Module 2: LMMs

Wednesday, September 6, 2023 13:31

BIOS526_M2_LMMs_...

Module 2: Linear Mixed Models

BIOS 526
Instructor: Benjamin Risk

Reading

- Ruppert, D., M. Wand, R. Carroll, *Semiparametric Regression*. 4.1 - 4.8 (4.9 is also interesting)
- Wood, S. *Generalized Additive Models*. Chapter 2.
- Reference for syntax: Table 2 in Bates et al. (2015), Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. → useful table in it

Concepts

- Mixed models for data that are not independent, e.g., clustered data, repeated measures. e.g. individual with repeated measurements
- Structure and notation for clustered data.
- Random intercept model: motivation and interpretation.
- Shrinkage estimation and BLUPs of random effects.
- Random slope model.
- Hierarchical formulation of random effect model.

Examples of Clustered Data

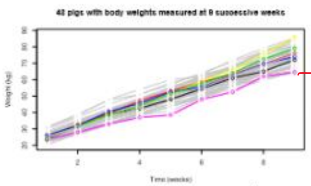
i = subject
j = repeated measurement

1. **Longitudinal Data**: add subscript j
   E.g., observations y_{ij}
   Repeated measurements (level-1), e.g., $j = 1, \dots, r$,
   on each subject (i.e., group, level-2), e.g., $i = 1, \dots, n$.
   - In a sample of students across years, *annual* math score from each student.
   - In a sample of patients, CD4+ cell counts of each HIV patient *visit* past seroconversion.
2. **Multilevel Data**: observations (level-1) nested within groups (level-2).
   - In a sample of students across years from multiple schools, math scores of students from each school.
   - In a sample of multiple time points of medical errors across multiple hospitals, occurrence of medical errors from each hospitals.

Clusters or groups represent a collection of units from a **population** of similar units.

Longitudinal data: Pig Weight

Pigs are measured for 9 successive weeks



Let y_{ij} be the weight (kg) at the j^{th} week for the i^{th} pig.

week 1 … 9

Pig Weight Data Structure

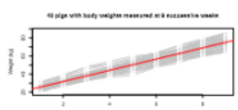
Multilevel data are often represented in the *long* format. Data are grouped by the variable *id*.

as opposed to wide format

```
> pig[1:13,]
   id weeks weight
1   1     1   24.0
2   1     2   32.0
3   1     3   39.0
4   1     4   42.5
5   1     5   48.0
6   1     6   54.5
7   1     7   61.0
8   1     8   65.0
9   1     9   72.0
10  2     1   22.5
11  2     2   30.5
12  2     3   40.5
13  2     4   45.0
```

this is how we'll analyze in R

Approach 1: Incorrect approach ignoring clustered structure



$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij} \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

for the week & pig

Note:

- $\hat{\beta}_1$ is an unbiased and consistent estimator of β_1

Issues:

- Errors are not independent $\rightarrow$ incorrect standard error estimates.
- σ^2 conflates within and between pig variability.

IID is wrong!

incorrectly assumes

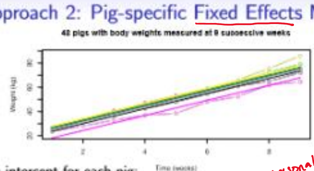
Limitations:

- Cannot forecast individual pig's growth curve.

It's not about the β_1 estimate itself but moreso the inference

separate intercept for each subject

Approach 2: Pig-specific Fixed Effects Model



Separate intercept for each pig:

$$y_{ij} = \boxed{\beta_{0i}} + \beta_1 x_{ij} + \epsilon_{ij} \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

a separate intercept for every pig

reasonable bc we have β_{0i} so remove pig-specific effect

Interpretations:

- β_{0i} is the pig-specific weight at zero and β_1 is the constant slope.
- σ^2 captures within pig variability.

Limitations:

- Estimating lots of parameters: subject-specific coefficients don't leverage population information and have less precision because of smaller sample size.
- Cannot forecast the growth curve of a new pig.

Pig Data: Fit Comparison

```
> fit.lm = lm (weight~weeks, data = dat)
> summary(fit.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.35561    0.46054   42.03   <2e-16 ***
weeks        6.20990    0.08184   75.88   <2e-16 ***

Residual standard error: 4.392 on 430 degrees of freedom
Multiple R-squared: 0.9305,     Adjusted R-squared: 0.9303
F-statistic:  5757 on 1 and 430 DF,  p-value: < 2.2e-16

> fit.strat = lm (weight~weeks+factor(id)-1, data = dat)
> summary(fit.strat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
weeks         6.20990    0.03906  158.97   <2e-16 ***
factor(id)1  17.61719    0.72557   24.28   <2e-16 ***
factor(id)2  20.28385    0.72557   27.96   <2e-16 ***
.
factor(id)48 25.67274    0.72557   35.38   <2e-16 ***
---
Residual standard error: 2.096 on 383 degrees of freedom
Multiple R-squared: 0.9859,     Adjusted R-squared: 0.9841
F-statistic: 557.8 on 48 and 383 DF,  p-value: < 2.2e-16
```

↳ don't include overall intercept

Here, SE of weeks: SE changes (↓) which helps us

Also $\hat{\beta}_{weeks}$ is the same (here, all data are balanced — 9 obs per pig)

Pig Data: Summary

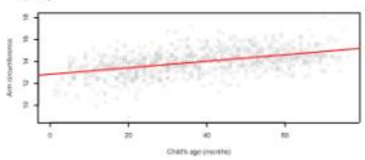
- Data are balanced (same number of observations for each pig).
- Here, the slope of week is the same in the model with a single intercept and the model with an intercept for each pig.
- Here, controlling for group-specific intercepts gives a smaller standard error for the slope of weeks.
- Note that oftentimes, the standard error will be larger. Pseudo-replication = treating clustered observations as independent.

↳ Here, the week covariate ≠ the same for all pigs. Turns out that the SE ↓

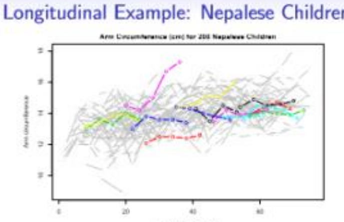
Longitudinal Example: Nepalese Children

Study Design:

1. Time-varying variables of 200 children collected at (up to) 5 time points about 4 months apart:
   - age (month), indicator for current breastfeeding status, arm circumference (cm), height (cm), weight (kg).
2. Time-invariant baseline information:
   - sex of the child, mother's age at birth, indicator of mother's literacy, parity.



Longitudinal Example: Nepalese Children



Scientific questions about arm circumference and age:

- What is the overall trend?
- How much do growth patterns differ between children?
- Do maternal covariates explain variability in growth patterns between children?
- How do we predict the growth pattern of a new child?

Nepalese Data: Fit Comparison

Ignores repeated measures

Incorrectly assumes IID

```
> fit.incorrect = lm (arm~age, data = nepal)
> summary (fit.incorrect)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.838182   0.075987  168.95   <2e-16 ***
age          0.029789   0.001798   16.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9849 on 880 degrees of freedom

> nepal$fid = factor(nepal$id)
> fit.fixedeffects = lm (arm~age+fid, data = nepal)
> summary (fit.fixedeffects)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.626381   0.287480  43.921  < 2e-16 ***
age          0.031354   0.003073  10.204  < 2e-16 ***
fid2        -0.276657   0.354961  -0.779 0.436013
.
fid199       0.950442   0.344728   2.757 0.005988 **
fid200      -2.112362   0.364621  -5.793 1.05e-08 ***
Residual standard error: 0.4972 on 684 degrees of freedom
```

fixed effects model

$Y_{ij} = \beta_{fid_1} + \beta_{fid_j} + \beta_{age} age_{ij}$

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

Fit Comparison

- Note that the effects of age are different.
- Here, controlling for group-specific intercepts gives a larger standard error. (Allows for valid inference.)

$H_0: \beta_{age} = 0$ if H_0 is true, then type I error equal to α

i.e. it's valid to reject H_0 if $p < \alpha$

p-values are usually under uniform distr.

Mixed model: Random Intercept

Consider the random intercept model with a vector of predictors x_{ij}:

$$y_{ij} = \mu + \theta_i + x_{ij}'\beta + \epsilon_{ij}$$

$$\theta_i \overset{iid}{\sim} N(0, \tau^2), \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad \theta_i \perp\!\!\!\perp \epsilon_{ij}$$

is independent ($\perp$ is orthogonal) but here, either works $\perp\!\!\!\perp$ or $\perp$ bc Gaussian?

- μ = overall intercept (grand mean when all $x_{ij} = 0$) (like population mean)
- θ_i = subject-specific difference from μ. aka "random effect"
- $\beta_{0i} = \mu + \theta_i$ = group i's intercept. (or participant i's intercept)
- β is the vector of coefficients that do not vary between groups.
- τ^2 = random effect variance: between-group variability in the intercepts.
- σ^2 = residual variance: within-group variability in the residuals. Measurement error.

Mixed model: θ_i is a random variable. β are fixed.

both random + fixed

Mixed model: Random Intercept

The following two models are *equivalent*:

Model 1: $y_{ij} = (\mu + \theta_i) + x_{ij}'\beta + \epsilon_{ij}, \quad \theta_i \overset{iid}{\sim} N(0, \tau^2) \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

random fixed

Model 1 is often referred to as a mixed model formulation where we assume the random coefficients θ_i have mean zero.

