Random Effects, Mixed Effects and Nested Models[1]

STA442/2101 Fall 2012

[1] See last slide for copyright information.

Background Reading

- Davison's *Statistical models*: Section 9.4
- The White Whale: Chapters 25 and 26

Overview

General Mixed Linear Model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
- $\mathbf{Z}$ is an $n \times q$ matrix of known constants
- $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ unknown but often diagonal
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

Random vs. fixed effects

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- Elements of $\boldsymbol{\beta}$ are called fixed effects.
- Elements of $\mathbf{b}$ are called random effects.
- Models with both are called *mixed*.

Main application

A random factor is one in which the values of the factor are a random sample from a populations of values.

- Randomly select 20 fast food outlets, survey customers in each about quality of the fries. Outlet is a random effects factor with 20 values.
- Randomly select 10 schools, test students at each school. School is a random effects factor with 10 values.
- Randomly select 15 naturopathic medicines for arthritis (there are quite a few), and then randomly assign arthritis patients to try them. Drug is a random effects factor.
- Randomly select 15 lakes. In each lake, measure how clear the water is at 20 randomly chosen points. Lake is a random effects factor.

One random factor

A nice simple example

- Randomly select 6 farms.
- Randomly select 10 cows from each farm, milk them, and record the amount of milk from each one.
- The one random factor is Farm.
- Total $n = 60$

The idea is that "Farm" is a kind of random shock that pushes all the amounts of milk in a particular farm up or down by the same amount.

Farm is a random shock

White Whale Equation 25.38, p. 1047 (almost)

$$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij},$$

where

$\mu_.$ is an unknown constant parameter

$\tau_i \sim N(0, \sigma^2_\tau)$

$\epsilon_{ij} \sim N(0, \sigma^2)$

τ_i and ϵ_{ij} are all independent.

$\sigma^2_\tau \geq 0$ and $\sigma^2 > 0$ are unknown parameters.

$i = 1, \ldots q$ and $j = 1, \ldots, k$

There are $q = 6$ farms and $k = 10$ cows from each farm.

General Mixed Linear Model Notation

$$
\begin{aligned}
Y_{ij} &= \mu_{.} + \tau_i + \epsilon_{ij} \\
\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} \;+\; \mathbf{Z}\mathbf{b} \;+\; \boldsymbol{\epsilon}
\end{aligned}
$$

$$
\begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{1,3} \\ \vdots \\ Y_{5,9} \\ Y_{5,10} \end{pmatrix}
=
\begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} (\mu_{.}) +
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \end{pmatrix}
+
\begin{pmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{1,3} \\ \vdots \\ \epsilon_{5,9} \\ \epsilon_{5,10} \end{pmatrix}
$$

Distribution of $Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$

- $Y_{ij} \sim N(\mu_., \sigma^2_\tau + \sigma^2)$
- $Cov(Y_{ij}, Y_{i,j'}) = \sigma^2_\tau$ for $j \neq j'$
- $Cov(Y_{ij}, Y_{i',j'}) = 0$ for $i \neq i'$
- Observations are not all independent.
- Covariance matrix of $\mathbf{Y}$ is block diagonal: Matrix of matrices
  - Off-diagonal matrices are all zeros
  - Matrices on the diagonal $(k \times k)$ have the *compound symmetry* structure

$$\left(\begin{array}{ccc} \sigma^2 + \sigma^2_\tau & \sigma^2_\tau & \sigma^2_\tau \\ \sigma^2_\tau & \sigma^2 + \sigma^2_\tau & \sigma^2_\tau \\ \sigma^2_\tau & \sigma^2_\tau & \sigma^2 + \sigma^2_\tau \end{array} \right)$$

(Except it's 10×10.)

