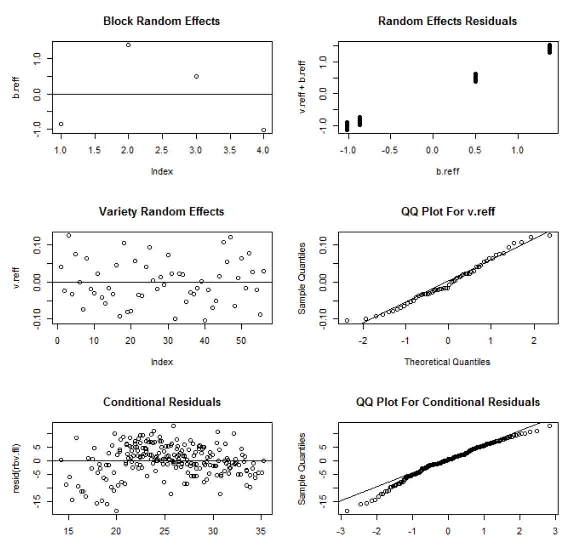
$$\begin{aligned} \hat\beta_0 &= 24.42 & \hat\sigma^2 &= 31.38 \\ \hat\beta_1 &= -0.21 & \hat\tau_1^2 &= 0.16 \\ \hat\beta_2 &= 0.48 & \hat\tau_2^2 &= 1.38 \end{aligned}$$

The generative model for crossed random effects



Graphing the random effects and the residuals…

```
par(mfcol=c(3,2))
b.reff <- ranef(rbv.fll)$Block[[1]]
v.reff <-
  ranef(rbv.fll)$variety[[1]]
plot(b.reff); abline(0,0)
plot(v.reff); abline(0,0)
plot(resid(rbv.fll)~fitted(rbv.fll))
abline(0,0)
vv.reff <- v.reff
b.reff <-
rep(ranef(rbv.fll)$Block[[1]],
  rep(56,4))
v.reff <-
  rep(ranef(rbv.fll)$variety[[1]],4)
plot(v.reff+b.reff ~ b.reff)
qqnorm(vv.reff)
qqline(vv.reff)
qqnorm(resid(rbv.fll))
qqline(resid(rbv.fll))
```



Nested Random Effects – Level 1 units belong to Level 2 clusters, which belong to Level 3 clusters…

- We make a model for some coefficients at level 2, as also being random effects:

Level 1:

$$y_i = \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i \quad , \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Level 2:

$$\alpha_{0j} = \beta_{00k[j]} + \beta_{01} u_j + \eta_{0j} \quad , \eta_{0j} \overset{iid}{\sim} N(0, \tau_0^2)$$

$$\alpha_{1j} = \beta_{10} + \eta_{1j} \quad , \eta_{1j} \overset{iid}{\sim} N(0, \tau_1^2)$$

Level 3:

$$\beta_{00k} = \mu_{000} + \eta_{00k} \quad , \eta_{00k} \overset{iid}{\sim} N(0, \omega_0^2)$$

- Again, two cluster indices (*j* and *k*)
- Can have fixed effects (or not) at any level…

Example: Oxide Layer Thickness in Semiconductors

- We wish to estimate the variance components to determine the assignable causes of the observed variability*.
- 72 observations, 5 variables:
  - Source – one of two sources (1 or 2)
  - Lot – Unique lot identifier (nested within Source)
  - Wafer – Unique wafer I.D. (nested within Lot)
  - Site – factor (1, 2, or 3) indicating site of measurement on wafer (these are just replications on each wafer)
  - Thickness - thickness of the oxide layer on each silicon wafer

*Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, SAS Institute, Cary, NC.

Looking at the data…

```
data(Oxide) # from nlme library
View(Oxide)
   Source Lot Wafer Site Thickness
1       1   1    1    1      2006
2       1   1    1    2      1999
3       1   1    1    3      2007
4       1   1    2    1      1980
5       1   1    2    2      1988
6       1   1    2    3      1982
7       1   1    3    1      2000
8       1   1    3    2      1998
9       1   1    3    3      2007
10      1   2    1    1      1991
11      1   2    1    2      1990
12      1   2    1    3      1988
13      1   2    2    1      1987
14      1   2    2    2      1989
15      1   2    2    3      1988
16      1   2    3    1      1985
17      1   2    3    2      1983
18      1   2    3    3      1989
19      1   3    1    1      2000
20      1   3    1    2      2004
21      1   3    1    3      2004
22      1   3    2    1      2001
23      1   3    2    2      1996
24      1   3    2    3      2004
25      1   3    3    1      1999
26      1   3    3    2      2000
27      1   3    3    3      2002
.
.
.
70      2   8    3    1      1990
71      2   8    3    2      1989
72      2   8    3    3      1992
```


Fitting the model…

```
# wrong model: crossed R.E.s
fit0 <- lmer(Thickness ~
  (1|Lot) + (1|Wafer),
  data=Oxide)
# one way to get right model
Oxide$Wafer.in.Lot <-
  with(Oxide, paste(Lot,
  Wafer,sep="."))
# or easier…
Oxide$Wafer.in.Lot <-
  with(Oxide,
  interaction(Lot,Wafer))
fit1 <- lmer(Thickness ~
  (1|Lot) +
  (1|Wafer.in.Lot),
  data=Oxide)
```

```
# an R equivalent would be:

fit2 <- lmer(Thickness ~
  (1|Lot/Wafer), data=Oxide)
Summary(fit2)

# Random effects:
# Groups    Name     Variance SD
# Wafer:Lot (Intcpt)  35.87   5.99
# Lot       (Intcpt) 129.911 1.40
# Residual            12.57   3.55

# Fixed effects:
#             Estimate   SE
# (Intercept) 2000.15   4.23
```

The model fit2 for the Oxide data...