Model 2: $y_{ij} = \beta_{0i} + x_{ij}'\beta + \epsilon_{ij}, \quad \beta_{0i} \overset{iid}{\sim} N(\mu, \tau^2) \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.

still a mixed model

Model 2 is often referred to as a hierarchical model formulation, where the random coefficients β_{0i} have a *higher-level* mean μ. We have a level for y_{ij} and also for β_{0i}

Assumptions:

- $\epsilon_{ij} \perp\!\!\!\perp \theta_i$ (where $\perp\!\!\!\perp$ = independent) for all i and j.
- θ_i are independent Normal for all i. aka $\theta_i \overset{iid}{\sim} N(0, \tau^2)$
- $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

and identically distributed

Properties of the Random Intercept Model

fixed

$$y_{ij} = \mu + \theta_i + \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \epsilon_{ij}, \quad \theta_i \overset{iid}{\sim} N(0,\tau^2) \quad \epsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$$

- Overall (average) trend: Integrate out the random effects
  population
  $$E[\, y_{ij}\,] = \mu + \boldsymbol{x}_{ij}'\boldsymbol{\beta}$$
- Total variability around the overall trend:
  $$Var[\, y_{ij}\,] = \tau^2 + \sigma^2$$
- Conditional (group-specific) trend:
  $$E[\, y_{ij} \mid \theta_i\,] = \mu + \theta_i + \boldsymbol{x}'\boldsymbol{\beta}$$
- Conditional (within-group) residual variance:
  $$Var[\, y_{ij} \mid \theta_i\,] = \sigma^2$$

$$E[y_{ij}] = E[\mu + \theta_i + x_{ij}'\beta]$$
$$= E[\mu] + E[\theta_i] + E[x_{ij}'\beta]$$
$$\boxed{= \mu + 0 + x_{ij}'\beta}$$

$$var[y_{ij}] = E[\,y_{ij} - E(y_{ij})]^2$$
$$= E[\theta_i + \varepsilon_{ij}]^2$$
$$= E\,\theta_i + 2E\,\theta_i\varepsilon_i + E\varepsilon_{ij}^2$$

Review derivations

$$var(y_{ij}|\theta_i) = E[\,y_{ij} - \mu + \theta_i + x_{ij}'\beta]^2$$
$$= E\,\varepsilon_{ij}^2 = \sigma^2$$

Pig Data Approach 3: Random Intercept Model

use lmerTest not lme4?

```
> library(lmerTest)
> fit.randomeffects = lmer(weight~weeks+(1|id), data = pig)
> summary(fit.randomeffects)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: weight ~ weeks + (1 | id)
   Data: pig

REML criterion at convergence: 2033.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.7390 -0.5456  0.0184  0.5122  3.9313

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 15.142   3.891
 Residual              4.395   2.096
Number of obs: 432, groups:  id, 48

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  19.35561    0.60314  58.55889   32.09   <2e-16 ***
weeks         6.20990    0.03906 383.00000  158.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr)
weeks -0.326
```

1|id
clustering variable

$\hat{\tau}^2$

$\hat{\sigma}^2$

Compared to a model fitted with group dummy variables, the *weeks* slope estimate and SE are identical (because the 'week' variable is the same for all pigs)

Nepalese Children: Random Intercept Model

1|id

```
> fit = lmer (arm ~ age + (1|id), data = nepal)
> random.eff.nepal = ranef (fit)$id[,1]
> summary(fit)

Linear mixed model fit by REML.
Formula: arm ~ age + (1 | id)
   Data: nepal
  AIC  BIC logLik deviance REMLdev
 1821 1840 -906.6     1799    1813

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.78073  0.88359
 Residual             0.24807  0.49806

Number of obs: 882, groups: id, 197

Fixed effects:
            Estimate Std. Error t value
(Intercept) 12.753789   0.109667  116.30
age          0.031697   0.002357   13.45

Correlation of Fixed Effects:
    (Intr)
age -0.803
```

extracting BLUPs?

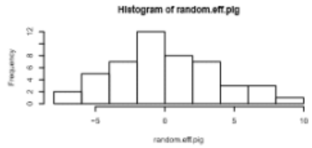
changed a little, relative to fixed effects model

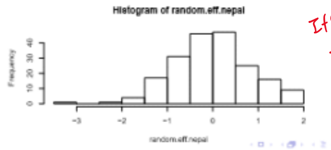
Nepalese Children: Random Intercept Model

- The fixed effect model had an *age* slope estimate of 0.0313 and a SE of 0.00307
- Here, data are not balanced. i.e. diff # of pts for diff babies
- We see a decrease in SE of slope of age with mixed model compared to fixed effects model.

Random Effect Histograms

Model formulation assumes random effect is normally distributed
↳ usually pretty reasonable assumption but sometimes there are exceptions





It's ok these don't look normal

Nepalese Children: Random Intercept Model Interpretation

```
> fit = lmer (arm ~ age + (1|id), data = nepal)

Linear mixed model fit by REML
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.78073  0.88359
 Residual             0.24807  0.49806

Fixed effects:
            Estimate Std. Error t value
(Intercept) 12.753789   0.109667  116.30
age          0.031697   0.002357   13.45
```

$\hat{\tau}^2$

→ just $\sqrt{\ }$ of variance aka $\sqrt{0.78073}$

We found a 0.032 cm ($CI_{95\%}$ 0.027, 0.037) increase in arm circumference per month after controlling for a child's arm circumference at birth.

We also found evidence of heterogeneity in arm circumference at birth. The estimated population-average arm circumference at birth is 12.8 cm, and the standard deviation of the random effect is 0.88 cm.

Nepalese Children: Random Intercept Model Interpretation

Consider another model with an indicator for mother's literacy.

```
> fit2 = lmer(arm~age+lit+(1|id), data = nepal)

> summary (fit2)
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.74712  0.86436
 Residual             0.24824  0.49823

Fixed effects:
            Estimate Std. Error t value
(Intercept) 12.710555   0.109304  116.29
age          0.031789   0.002338   13.60
lit          0.930247   0.316301    2.94
```

We found literacy to be significantly associated with arm circumference as a main effect. Also note that there is a small decrease in the degree of heterogeneity (from 0.78 to 0.75). Therefore mother's literacy may help explain some of the observed between-children variation in arm circumference at birth. Also the intercept estimate 12.71 now corresponds to the population-average arm circumference at birth from mothers who are illiterate.

9/11 Start

Covariance Structure

A random intercept model is also known as a two-level variance component model. Note that

conditional formulation

$$y_{ij} = \mu + \theta_i + \beta x_{ij} + \epsilon_{ij}, \quad \theta_i \overset{iid}{\sim} N(0, \tau^2), \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \quad \theta_i \perp\!\!\!\perp \epsilon_{ij}$$

can be re-written as

$$y_{ij} = \mu + \beta x_{ij} + \overbrace{\epsilon^*_{ij}}^{\theta_i + \varepsilon_{ij}}, \quad \epsilon^*_{ij} \sim N(0, \tau^2 + \sigma^2).$$ marginal formulation

Let $\epsilon^* = [\epsilon^*_{11}, \epsilon^*_{12}, \ldots, \epsilon^*_{1r}, \epsilon^*_{21}, \ldots, \ldots]'$

What is Cov ϵ^*, or equivalently, Cov $\mathbf{Y}$?

(something here he work

Approach 1: (incorrect)
- ignore repeated measurements

Approach 2: (okay)
- separate intercept (fixed effects) for each subject

Approach 3: mixed

Covariance Structure

$\text{Cov}(X,Y) = E[(X-E(X))(Y-E(Y))]$

$\varepsilon_{ij}^* = \theta_i + \varepsilon_{ij}$

$\text{Cov}(\varepsilon_{ij}^*, \varepsilon_{ij}^*) = \tau^2 + \sigma^2$

$\text{Cov}(\varepsilon_{ij}^*, \varepsilon_{ij'}^*) = \text{Cov}(\theta_i + \varepsilon_{ij}, \theta_i + \varepsilon_{ij'}) = \text{Cov}(\theta_i, \theta_i)$
$= \tau^2$

same subject

diff measurement

EXAM

$$Cov(\varepsilon_{ij}^*, \varepsilon_{ij'}^*) = Cov(\theta_i + \varepsilon_{ij}, \theta_i + \varepsilon_{ij'}) = Cov(\theta_i, \theta_i) = \tau^2$$

same subject, diff measurement

$$Cov(\varepsilon_{ij}^*, \varepsilon_{i'j}^*) = Cov(\theta_i + \varepsilon_{ij}, \theta_{i'} + \varepsilon_{i'j}) = 0$$

diff subject, same measurement

new subject, so new realization of measurement error

Covariance Structure

this is all obs of subject 1

$\varepsilon_{ij}^* = \theta_i + \varepsilon_{ij}$

$$
\begin{array}{c|cccccc}
 & \varepsilon_{11}^* & \varepsilon_{12}^* & \varepsilon_{13}^* & \varepsilon_{21}^* & \varepsilon_{22}^* & \varepsilon_{23}^* \\
\hline
\varepsilon_{11}^* & \sigma^2+\tau^2 & \tau^2 & \tau^2 & 0 & 0 & 0 \\
\varepsilon_{12}^* & \tau^2 & \sigma^2+\tau^2 & \tau^2 & 0 & 0 & 0 \\
\varepsilon_{13}^* & \tau^2 & \tau^2 & \sigma^2+\tau^2 & 0 & 0 & 0 \\
\varepsilon_{21}^* & 0 & 0 & 0 & \sigma^2+\tau^2 & \tau^2 & \tau^2 \\
\varepsilon_{22}^* & 0 & 0 & 0 & \tau^2 & \sigma^2+\tau^2 & \tau^2 \\
\varepsilon_{23}^* & 0 & 0 & 0 & \tau^2 & \tau^2 & \sigma^2+\tau^2
\end{array}
$$

this is all obs of subject 2

Random Intercept Model in Matrix Form

Consider the mixed model with random intercepts for n groups and define $N = \sum_{i=1}^n r_i$.