Statistics based on $Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$

And their distributions

$$\overline{Y}_i \sim N(\mu_., \tfrac{\sigma^2}{k} + \sigma_\tau^2) \qquad Cov(\overline{Y}_i, Y_{ij} - \overline{Y}_i) = 0$$

$$SSTR = k \sum_{i=1}^{q} (\overline{Y}_i - \overline{Y}_.)^2 \qquad \frac{SSTR}{\sigma^2 + k\sigma_\tau^2} \sim \chi^2(q-1)$$

$$SSE = \sum_{i=1}^{q} \sum_{j=1}^{k} (Y_{ij} - \overline{Y}_i)^2 \qquad \frac{SSE}{\sigma^2} \sim \chi^2(qk - q)$$

Last fact is true even though Y_{ij} are not independent, because

$$\overline{Y}_i = \mu_. + \tau_i + \overline{\epsilon}_i$$

Expected mean squares

Since $\frac{SSTR}{\sigma^2+k\sigma_\tau^2} \sim \chi^2(q-1)$, have $E\left(\frac{SSTR}{\sigma^2+k\sigma_\tau^2}\right) = q-1$.

$$
\begin{aligned}
E(MSTR) &= E\left(\frac{SSTR}{q-1}\right) \\
&= E\left(\frac{\sigma^2+k\sigma_\tau^2}{q-1} \times \frac{SSTR}{\sigma^2+k\sigma_\tau^2}\right) \\
&= \frac{\sigma^2+k\sigma_\tau^2}{q-1}\, E\left(\frac{SSTR}{\sigma^2+k\sigma_\tau^2}\right) \\
&= \frac{\sigma^2+k\sigma_\tau^2}{q-1}\, q-1 \\
&= \sigma^2+k\sigma_\tau^2
\end{aligned}
$$

Similarly, $E(MSE) = \sigma^2$.

Components of variance

The proportion of variance in performance explained by the school is

$$Corr(Y_{ij}, Y_{i,j'}) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2}$$

Estimate with

$$\frac{MSTR - MSE}{MSTR + (k-1)MSE}$$

There is a confidence interval based on the F distribution.

Testing $H_0 : \sigma^2_\tau = 0$

$Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$

Is this a strange thing to do?

$SSTR = k\sum_{i=1}^{q}(\overline{Y}_i - \overline{Y}_.)^2$	$\frac{SSTR}{\sigma^2+k\sigma^2_\tau} \sim \chi^2(q-1)$	$E(MSTR) = \sigma^2 + k\sigma^2_\tau$
$SSE = \sum_{i=1}^{q}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_i)^2$	$\frac{SSE}{\sigma^2} \sim \chi^2(qk-q)$	$E(MSE) = \sigma^2$

- $SSTR$ and SSE are independent.
- If H_0 is true, $F^* = \frac{MSTR}{MSE} \sim F(q-1, qk-q)$
- Expected mean squares suggest it will be big when H_0 is false.
- It's the same as the F statistic for a fixed effects model.
- This is the only case where it happens.
- And under H_1, the distribution is *not* a non-central F.

Distribution when $H_0 : \sigma_\tau^2 = 0$ is false

$$\frac{SSTR}{\sigma^2+k\sigma_\tau^2} \sim \chi^2(q-1) \qquad \frac{SSE}{\sigma^2} \sim \chi^2(qk-q)$$

- Note $\frac{MSTR/(\sigma^2+k\sigma_\tau^2)}{MSE/\sigma^2} \sim F(q-1, qk-q)$ whether H_0 is true or not.
- Reject when

$$F^* = \frac{MSTR}{MSE} > f_c$$

$$\Leftrightarrow \quad \frac{MSTR/(\sigma^2+k\sigma_\tau^2)}{MSE/\sigma^2} > \frac{\sigma^2}{\sigma^2+k\sigma_\tau^2}\, f_c$$

Power is a tail area of the *central* F distribution.

Mixed models

The classical approach

- There can be both fixed and random factors in the same experiment
- The interaction of any random factor with another factor (whether fixed or random) is random.
- F-tests are often possible, but they don't always use Mean Squared Error in the denominator of the F statistic.
- Often, it's the Mean Square for some interaction term.
- The choice of what error term to use is relatively mechanical for balanced models with equal sample sizes — based on expected mean squares.