Level 1:

$$y_i = \alpha_{0j[i]} + \epsilon_i \quad , \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Level 2:

$$\alpha_{0j} = \beta_{00k[j]} + \eta_{0j} \quad , \eta_{0j} \overset{iid}{\sim} N(0, \tau_0^2)$$

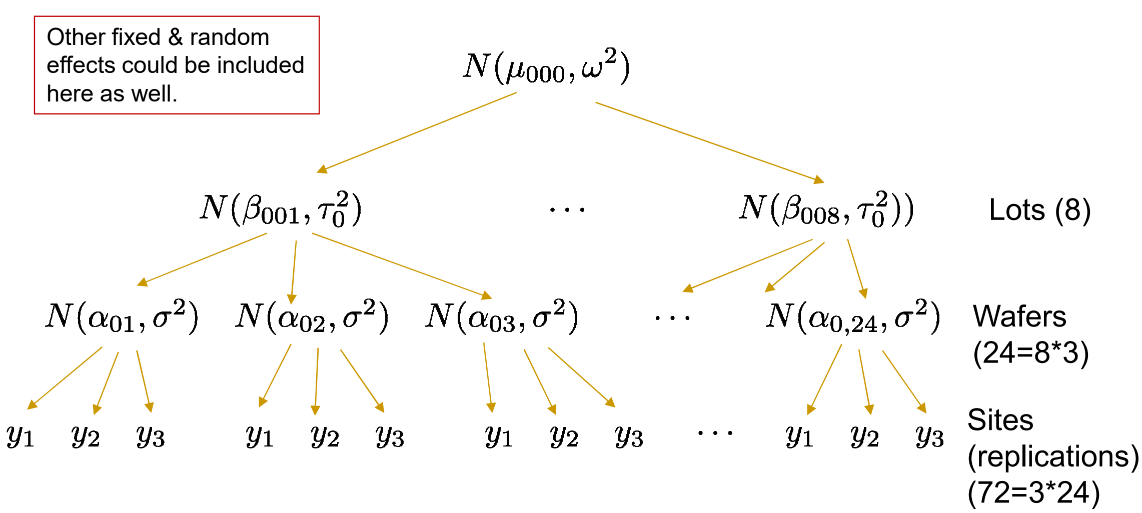
Level 3:

$$\beta_{00k} = \mu_{000} + \eta_{00k} \quad , \eta_{00k} \overset{iid}{\sim} N(0, \omega_0^2)$$

with estimates

$$\hat{\mu}_{000} = 2000.5$$
$$\hat{\sigma}^2 = 12.57$$
$$\hat{\tau}_0^2 = 35.87$$
$$\hat{\omega}_0^2 = 129.91$$

The generative model for nested random effects



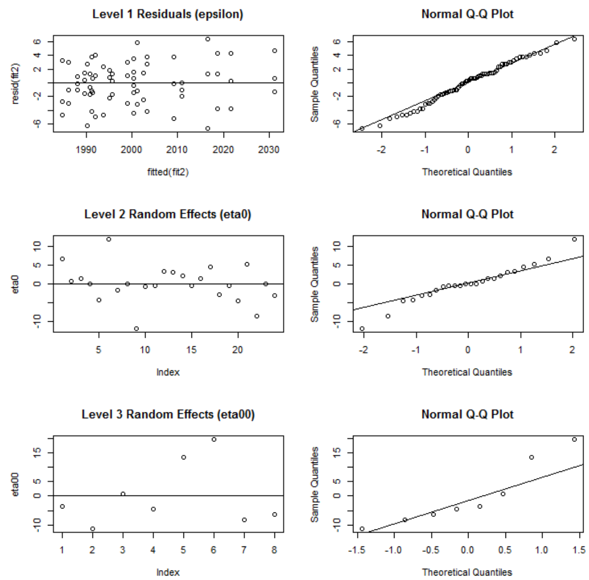
Plotting residuals and random effects for this model…

```
par(mfrow=c(3,2))

plot(resid(fit2) ~ fitted(fit2))
abline(0,0)
qqnorm(resid(fit2))
qqline(resid(fit2))

eta0 <- ranef(fit2)$"Wafer:Lot"[[1]]
plot(eta0)
abline(0,0)
qqnorm(eta0)
qqline(eta0)

eta00 <- ranef(fit2)$Lot[[1]]
plot(eta00)
abline(0,0)
qqnorm(eta00)
qqline(eta00)
```



Outline

- Random Effect Configurations
  - *Most of our models so far*: Level 1 units (observations) belong to a single set of mutually exclusive Level 2 groups (clusters)
  - *Crossed Random Effects*: Level 1 units can belong to clusters in two (or more) different sets of mutually exclusive Level 2 clusters
  - *Nested Random Effects*: Level 1 units belong to Level 2 clusters, and Level 2 clusters belong to Level 3 clusters (and so on, perhaps…)