$$\mathbf{y} = \underbrace{\mathbf{Z}}_{n \times n}\boldsymbol{\theta} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \\ 1 & 0 & 0 & \\ 0 & 1 & 0 & \\ 0 & 1 & 0 & \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \end{bmatrix} + \begin{bmatrix} 1 & x_{111} & x_{112} \\ 1 & x_{121} & x_{122} \\ \vdots & \vdots & \vdots \\ 1 & & \end{bmatrix} \begin{bmatrix} \hat{\beta} \end{bmatrix} \ldots$$

where

- $\mathbf{y} = N \times 1$ vector of response.
- $\mathbf{Z} = N \times n$ design matrix of indicator variables for each group. (subject)
- $\boldsymbol{\theta} = n \times 1$ vector of random intercepts.
- $\mathbf{X} = N \times p$ design matrix of fixed effects (including overall intercept).
- $\boldsymbol{\beta} = p \times 1$ vector of fixed effects.
- $\boldsymbol{\epsilon} = N \times 1$ vector of residual error.

Assumptions

- $\boldsymbol{\theta} \sim N(\mathbf{0}, \tau^2 \mathbf{I}_{n \times n})$.
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{N \times N})$.

Intraclass Correlation (ICC) = ρ

Note that the within-group covariance is

$$Cov(\, y_{ij}, y_{ij'}) = \tau^2.$$

So the correlation between observations within the same group is

$$\frac{Cov(y_{ij}, y_{ij'})}{\sqrt{Var(y_{ij})Var(y_{ij'})}} \qquad \rho = Corr(\, y_{ij}, y_{ij'}) = \frac{\tau^2}{\tau^2 + \sigma^2} \text{ for all } j \neq j'. \tag{1}$$

The value ρ is often called the intraclass correlation. It measures the degree of similarity among same-group observations compared to the residual error σ^2.

Application: reproducibility studies.

Example: Multiple scans of a subject's brain, and measure the connections between brain regions. We assume differences between the scans are due to measurement error. Then σ^2 quantifies measurement error, ρ = reproducibility.

ICC, cont.

$$\rho = Corr(\, y_{ij}, y_{ij'}) = \frac{\tau^2}{\tau^2 + \sigma^2} \text{ for all } j \neq j'. \tag{2}$$

- $\rho \to 0$ when $\tau^2 \to 0$ (*i.e. same intercept*).
- $\rho \to 0$ when $\sigma^2 \to \infty$ (*i.e. growing measurement error*).
- $\rho \to 1$ when $\tau^2 \to \infty$ (*i.e. large separation in intercepts*).
- $\rho \to 1$ when $\sigma^2 \to 0$ (*i.e. zero measurement error*).

The above intraclass correlation has an exchangeable structure because the correlation is constant between any pair of within-group observations.

ICC, cont.

$$\rho = Corr(\, y_{ij}, y_{ij'}) = \frac{\tau^2}{\tau^2 + \sigma^2} \text{ for all } j \neq j'. \qquad (2)$$

- $\rho \to 0$ when $\tau^2 \to 0$ (*i.e. same intercept*).
- $\rho \to 0$ when $\sigma^2 \to \infty$ (*i.e. growing measurement error*).
- $\rho \to 1$ when $\tau^2 \to \infty$ (*i.e. large separation in intercepts*).
- $\rho \to 1$ when $\sigma^2 \to 0$ (*i.e. zero measurement error*).

The above intraclass correlation has an exchangeable structure because the correlation is constant between *any pair* of within-group observations.

Simulated Pig Data Ex 1 – Between subject variability

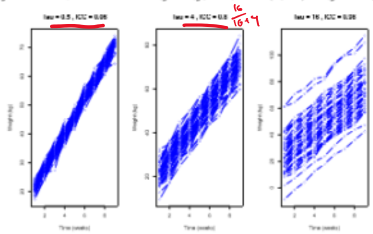
carry between subject variability

$$y_{ij} = 15 + \theta_i + 6.2 \times \text{weeks}_{ij} + \epsilon_{ij}, \quad \theta_i \sim N(0, \tau^2) \quad \epsilon_{ij} \sim N(0, 4)$$

σ^2

$\frac{16}{16+4}$

$\frac{0.25}{4.25}$

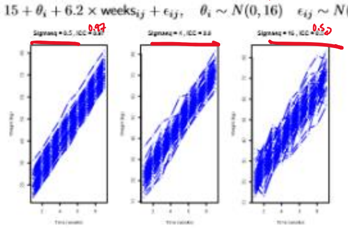


Simulated Pig Data Ex 2 – Measurement error

τ^2 $(\tau = 4)$

$$y_{ij} = 15 + \theta_i + 6.2 \times \text{weeks}_{ij} + \epsilon_{ij}, \quad \theta_i \sim N(0, 16) \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

0.97 0.50



Shrinkage and Random Effects

To simplify the derivation and make connections to ridge regression, we first consider a special case:

Consider a random effects model without fixed effects:

RE model

$$y_{ij} = \theta_i + \epsilon_{ij}, \quad \theta_i \overset{iid}{\sim} N(0, \tau^2) \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

$\boxed{P(x,y) = P(x|y)P(y)}$

The joint density of the data and random effects is given by

$$\prod_{i,j} f(y_{ij}, \theta_i) = \prod_{i,j} f(y_{ij}|\theta_i) \times \prod_i g(\theta_i)$$

(drop terms that are equal to 0 when taking derivative with respect to θ)

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - \theta_i)^2\right] \times \exp\left[-\frac{1}{2\tau^2}\boldsymbol{\theta}'\boldsymbol{\theta}\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - \theta_i)^2 - \frac{1}{2\tau^2}\boldsymbol{\theta}'\boldsymbol{\theta}\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2}\left[\sum_{i,j}(y_{ij} - \theta_i)^2 + \frac{\sigma^2}{\tau^2}\boldsymbol{\theta}'\boldsymbol{\theta}\right]\right]$$

$$\frac{\partial}{\partial \theta}\left[-2y'z\theta + \theta'z'z\theta + \frac{\sigma^2}{\tau^2}\theta'\theta\right]$$

$$-2z'y + 2z'z\theta + \frac{2\sigma^2}{\tau^2}\theta$$

set to zero

$$0 = -2z'y + 2z'z\theta + \frac{2\sigma^2}{\tau^2}\theta$$

$$z'y = \left(z'z\theta + \frac{\sigma^2}{\tau^2}I\right)\theta$$

Shrinkage and Random Effects

Then maximizing the log likelihood is equivalent to

$$\arg\min \left[\sum_{i,j}(y_{ij} - \theta_i)^2 + \frac{\sigma^2}{\tau^2}\sum_i \theta_i^2\right]$$

Consider the matrix formulation

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

where $\mathbf{Z} \in \mathbb{R}^{nr \times n}$ design matrix of indicator variables denoting the ijth observation belongs to group i, for clarity we assume r observations in all groups. Then

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

where $\mathbf{Z} \in \mathbb{R}^{nr \times n}$ design matrix of indicator variables denoting the ijth observation belongs to group i, for clarity we assume r observations in all groups. Then

$$\arg\min \left[(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}) + \frac{\sigma^2}{\tau^2}\boldsymbol{\theta}'\boldsymbol{\theta} \right]$$

Given values of σ^2 and τ^2, it's easy to find the closed-form solution to this. We will see it again in ridge regression in module 6:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\tau^2}\mathbf{I} \right)^{-1} \mathbf{Z}'\mathbf{y}$$

$$-2Z'y + 2Z'Z\theta + 2\frac{\sigma^2}{\tau^2}\theta$$

$$Z'y = \left(Z'Z\theta + \frac{\sigma^2}{\tau^2}I\theta\right)$$

$$\left(Z'Z + \frac{\sigma^2}{\tau^2}I\right)^{-1} Z'y = \hat{\theta}$$

Shrinkage and Random Effects, cont.

This is equivalent to

$$\hat{\theta}_i = \frac{\sum_{j=1}^{r} y_{ij}}{r + \boxed{\sigma^2/\tau^2}}$$

without this, it's FE

Note that recall we have special case where pop mean = 0

- $\hat{\theta}_i \to 0$ when $\tau^2 \to 0$ (*i.e.* shrinks all random intercepts to zero). aka μ → μ
- $\hat{\theta}_i \to \bar{y}_{i\cdot}$ when $\tau^2 \to \infty$ (*i.e.* no shrinkage = raw group mean estimates) i.e. participant
- $\hat{\theta}_i \to \bar{y}_{i\cdot}$ when $\sigma^2 \to 0$ (*i.e.* no shrinkage = raw group mean estimates). aka μ
- $\hat{\theta}_i \to \bar{y}_{i\cdot}$ when $r \to \infty$ (*i.e.* no shrinkage = raw group mean estimates) # obs of each subject

τ^2 controls the amount of shrinkage and how much information to borrow across groups

What happens if groups differ a lot?