Two-factor random effects: A and B both random

Equal sample sizes: k per cell

Effect	Expected Mean Square	F test
A	$\sigma^2 + kb\sigma_\alpha^2 + k\sigma_{\alpha\beta}^2$	$\frac{MSA}{MSAB}$
B	$\sigma^2 + ka\sigma_\beta^2 + k\sigma_{\alpha\beta}^2$	$\frac{MSB}{MSAB}$
$A \times B$	$\sigma^2 + k\sigma_{\alpha\beta}^2$	$\frac{MSAB}{MSE}$
Error	σ^2	

Mixed model: A fixed and B random

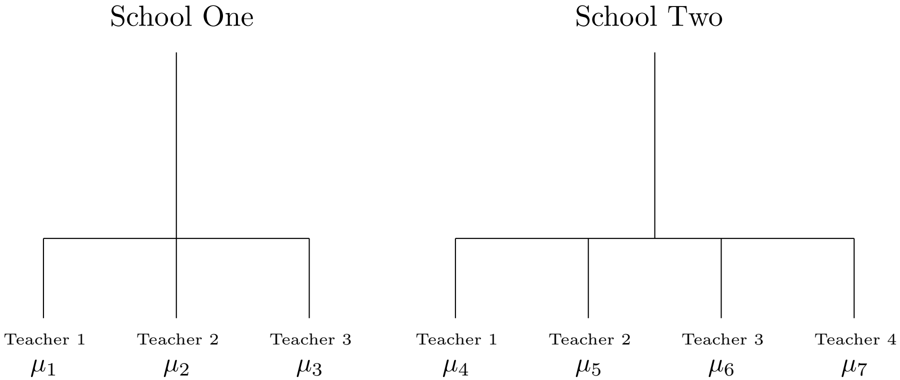
Equal sample sizes: k per cell

Effect	Expected Mean Square	F test
A	$\sigma^2 + kb\frac{\sum \alpha_i^2}{a-1} + k\sigma^2_{\alpha\beta}$	$\frac{MSA}{MSAB}$
B	$\sigma^2 + ka\sigma^2_{\beta}$	$\frac{MSB}{MSE}$
$A \times B$	$\sigma^2 + k\sigma^2_{\alpha\beta}$	$\frac{MSAB}{MSE}$
Error	σ^2	

Extensions to larger designs are natural

- Mean squares are all independent, and multiples of chi-squared (if the design is balanced).
- Look at expected mean squares to see which variance terms will cancel in numerator and denominator under H_0.
- Calculation of expected mean squares can be automated.
- Extends to nested designs.

A Nested design with fixed effects



Schools $H_0 : \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \frac{1}{4}(\mu_4 + \mu_5 + \mu_6 + \mu_7)$

Teachers within Schools $H_0 : \mu_1 = \mu_2 = \mu_3$ and $\mu_4 = \mu_5 = \mu_6 = \mu_7$

Tests of nested effects are tests of contrasts

You can specify the contrasts yourself, or you can take advantage of `proc glm`'s syntax for nested models.

```
proc glm;
     class school teacher;
     model score = school teacher(school);
```

The notation `teacher(school)` should be read "teacher within school."

Easy to extend the ideas

- Can have more than one level of nesting. You could have climate zones, lakes within climate zones, fishing boats within lakes, ...
- There is no problem with combining nested and factorial structures. You just have to keep track of what's nested within what.
- Factors that are not nested are sometimes called "crossed."

Nesting and random effects

- Nested models are often viewed as random effects models, but there is no necessary connection between the two concepts.
- It depends on how the study was conducted. Were the two schools randomly selected from some population of schools, or did someone just pick those two (maybe because there are just two schools)?
- Random effects, like fixed effects, can either be nested or not; it depends on the logic of the design.

Sub-sampling

- Sub-sampling is an interesting case of nested and purely random effects
- For example, we take a random sample of towns, from each town we select a random sample of households, and from each household we select a random sample of individuals to test, or measure, or question.