Shrinkage and Random Effects - EDF

skipped

In penalized regression, the notion of effective degrees of freedom is useful for generalizing the notion of the number of parameters to models in which parameter estimates are shrunk towards zero.

Recall in multiple regression, $\text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') =$ number of parameters.

For ridge regression, $\text{EDF} = \text{tr}\left[\mathbf{X}\left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\right]$.

The notion of effective degrees of freedom (EDF) can be extended to understanding random effects:

$$\text{EDF} = \text{tr}\left[\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1}\mathbf{Z}'\right].$$

The amount of shrinkage depends on the ratio of between-group versus within-group variation.

Shrinkage and Random Effects - EDF

skip

For the pig data, we have $\mathbf{Z}'\mathbf{Z} = 9 \times \mathbf{I}_{48\times48}$. So

$$\begin{aligned}
\text{EDF} &= \text{trace}\left[\mathbf{Z}\left(9 + \frac{\sigma^2}{\tau^2}\right)^{-1} \times \mathbf{I}\mathbf{Z}'\right] = \text{trace}\left[\left(9 + \frac{\sigma^2}{\tau^2}\right)^{-1}\mathbf{Z}\mathbf{Z}'\right] \\
&= \left(\frac{9\tau^2 + \sigma^2}{\tau^2}\right)^{-1}\text{trace}[\mathbf{Z}\mathbf{Z}'] = 48 \times 9\left(\frac{\tau^2}{9\tau^2 + \sigma^2}\right) \\
&= 48\left(\frac{9}{9 + \sigma^2/\tau^2}\right)
\end{aligned}$$

Shrinkage and Random Effects - EDF Skip

$$\text{EDF} = 48\left(\frac{9}{9+\sigma^2/\tau^2}\right)$$

EDF $\to$ 48 (less shrinkage) when:

- $\sigma^2/\tau^2 \to 0$
- Within-pig variation $\sigma^2 \ll$ between-pig variation τ^2.
- Clear separation of the pig-specific intercepts. Estimate the intercepts close to fixed effects.

EDF $\to$ 0 (more shrinkage) when:

- $\sigma^2/\tau^2 \to \infty$
- Within-pig variation $\sigma^2 \gg$ between-pig variation τ^2.
- Random residual error σ^2 dominates. Make estimates of the pig-specific intercepts more similar to each other, as overall mean is more informative.

Random effects are a sort of compromise between "Approach 1" (one intercept) and "Approach 2" (intercept for each subject).

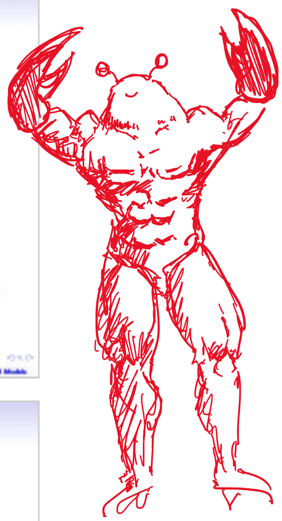
Shrinkage and Random Effects - EDF Skip

Let n be the number of subjects/groups, and r be the number of observations within each group. Then for a simple random intercept model with no fixed effect:

$$\text{EDF} = n\left(\frac{r}{r+\sigma^2/\tau^2}\right).$$

Also note that EDF $\to n$ when r increases. Less shrinkage is experienced because with large r, we have sufficiently large sample size per group to estimate their own intercepts. So there is no need to rely on the normality assumption to borrow information between groups.

Shrinkage and Borrowing Information



In Slide 30, we assumed the population mean was 0. Now assume the random effects are centered around a common mean μ:

$$y_{ij} = \theta_i + \epsilon_{ij}, \quad \theta_i \sim N(\mu, \tau^2) \quad \epsilon_{ij} \sim N(0, \sigma^2).$$

↳ hierarchical formulation

The joint density of the data and random effects is then

$$\begin{aligned}\prod_{i,j} f(y_{ij}, \theta_i) &= \prod_{i,j} f(y_{ij}|\theta_i) \times \prod_i g(\theta_i)\\ &\propto \exp\left[-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{Z}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{Z}\boldsymbol{\theta}) + \frac{\sigma^2}{\tau^2}(\boldsymbol{\theta}-\boldsymbol{\mu})'(\boldsymbol{\theta}-\boldsymbol{\mu})\right]\right]\\ &\propto \exp\left[-\frac{1}{2\sigma^2}\left[-2\mathbf{y}'\mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\theta}'(\mathbf{Z}'\mathbf{Z})\boldsymbol{\theta} + \frac{\sigma^2}{\tau^2}\boldsymbol{\theta}'\boldsymbol{\theta} - 2\frac{\sigma^2}{\tau^2}\boldsymbol{\mu}'\boldsymbol{\theta}\right]\right]\\ &= \exp\left[-\frac{1}{2\sigma^2}\left[\boldsymbol{\theta}'(\mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\tau^2}\mathbf{I})\boldsymbol{\theta} - 2(\mathbf{y}'\mathbf{Z} + \frac{\sigma^2}{\tau^2}\boldsymbol{\mu}')\boldsymbol{\theta}\right]\right]\end{aligned}$$

Recall the *completing the squares* property: let A be a symmetric and invertible matrix, then

$$\boldsymbol{\theta}'\mathrm{A}\boldsymbol{\theta} - 2\boldsymbol{\alpha}'\boldsymbol{\theta} = (\boldsymbol{\theta} - \mathrm{A}^{-1}\boldsymbol{\alpha})'\mathrm{A}(\boldsymbol{\theta} - \mathrm{A}^{-1}\boldsymbol{\alpha}) - \boldsymbol{\alpha}'\mathrm{A}^{-1}\boldsymbol{\alpha}.$$

Shrinkage and Borrowing Information, cont.

The joint density is a multivariate Normal density:

$$\prod_{i,j} f(y_{ij}|\theta_i) \times \prod_i g(\theta_i) \propto \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \mathrm{A}^{-1}\boldsymbol{\alpha})'\mathrm{A}(\boldsymbol{\theta} - \mathrm{A}^{-1}\boldsymbol{\alpha})\right]$$

where $\mathrm{A} = (\mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\tau^2}\mathbf{I})$ and $\boldsymbol{\alpha} = (\mathbf{Z}'\mathbf{y} + \frac{\sigma^2}{\tau^2}\boldsymbol{\mu})$.

For maximizing $\boldsymbol{\theta}$, this function is maximized at the mean:

$$\hat{\boldsymbol{\theta}} = \mathrm{A}^{-1}\boldsymbol{\alpha} = (\mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\tau^2}\mathbf{I})^{-1}(\mathbf{Z}'\mathbf{y} + \frac{\sigma^2}{\tau^2}\boldsymbol{\mu}). \qquad (3)$$

Let r_i = number of replicates for the ith group. Then,

for $\theta_i \sim N(\mu, \tau^2)$
$\hat{\theta}_i$ borrows information from the pop.

$$\hat{\theta}_i = \frac{(\sigma^2/\tau^2)\mu + \sum_{j=1}^{r_i} y_{ij}}{r_i + \sigma^2/\tau^2}.$$

Note that

- $\hat{\theta}_i \to \mu$ when $\tau^2 \to 0$ (*shrink all random intercepts to a common mean*).
- $\hat{\theta}_i \to \bar{y}_{i\cdot}$ when $\tau^2 \to \infty$ (*no shrinkage = raw mean estimates*).

i.e. τ^2 is .... ??

Shrinkage and Borrowing Information, cont. ii

We can also express $\hat{\theta}_i$ as

inverse of inter-subject variance → w/in subject variance

$$\hat{\theta}_i = \frac{(1/\tau^2)\mu + (r_i/\sigma^2)\bar{y}_{i\cdot}}{1/\tau^2 + (r_i/\sigma^2)}.$$

Since (σ^2/r_i) is the sample variance of the estimated sample mean $\bar{y}_{i\cdot}$, the above form shows that random effects can be viewed as a weighted average of:

1. standard estimate without penalization: $\bar{y}_{i\cdot}$
2. overall mean μ.

with their corresponding inverse-variances as weights!

Finally, express $\hat{\theta}_i$ in terms of intraclass correlation $\rho = \tau^2/(\tau^2+\sigma^2)$

$$\hat{\theta}_i = \frac{\rho^{-1}\mu + r_i(1-\rho)^{-1}\bar{y}_{i\cdot}}{\rho^{-1} + r_i(1-\rho)^{-1}}$$

and less shrinkage is expected for $\rho \to 1$.

Best Linear Unbiased Prediction

For the random intercept model

$$y_{ij} = \theta_i + \epsilon_{ij}, \quad \theta_i \overset{iid}{\sim} N(\mu, \tau^2) \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

we wish to estimate the unobserved random variable θ_i.

We can also derive the estimators using the MVN (multi-variate normal) distribution. Assume τ^2 and σ^2 are known. Then

$$\begin{array}{l} nr\times 1 \\ n\times 1 \end{array} \begin{bmatrix} \mathbf{y}_i \\ \theta_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \mathbf{1}_{r_i} \\ \mu \end{bmatrix}, \begin{bmatrix} \tau^2 \mathbf{1}_{r_i}\mathbf{1}'_{r_i} + \sigma^2 \mathbf{I}_{r_i\times r_i} & \tau^2 \mathbf{1}_{r_i} \\ \tau^2 \mathbf{1}'_{r_i} & \tau^2 \end{bmatrix} \right)$$

because $cov(y_{ij}, \theta_i) = cov(\theta_i + \epsilon_{ij}, \theta_i) = \tau^2$.

To make a prediction of θ_i given the data $\mathbf{y}_i$, we can use the conditional distribution of the multivariate normal density. Specifically our estimator will be

$$\hat{\theta}_i = E[\theta_i \,|\, \mathbf{y}_i].$$

Best Linear Unbiased Prediction: BLUPs

$$\begin{aligned}
\hat{\theta}_i = E[\theta_i \,|\, \mathbf{y}_i] &= \mu + \tau^2 \mathbf{1}'_{r_i}[\tau^2 \mathbf{1}_{r_i}\mathbf{1}'_{r_i} + \sigma^2 \mathbf{I}_{r_i\times r_i}]^{-1}[\mathbf{y}_i - \mu \mathbf{1}_{r_i}] \\
&= \mu + \tau^2 \mathbf{1}'_{r_i}\frac{1}{\sigma^2}\left[\mathbf{I}_{r_i\times r_i} - \frac{\tau^2}{\sigma^2 + n\tau^2}\mathbf{1}_{r_i}\mathbf{1}'_{r_i}\right][\mathbf{y}_i - \mu \mathbf{1}_{r_i}] \\
&= \mu + \frac{\tau^2}{\sigma^2}\left(1 - \frac{r_i\tau^2}{\sigma^2 + r_i\tau^2}\right)\mathbf{1}'_{r_i}[\mathbf{y}_i - \mu \mathbf{1}_{r_i}] \\
&= \mu + \frac{\tau^2}{\sigma^2}\left(\frac{\sigma^2}{\sigma^2 + r_i\tau^2}\right)(r_i\bar{y}_{i\cdot} - r_i\mu) \\
&= \mu + \left(\frac{\tau^2}{\sigma^2 + r_i\tau^2}\right)(r_i\bar{y}_{i\cdot} - r_i\mu) \\
&= \frac{\sigma^2\mu + \tau^2 r_i \bar{y}_{i\cdot}}{\sigma^2 + r_i\tau^2}.
\end{aligned}$$

This is equivalent to (3). (Apply the Sherman-Morrison matrix inverse formula.)

eBLUPs

- For known variance parameters, $\hat{\theta}_i$ is the BLUP: Best Linear Unbiased Predictor.
- They are unbiased in the sense that $E(\hat{\theta}_i) = E(\theta_i) = \mu$, see Robinson 1991 (in course files /Readings).
- They are "best" in the sense that the conditional expectation minimizes the mean-squared error $E(\hat{\theta}_i - \theta_i)^2$ among the class of linear unbiased estimators.
- Note: in ordinary linear regression, $y_i = \boldsymbol{x}'_i\boldsymbol{\beta} + \epsilon_{ij}$, the least-squares estimate of $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the Best Linear Unbiased Estimator (BLUE).
- In practice, we can't estimate BLUPs because their variances are not known.
- We use $\hat{\sigma}_2$ and $\hat{\tau}^2$ in place of their true values.
- The resulting random effects estimators are eBLUPs: estimated Best Linear Unbiased Predictors
- Connections to Bayesian statistics: see Robinson (1991).

BLUPs: Unbiased but... biased?

Let's go back to the model $\theta_i \sim N(0, \sigma^2)$ (slide 30), where we assume mean 0 to simplify the formulae.

Assume the conditional model $y_{ij} \mid \theta_i = \theta_i + \epsilon_{ij}$ such that $E[y_{ij} \mid \theta_i] = \theta_i$. Additionally assume σ^2 and τ^2 known.

From this perspective, the random intercepts are biased. For $\tau^2 > 0$,

$$E[\hat{\theta}_i \mid \theta_i] \mid = \mid E\left[\frac{\sum_{j=1}^r y_{ij}}{r + \sigma^2/\tau^2} \mid \theta_i\right] \mid < \mid E\left[\frac{\sum_{j=1}^r y_{ij}}{r} \mid \theta_i\right] \mid = \mid \theta_i \mid.$$

However, the variances are smaller.

$$Var[\hat{\theta}_i \mid \theta_i] = Var\left[\frac{\sum_{j=1}^r y_{ij}}{r + \sigma^2/\tau^2} \mid \theta_i\right] < Var\left[\frac{\sum_{j=1}^r y_{ij}}{r} \mid \theta_i\right].$$

We see a trade-off between bias and variance. Some bias is introduced, but we get smaller standard error.

skip

BLUPs: Matrix formulation

BLUPs can be derived as the conditional distribution of $\boldsymbol{\theta}$ given the data $\mathbf{y}$. Consider the joint distribution of $[\mathbf{y}, \boldsymbol{\theta}]$:

$$\begin{bmatrix}\mathbf{y}\\ \boldsymbol{\theta}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{X}\beta\\ 0\end{bmatrix}, \begin{bmatrix}\tau^2\mathbf{Z}\mathbf{Z}' + \sigma^2\mathbf{I}_{N\times N} & \tau^2\mathbf{Z}\\ \tau^2\mathbf{Z}' & \tau^2\mathbf{I}_{n\times n}\end{bmatrix}\right)$$

Then

$$E[\boldsymbol{\theta}|\mathbf{y}] = (\tau^2\mathbf{Z}')\left(\tau^2\mathbf{Z}\mathbf{Z}' + \sigma^2\mathbf{I}_{N\times N}\right)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

E.g., see "Conditional Distributions" at https://en.wikipedia.org/wiki/Multivariate_normal_distribution

In practice, replace $\boldsymbol{\beta}$, σ^2, and τ^2 by their estimates.

start

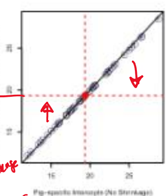
Shrinkage: Pig Data

$$\text{weight}_{ij} = \beta_0 + \theta_i + \beta_1 \text{week}_{ij} + \epsilon_{ij} \qquad \theta_i \overset{iid}{\sim} N(0, \tau^2) \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

$$\hat{\tau}^2 = 15.1, \ \hat{\sigma}^2 = 4.39, \ \hat{\sigma}^2/\hat{\tau}^2 = 0.29, \ \underbrace{\text{ICC} = 0.77}_{\text{"good"}}, \ r_i = 9 \text{ for all } i$$

approach 3: mixed

overall mean

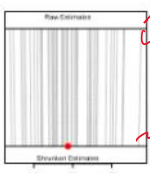


y=x: no shrinkage

↳ corresponds to approach 2: FE

approach 2 FE

$\hat{\beta}_i$



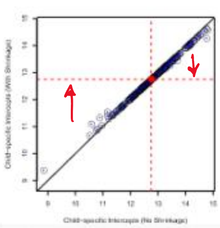
$\hat{\beta}_0 + \hat{\theta}_i$

Shrinkage: Nepalese Data

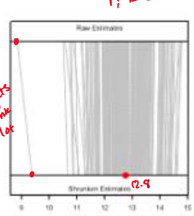
$$\text{armc}_{ij} = \beta_0 + \theta_i + \beta \text{age}_{ij} + \epsilon_{ij} \qquad \theta_i \overset{iid}{\sim} N(0, \tau^2) \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

$$\hat{\tau}^2 = 0.78, \ \hat{\sigma}^2 = 0.25, \ \hat{\sigma}^2/\hat{\tau}^2 = 0.32, \ \text{ICC} = 0.76, \ \underbrace{r_i \in \{1, \ldots, 5\}}_{r_i \text{ is smaller}}$$

$\hat{\beta}_0 + \hat{\theta}_i$



gets shrunk a lot



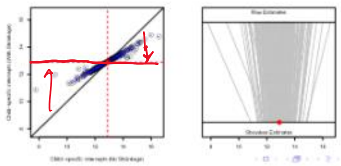
Shrinkage: Nepalese Data with Noise

What happens if we add more random noise to the outcome? More Shrinkage!

$$\text{armc}^*_{ij} = \beta_0 + \theta_i + \beta_1 \text{age}_{ij} + \epsilon_{ij} \qquad \theta_i \overset{iid}{\sim} N(0, \tau^2) \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

where armc* = arm $+N(0,2)$. added more noise

$\hat{\tau}^2 = 0.74 \quad \hat{\sigma}^2 = 2.27 \quad \hat{\sigma}^2/\hat{\tau}^2 = 3.07 \quad \text{ICC} = 0.24$

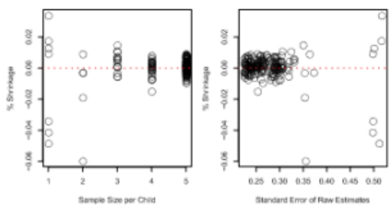


Shrinkage and Borrowing Information, cont. iii

The Nepalese children dataset contains missing data. Not all 200 children have complete 5 visits. $\frac{r_i}{\sigma^2}$

More shrinkage when have fewer visits (use more pop. info)

Note how the amount of shrinkage is related to the standard error of the fixed effects model (raw estimates = slide 12)



Normality assumption and shrinkage

summary:

Shrinkage occurs according to the conditional mean determined by the normality assumption.