Another good example

Of sub-sampling

- We are studying waste water treatment, specifically the porosity of "flocks," nasty little pieces of something floating in the tanks.
- We randomly select a sample of flocks, and then cut each one up into very thin slices. We then randomly select a sample of slices (called "sections") from each flock, look at it under a microscope, and assign a number representing how porous it is (how much empty space there is in a designated region of the section).
- The explanatory variables are flock and section. The research question is whether section is explaining a significant amount of the variance in porosity – because if not, we can use just one section per flock, and save considerable time and expense.

SAS proc nested

SAS `proc nested` is built specifically for pure random effects models with each explanatory variable nested within all the preceding ones.

```
proc sort; by flock section; /* Data must be sorted */
proc nested;
     class flock section;
     var por;
```

You could use `proc glm`, but the `proc nested` syntax is easier and the output is nicer for this special case.

Repeated measures

Sometimes called *within-cases*

- Sometimes an individual is tested under more than one condition, and contributes a response for each value of a categorical explanatory variable.
- One can view "subject" as just another random effects factor, because subjects supposedly were randomly sampled.
- Subject would be nested within sex, but might cross stimulus intensity.
- This is the classical (old fashioned) way to analyze repeated measures.

Sometimes there is a lot of pretending

- Of course lots of the time, nothing is randomly selected – but people use random effects models anyway. Why pretend?
- Sometimes they are thinking that in a better world, lakes *would have been* randomly selected.
- Or sometimes, the scientists are thinking that they really would like to generalize to the entire population of lakes, and therefore should use statistical tools that support such generalization, even if there was no random sampling.
- But remember that no statistical method can compensate for a biased sample.
- Often the scientists are quite aware of this point, but they use random effects models anyway because it's just a tradition in certain sub-areas of research, and everybody expects to see them.

Problems with the classical approach

- Normality matters in a serious way.
- Sometimes (especially for complicated mixed models) a valid F-test for an effect of interest just doesn't exist.
- When sample sizes are unbalanced, everything falls apart.
  - Mean squares are independent of MSE, but not of one another.
  - Chi-squared variables involve matrix inverses, and variance terms no longer cancel in numerator and denominator.
  - What about covariates? Now it gets really complicated.
- Standard large-sample methods are no help.

A modern approach using the general mixed linear model

$$\mathbf{Y} \;=\; \mathbf{X}\boldsymbol{\beta} \;+\; \mathbf{Z}\mathbf{b} \;+\; \boldsymbol{\epsilon}$$

- $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_b\mathbf{Z}' + \sigma^2 I_n)$
- Estimate $\boldsymbol{\beta}$ as usual with $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- Estimate $\boldsymbol{\Sigma}_b$ and σ^2 by "restricted maximum likelihood."

Restricted maximum likelihood

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- Transform $\mathbf{Y}$ by the $q \times n$ matrix $\mathbf{K}$.
- The rows of $\mathbf{K}$ are orthoganal to the columns of $\mathbf{X}$, meaning $\mathbf{KX} = \mathbf{0}$.
- Then

$$\begin{aligned} \mathbf{KY} &= \mathbf{KX}\boldsymbol{\beta} + \mathbf{KZb} + \mathbf{K}\boldsymbol{\epsilon} \\ &= \mathbf{KZb} + \mathbf{K}\boldsymbol{\epsilon} \\ &\sim N(\mathbf{0}, \mathbf{KZ}\boldsymbol{\Sigma}_b\mathbf{Z}'\mathbf{K}' + \sigma^2\mathbf{KK}') \end{aligned}$$

- Estimate $\boldsymbol{\Sigma}_b$ and σ^2 by maximum likelihood.
- A big theorem says the result does not depend on the choice of $\mathbf{K}$.

Nice results from restricted maximum likelihood

- F statistics that correspond to the classical ones for balanced designs.
- For unbalanced designs, "F statistics" that are actually excellent F approximations — not quite F, but very close.
- SAS `proc mixed` does the job well, and has other virtues.
- Like $V(\boldsymbol{\epsilon})$ can be block diagonal, with useful structures …

Copyright Information