Effects of the normality assumption (amount of shrinkage) on random effects depend on

1. group-specific sample size, r_i
2. within-group residual error, σ^2
3. between-group heterogeneity. τ^2

Parameter Estimation: Maximum Likelihood Approach

matrix of 0's and 1's, random effect, fixed effects

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \boldsymbol{\theta} \sim N(\mathbf{0}, \tau^2\mathbf{I}) \quad \epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

Since $\boldsymbol{\theta}$ and ϵ are *random variables*, we can rewrite the above as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon^*, \quad \epsilon^* = \mathbf{Z}\boldsymbol{\theta} + \epsilon.$$

We know $Cov(\epsilon^*) = \mathbf{V} = \mathbf{Z}Cov(\boldsymbol{\theta})\mathbf{Z}' + Cov(\epsilon) = \tau^2\mathbf{ZZ}' + \sigma^2\mathbf{I}$.

equivalent to matrix in slide 24/26 ?

This is equivalent to integrating out the random effects. Then the marginal model is:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

Diagonals: $\sigma^2 + \tau^2$

off diagonals within subject : τ^2

off-block (between subjects) : 0

start 9/13

Generalized Least Squares

For known $\mathbf{V}$, the generalized least-squares problem is

$$\underset{\boldsymbol{\beta}}{\arg\min} \; (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

$\mathbf{V}^{-1}$: no longer diagonal i.e. $\sigma^2 I$

This is also called weighted least squares.

Note this is the kernel of the multivariate normal distribution $\rightarrow f(\tilde{y}) = (2\pi)^{-p/2} |V|^{-1/2} e^{-\frac{1}{2}(\tilde{y} - \tilde{X}'\beta) V^{-1} (\tilde{y} - \tilde{X}'\tilde{\beta})}$

"kernel" of multivariate

Then the value of $\boldsymbol{\beta}$ that maximizes the likelihood is given by the generalized least-squares estimate:

This is also called weighted least squares.

Note this is the kernel of the multivariate normal distribution. → $f(\tilde{y}) = (2\pi)^{-P/2}\,|V|^{-1/2}\, e^{-\frac{1}{2}(\tilde{y}-\tilde{x}'\tilde{\beta})V^{-1}(\tilde{y}-\tilde{x}'\tilde{\beta})}$

Then the value of β that maximizes the likelihood is given by the generalized least-squares estimate:

$$\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$ reminder: $\in \mathbb{R}^N$

"kernel" of multivariate normal distribution

This estimator is the best linear unbiased estimator (BLUE).

$N = \sum_{i=1}^{n} r_i$, r_i = # obs on ith subject

Check equation later if he corrected it

Parameter Estimation: Maximum Likelihood Approach

The log-likelihood $l(\sigma^2,\tau^2)$ in terms of σ^2 and τ^2 is:

$$l(\sigma^2,\tau^2) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\tilde{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\beta}).$$

Plug in $\tilde{\beta} = \boxed{(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}}$ substitute

It is then straightforward to maximize the above function over the 2-D domain of σ^2 and τ^2.

This method of substituting some unknown parameters (β) with their MLE fixed at some other parameters (σ^2 and τ^2) is known as a profile likelihood approach.

REML

The MLE estimate of variances are biased. An alternative is restricted maximum likelihood (REML)

$$\underbrace{l(\sigma^2,\tau^2)}_{\text{same but with}} - \underbrace{\frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|}_{\text{ADDITIONAL term}}$$ for lost degrees of freedom

↳ bigger estimates of SEs, more conservative

to account for the degrees of freedom in the fixed effects (e.g., Ch. 6 in Searle et al. 1992, "Variance Components").

REML can be unbiased.

In the simple case of estimating σ^2 from $\mathbf{X}_i \overset{iid}{\sim} N(0,\sigma^2)$, we have

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\hat{\sigma}^2_{REML} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Small samples: often prefer REML. → default, a little more conservative

Likelihood ratio tests and AIC: use ML.

↳ based on ML so use ML

library(lme4)

Parameter Estimation: MLE versus REML

subject random effect

fixed effect for week

```
> fit1 <- lmer (weight~weeks+(1|id), data = dat)
> summary (fit1)
Linear mixed model fit by REML
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 15.1418  3.8913
 Residual              4.3947  2.0964

Fixed effects:
            Estimate Std. Error t value
(Intercept) 19.35561    0.60311   32.09
weeks        6.20990    0.03906  158.97

> fit2 <- lmer (weight~weeks+(1|id), REML = FALSE,data = dat)
> summary(fit2)
Linear mixed model fit by maximum likelihood
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 14.8175  3.8493
 Residual              4.3833  2.0936
Number of obs: 432, groups: id, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) 19.35561    0.59737    32.4
weeks        6.20990    0.03901   159.2
```

$\hat{\tau}^2$, $\hat{\sigma}^2$, $\hat{\beta}_0$, $\hat{\beta}_1$

no p-value → need to load lmerTest to get p-values

smaller than REML

Note that the standard errors are larger for REML.

Fixed versus Random

Consider the model:

$$y_{ij} = \theta_i + \beta' x_{ij} + \epsilon_{ij},\ \epsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)\theta_i : \text{either fixed or random}$$

- Fixed effects: we can treat θ_i as fixed. Note: to make comparable to RE, we can use the sum-to-zero constraint, $\sum_{i=1}^{n}\theta_i = 0$, and estimate the intercept.
- We can treat θ_i as random, $\theta_i \overset{iid}{\sim} N(0,\tau^2)$.

A useful paradigm: one person's covariance structure is another person's mean structure.

think of as fixed

Random: Consider $E(y_{ij} - \beta' x_{ij})^2 = \sigma^2 + E\theta_i^2$. (model the variance)

Fixed: $E(y_{ij} - \theta_i - \beta' x_{ij})^2 = \sigma^2$. (model the mean structure)

Guidelines for choosing fixed vs random

approach 1 approach 2

- Are we interested in predicting subject effects?
  - RE leverages population info – lower prediction error if treat θ_i as random.
- If the experiment were repeated, would the same subjects (i.e., groups) be used?
  - If yes, suggests FE.
- Or are the subjects a random sample from a population of interest?
  - RE
- Are there enough subjects to estimate heterogeneity?
  - E.g., if two subjects, use FE.
- Are there enough repeated measurements to estimate FE?
  - E.g., two measurements for a subject, use RE
- Do some subjects have only 1 observation and/or is there different number of samples for each subject?
  - Consider RE to leverage subjects with more information.

Fixed versus Random

However, in scientific applications, we are often interested in inference on a fixed covariate, and the variable we are deciding to treat as fixed or random (subject, plot, etc.) is a "nuisance" variable.

In this case, the choice of fixed versus random may not have a big impact on inference. You can look at how sensitive your findings are to fixed versus random specification.

Pig data: data were balanced and t-statistics of week equivalent.

Nepal data: estimates of slope of age similar (t = 10.20 in FE, versus t = 13.45)

The big issue is that we need to account for repeated observations in clustered data, and both approaches allow for valid inference on fixed covariates of interest.

Contrast with a model estimating a single intercept (slides 7 and 12), which results in incorrect standard errors, resulting in invalid inference.

Keywords

not synonyms, but related

- Clustered / correlated / grouped / longitudinal / multi-level / hierarchical / nested data
- Random effect / (~~Bayesian~~) hierarchical / mixed / variance component model
- Between-group variability / heterogeneity / structured error
- Within-group correlation / intraclass correlation
- Within-group variability / unstructured (residual) error / measurement error
- Shrinkage / penalization / borrowing information / smoothing

borrow information from pop. to improve subject level predictions

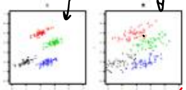
Quiz 3

$$y_{ij} = \mu + \theta_i + \epsilon_{ij}, \quad \theta_i \sim N(0, \tau^2) \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

(between group) (within group)

Part I

τ^2 bigger σ^2 bigger



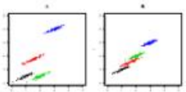
more info pg. 40

more shrinkage from lower ICC

1. Has the larger σ^2? B ✓
2. Has the larger intraclass correlation? A ✓
3. $\hat{\theta}_i$ will experience more shrinkage? B ~~A? τ^2 bigger?~~
4. Has the larger prediction SE for an observation from a within-sample group? B ✓

aka which has a larger σ^2 ?

Part II



Which plot:

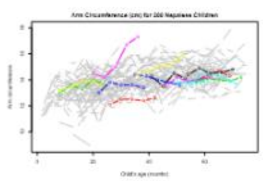
5. Has the larger τ^2? A ✓
6. Has the larger intraclass correlation? → $\frac{\tau^2}{\tau^2+\sigma^2}$
7. $\hat{\theta}_i$ will experience more shrinkage? ~~A~~ B
8. Has the larger prediction SE for an observation from an out-of-sample group? A

aka which has a larger τ^2

A, larger ICC for larger τ^2 given fixed σ^2

Random Slope Model

Nepalese Children Data



Scientific questions about arm circumference and age:

- What is the overall trend?
- How much do growth patterns differ between children?
- Do maternal covariates explain variability in growth patterns between children?

Random Intercept and Random Slope Model

Let $ageC_{ij}$ be the child's age in months minus 36.

$$\text{arm}_{ij} = \beta_0 + \theta_{0i} + (\beta_1 + \theta_{1i})\,\text{ageC}_{ij} + \epsilon_{ij}$$

$$\begin{bmatrix}\theta_{0i}\\ \theta_{1i}\end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau_1^2 & \rho\tau_1\tau_2\\ \rho\tau_1\tau_2 & \tau_2^2\end{bmatrix}\right), \ \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \ \boldsymbol{\theta}_i \perp\!\!\!\perp \epsilon_{ij}$$

$\text{Corr}(\theta_{0i}, \theta_{1i}) = \rho$

$\frac{\text{Cov}(\theta_{0i}, \theta_{1i})}{\tau_1 \tau_2} = \rho$ → $\text{Cov}(\theta_{0i}, \theta_{1i}) = \rho\,\tau_1\tau_2$

The above model treats both intercept and slope of age as child-specific. These random effects represent child-specific deviations from the overall trend. We typically assume θ_{01} and θ_{1i} are bivariate normal.

- τ_1^2 describes between-children variation in baseline arm circumferences at at age three.
- τ_2^2 describes between-children variation in the linear effects of age. aka variation of random slope
- ρ describes the correlation between child-specific intercept and slope.
- σ^2 describes within-child variation around a child-specific linear growth trend. measurement error

lme4 (no p-value)

Mixed Model: Nepalese Children

```
> nepal$ageC = nepal$age - 36
> fit = lmer (arm~ageC+(ageC|id), data = nepal)
Linear mixed model fit by REML
Random effects:
 Groups   Name        Variance   Std.Dev. Corr
 id       (Intercept) 0.71937744 0.848161
          ageCenter   0.00043572 0.020874 0.090
 Residual             0.22657451 0.475998
Number of obs: 882, groups: id, 197

Fixed effects:
            Estimate Std. Error t value
(Intercept) 13.943962   0.066677  209.13
ageC         0.032527   0.002754   11.81
```

(age C | id) : it fits the random intercept
the random slope,
& the correlation between them (ρ)

$\hat{\tau}_1^2$ $\hat{\tau}_1$ $\hat{\tau}_2^2$ $\hat{\sigma}^2$ $\hat{\rho}$

Population distribution of random intercepts and slopes:

correlation

- Child-specific intercept: $\beta_0 + \theta_{0i} \sim N(13.9, 0.85^2)$ arm circumference of ith child at 36 months
- Child-specific slope: $\beta_1 + \theta_{1i} \sim N(0.033, 0.021^2)$
  Very high heterogeneity in the age effects. The central 95% of this distribution includes zero. Thus it's possible that a child's arm circumference does not increase with age.
- $\rho = cor(\beta_{0i}, \beta_{1i}) = 0.09$

$= \beta_0 + \theta_{0i}$, $\beta_{1i} = \beta_1 + \theta_{1i}$

choosing between random intercept vs. random slope model

Comparing models: AIC, but doesn't work well

Is the model preferred to the model with a random intercept only?

AIC often used in model selection. Popular approach to choosing which variables should be included in a model.
Lower is better.

RoT: Difference of 2 or more is substantially better.

"rule of thumb"

To compare models with different variance structures, one approach is to use Akaike's Information Criterion:

log likelihood

$$AIC = -2\ell(\boldsymbol{\theta}) + 2p$$

covariates + # variance components

where $\ell(\boldsymbol{\theta})$ is the log likelihood for all parameters $\boldsymbol{\theta}$ and p is the number of parameters.

For nested models (one model contains a subset of parameters of the other model), we can use a likelihood ratio test. → ANOVA(reduced, full) LRT (likelihood ratio test)

Both these approaches use the MLE, so should use REML=FALSE

when specifying LMM, set to false

Comparing models: caveat

Testing the significance of a variance component is problematic because the null hypothesis is on the boundary of the parameter space, e.g., $\tau_2^2 = 0$.

This makes the χ_1^2 approximation of the LRT a poor approximation of the distribution of the test statistic under the null.

Generally, this makes the p-value too large (i.e., favors simpler models).

The homework describes a preferred approach.

For additional details, see Section 2.5, Pinheiro and Bates, *Mixed-Effects Models in S and S-Plus*, 2000.

R package: RLRsim: exactRLRT to simulate from the finite sample null dist.

AIC: lower = better

Compare to model without random slope

fits a random slope & random intercept along with their correlation (τ_1^2, τ_2^2, ρ)

```
> fit = lmer (arm~ ageC + (ageC|id), data = nepal,REML=FALSE)
> fit.randomintercept = lmer(arm~ageC+(1|id),data=nepal,REML=FALSE)
> AIC(fit)
[1] 1802.579
> AIC(fit.randomintercept)
[1] 1807.264
```

reduced, no intercept

lower is better

Likelihood ratio test:

```
> anova(fit.randomintercept,fit)
Data: nepal
Models:
fit.randomintercept: arm ~ ageC + (1 | id)
fit: arm ~ ageC + (ageC | id)
                    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
fit.randomintercept  4 1807.3 1826.4 -899.63   1799.3
fit                  6 1802.6 1831.3 -895.29   1790.6 8.6854      2      0.013 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

AIC much improved

also significant

Both AIC and LRT indicate model with random slopes is preferred.

Mixed Model: Nepalese Children

Other options (not recommended).

Assume Independent Random Effects:

```
> fit.indep = lmer (arm~ ageC + (1|id) + (0+ageC|id), data = nepal)
> summary(fit.indep)
Linear mixed model fit by REML ['lmerMod']
Formula: arm ~ ageC + (1 | id) + (0 + ageC | id)
   Data: nepal

REML criterion at convergence: 1804.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5914 -0.4923  0.0625  0.5651  2.9879

Random effects:
 Groups   Name        Variance  Std.Dev.
 id       (Intercept) 0.7170900 0.8468
 id.1     ageC        0.0004122 0.0203
 Residual             0.2279327 0.4774
Number of obs: 882, groups:  id, 197

Fixed effects:
            Estimate Std. Error t value
(Intercept) 13.94277    0.06650  209.66
ageC         0.03225    0.00273   11.81
```

this excludes ρ

- usually we will fit the model including ρ
- for the advanced test exactLRT in HW2, we may use this

?

σ^2

Fixed effect model

An alternative framework would treat the intercepts as fixed.

$$\text{arm}_{ij} = \beta_0 + \theta_{0i} + (\beta_1 + \theta_{1i})\,\text{ageC}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

We can use the sum-to-zero contrasts: → means that all the slopes sum to 0

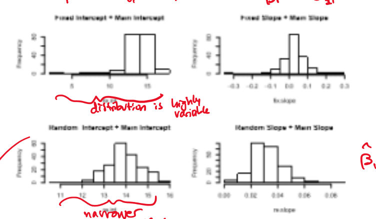
$$\sum_{i=1}^{n} \hat{\theta}_{0i} = 0$$

Then fixed effect interaction terms for each subject have similarities with random slopes (but don't leverage pop info), as the total age effect for each subject becomes $\hat{\beta}_1 + \hat{\theta}_{1i}$.

Distributions of Child-specific Intercepts and Slopes

$\hat{\beta}_0 + \hat{\theta}_{0i}$ → from fixed effects model

$\hat{\beta}_1 + \hat{\theta}_{1i}$ → from fixed effects



distribution is highly variable

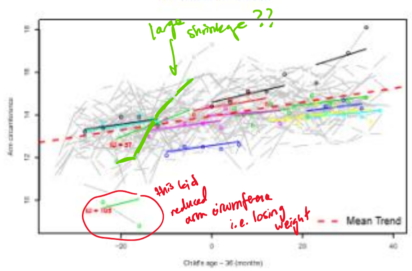
$\hat{\beta}_1 + \hat{\theta}_{1i}$ → from random effects

narrower range

why? Leveraging population information

$\hat{\beta}_0 + \hat{\theta}_{0i}$ → from RANDOM effects model

Mixed Model



large shrinkage ??

this kid reduced arm circumference i.e. losing weight

- Mean trend $= \beta_0 + \beta_1\,\text{ageC}_{ij}$
- i^{th} individual trend $= \beta_0 + \theta_{0i} + (\beta_1 + \theta_{1i})\,\text{ageC}_{ij}$

Go to R

M2-LMM.R

the random slopes are

ageC:id2
ageC:id2 for fixed effects

ageC for random

??? go back to recording

Mixed Model: Drop ID 57 and ID 105

```
> fit = lmer (arm~ageC+(ageC|id), data = subset(nepal, id != 57 & id != 105) )

Random effects:
 Groups   Name        Variance   Std.Dev. Corr
 id       (Intercept) 0.66724471 0.816850
          ageC        0.00029806 0.017264 0.123
 Residual             0.22133729 0.470465

Fixed effects:
            Estimate Std. Error t value
(Intercept) 13.949133   0.063888  218.34
ageC         0.031182   0.002581   12.08
```

- Changes in the fixed effects are minor:

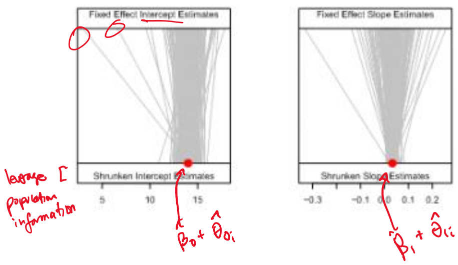
Mixed Model: Drop ID 57 and ID 105

```
> fit = lmer (arm~ageC+(ageC|id), data = subset(nepal, id != 57 & id != 105) )

Random effects:
 Groups   Name        Variance   Std.Dev. Corr
 id       (Intercept) 0.66724471 0.816850
          ageC        0.00029806 0.017264 0.123
 Residual             0.22133729 0.470465

Fixed effects:
            Estimate Std. Error t value
(Intercept) 13.949133   0.063888  218.34
ageC         0.031182   0.002581   12.08
```

- Changes in the fixed effects are minor:
  - Intercept: 13.944 → 13.949.
  - AgeC: 0.0325 → 0.0312.
- As expected, heterogeneity standard deviations become smaller:
  - $\hat{\tau}_1^2$ Intercept: 0.848 → 0.817.
  - $\hat{\tau}_2^2$ AgeC: 0.021 → 0.017.

- In HW2, will look at more diagnostic plots

Shrinkage!



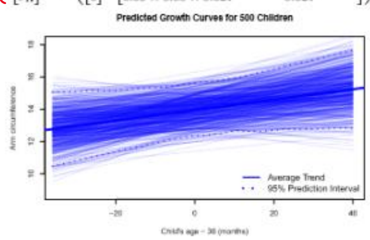
average [population information

$\hat{\beta}_0 + \hat{\theta}_{0i}$ $\hat{\beta}_1 + \hat{\theta}_{1i}$

Simulating Out-of-Sample Growth Curves

$$y_{ij} = (\overset{\hat{\beta}_0}{13.94} + \theta_{0i}) + (\overset{\hat{\beta}_1}{0.0325} + \theta_{1i})\, x_{ij} + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0, 0.48^2)$$

$$\begin{bmatrix}\theta_{0i}\\ \theta_{1i}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.85^2 & 0.09\times 0.85\times 0.021\\ 0.09\times 0.85\times 0.021 & 0.021^2\end{bmatrix}\right).$$



Hierarchical Formulation

Just different way of doing things

Consider the following model now including interactions:

$$\text{arm}_{ij} = \beta_0 + \theta_{0i} + (\beta_1 + \theta_{1i})\,\text{ageC}_{ij} + \epsilon_{ij} \tag{4}$$

$$[\theta_{0i}, \theta_{1i}]' \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma}), \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

In (4), the random effects are viewed as deviations from population averages. The model can also be written in a hierarchical (multilevel) model form:

$$\text{arm}_{ij} = \beta_{0i} + \beta_{1i}\,\text{ageC}_{ij} + \epsilon_{ij} \tag{5}$$

$$[\beta_{0i}, \beta_{1i}]' \sim N([\beta_0, \beta_1]', \mathbf{\Sigma}), \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

Equation (5) can also be written as:

earlier in lec this was β_0 → β_1

Level 1: $\beta_{0i} = \mu_0 + \theta_{0i} \qquad \beta_{1i} = \mu_1 + \theta_{1i}$

Level 2: $\text{arm}_{ij} = \beta_{0i} + \beta_{1i}\,\text{ageC}_{ij} + \epsilon_{ij}$

$$[\theta_{0i}, \theta_{1i}]' \sim N(\mathbf{0}, \mathbf{\Sigma}), \qquad \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

Hierarchical Formulation: Back to Random Intercepts

First consider the random intercept model with covariate age and lit (indicator for mother's literacy).

↳ lit_i : i^{th} child's mother's literacy

$$\text{arm}_{ij} = \beta_0 + \theta_{0i} + \beta_1 \text{ageC}_{ij} + \beta_2 \text{lit}_{ij} + \epsilon_{ij}$$
$$\theta_{0i} \overset{iid}{\sim} N(0, \tau^2), \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

What is the interpretation of β_2?

- Because lit_{ij} is an indicator variable, β_2 describes the difference in intercept (arm circumference at age 3) between literate mothers and illiterate mothers (reference).

However lit_{ij} is constant within each child. We can drop the j subscript and rewrite the model as

overall time-invariant effect

$$\beta_{0i} \sim N(\beta_0 + \beta_2 \text{lit}_i, \tau^2)$$
$$\text{arm}_{ij} = \beta_{0i} + \beta_1 \text{ageC}_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

Therefore an equivalent interpretation of β_2 is

- β_2 describes the difference in population averages in intercepts between literate and illiterate mothers.

Hierarchical Formulation

The hierarchical (multilevel) formulation is particularly useful when covariates are available or collected at different levels. Higher level (i) covariate values are constant in lower level (j).

Consider the following model: add literacy & sex

Level 1: $\beta_{0i} = \mu_0 + \alpha_{01} lit_i + \alpha_{02} sex_i + \theta_{0i}$

$\beta_{1i} = \mu_1 + \alpha_{11} lit_i + \alpha_{12} sex_i + \theta_{1i}, \quad [\theta_{0i}, \theta_{1i}]' \overset{iid}{\sim} N(\mathbf{0}, \Sigma)$

age Cij, weight Cij; overall slope age; interaction age-lit; interaction age-sex; subject specific deviation

Level 2: $\text{arm}_{ij} = \beta_{0i} + \beta_{1i} \text{ageC}_{ij} + \beta_2 \text{weightC}_{ij} + \epsilon_{ij}, \quad \epsilon \overset{iid}{\sim} N(0, \sigma^2)$

Note how we explicitly present covariates lit and sex as predictors that explain between-subjects heterogeneity. For example,

- α_{12} is the effect of a child's sex on the association between age and arm circumference after controlling for child-specific intercept and weight.
- β_2 is the effect of a child's weight on arm circumference adjusting for individual linear growth trend in age.

Cross-level Interactions

Level 1: $\beta_{0i} = \mu_0 + \alpha_{01} lit_i + \alpha_{02} sex_i + \theta_{0i}$

$\beta_{1i} = \mu_1 + \alpha_{11} lit_i + \alpha_{12} sex_i + \theta_{1i}, \quad [\theta_{0i}, \theta_{1i}]' \overset{iid}{\sim} N(\mathbf{0}, \Sigma)$

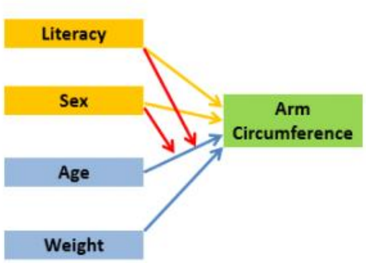
Level 2: $\text{arm}_{ij} = \beta_{0i} + \beta_{1i} \text{ageC}_{ij} + \beta_2 \text{weightC}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

By substituting Level 1 regressions into Level 2:

$$\begin{aligned}
\text{arm}_{ij} &= \mu_0 + \alpha_{01} lit_i + \alpha_{02} sex_i + \theta_{0i} \\
&\quad + (\mu_1 + \alpha_{11} lit_i + \alpha_{12} sex_i + \theta_{1i}) \text{ageC}_{ij} + \beta_2 \text{weightC}_{ij} + \epsilon_{ij} \\
&= \mu_0 + \alpha_{01} lit_i + \alpha_{02} sex_i + \theta_{0i} \\
&\quad + \mu_1 \text{ageC}_{ij} + \alpha_{11} lit_i \times \text{ageC}_{ij} + \alpha_{12} sex_i \times \text{ageC}_{ij} + \theta_{1i} \text{ageC}_{ij} \\
&\quad + \beta_2 \text{weightC}_{ij} + \epsilon_{ij}, \\
&\qquad [\theta_{0i}, \theta_{1i}]' \overset{iid}{\sim} N(\mathbf{0}, \Sigma), \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).
\end{aligned}$$

Note that α_{11} and α_{12} can be interpreted as an interaction between two variables across levels.

Cross-level Interactions



Ben rise

Cross-level Interactions

```
> nepal$wtC = scale(nepal$wt,center=TRUE,scale=FALSE)
> fit.sexwtlit = lmer (arm~sex+lit+sex*ageC + lit*ageC + wt+ (ageC|id), data = nepal)
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.689381 0.83029
          ageC        0.000427 0.02066  0.07
 Residual             0.224295 0.47360
Number of obs: 882, groups:  id, 197

Fixed effects:
            Estimate Std. Error t value
(Intercept) 13.785990   0.095554  144.27
sex2         0.024052   0.131598    0.18
lit          0.811968   0.322324    2.52
ageC         0.029034   0.003854    7.53
wt           0.009867   0.002751    3.59
sex2:ageC    0.002838   0.005493    0.52
lit:ageC     0.007405   0.015026    0.49
```

those interaction variables

Including more covariates, we expect these to decrease

- Heterogeneity decreases
- This is because we are now accounting for variation between subjects in the fixed effects, i.e., sex, lit, wt
- Mother's literacy and child's weight associated with baseline arm circum.