

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*.

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

1. Longitudinal Data:

E.g., observations y_{ij}

y_{ij} i subject
 j repeated measurement

Repeated measurements (**level-1**), e.g., $j = 1, \dots, r$,

on each subject (i.e., group, **level-2**), e.g., $i = 1, \dots, n$.

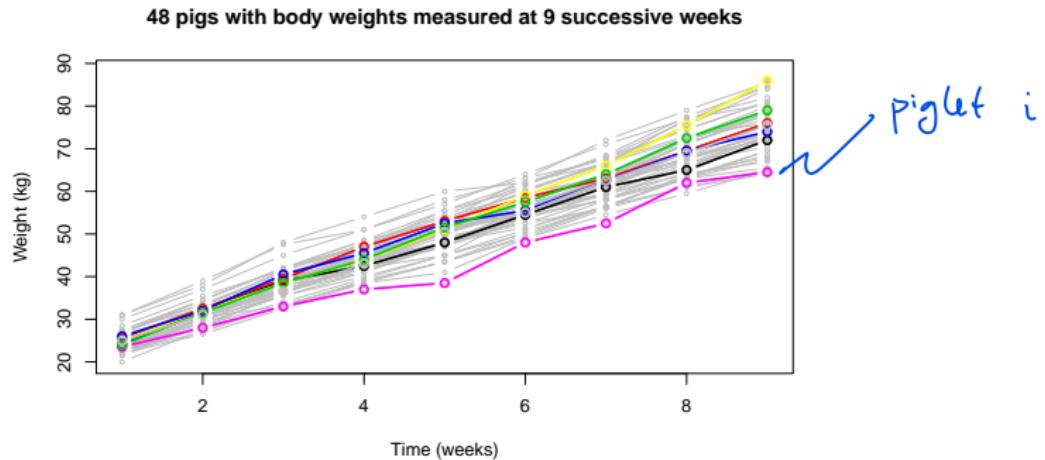
- In a sample of students across years, *annual math score* from each student.
- In a sample of patients, *CD4+ cell counts* of each *HIV patient visit* past seroconversion.

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



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

y_{ij} j^{th} week, week $1, \dots, 9$

Pig Weight Data Structure

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

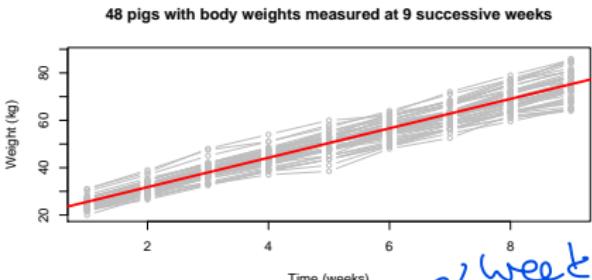
```
> pig[1:13,]  
  id weeks weight  
1  1      1   24.0  
2  1      2   32.0  
3  1      3   39.0  
4  1      4   42.5  
5  1      5   48.0  
6  1      6   54.5  
7  1      7   61.0  
8  1      8   65.0  
9  1      9   72.0  
10 2      1   22.5  
11 2      2   30.5  
12 2      3   40.5  
13 2     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

wide format:

pig-id week1 wk2 ... wk9
|
wt_{ij}
..

4 8

Approach 1: Incorrect approach ignoring clustered structure



Note:

- $\hat{\beta}_1$ UNBIASED and consistent of β_1

Issues:

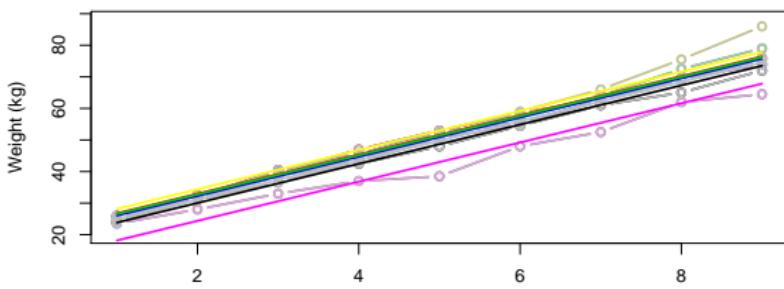
- IID is wrong — errors are not independent
- σ^2 here incorrectly includes within

Limitations: and between pig variability

- Cannot forecast individual pig's growth curve.

Approach 2: Pig-specific Fixed Effects Model

48 pigs with body weights measured at 9 successive weeks



Separate intercept for each pig:

$$y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{IID}}{\sim} N(0, \sigma^2)$$

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 weekly β
• β_{weekly} is same
(here, all data are balanced,
i.e., 9 obs/piglet)

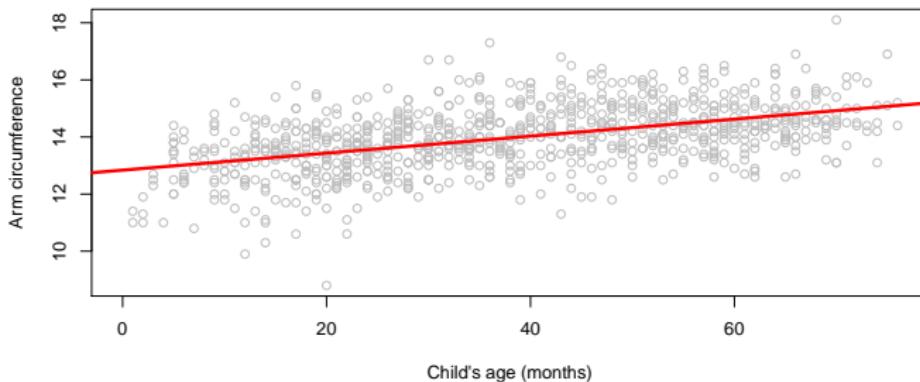
Pig Data: Summary

- 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 is the same for all pigs, turns out that the SE +es

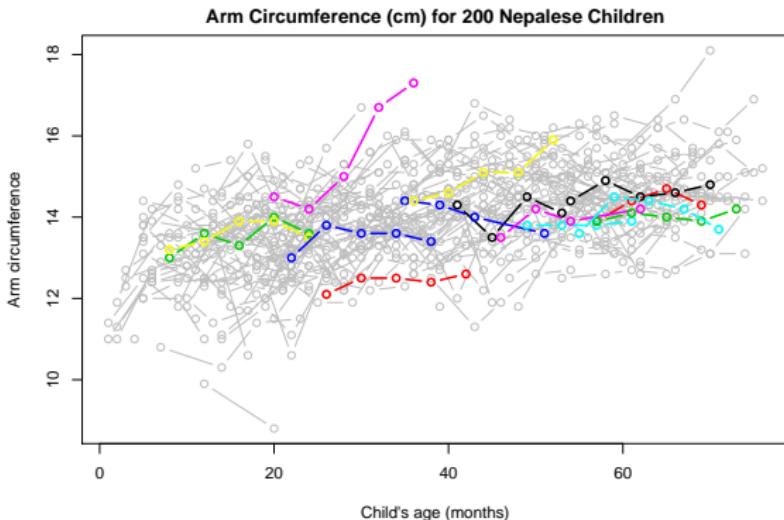
Longitudinal Example: Nepalese Children

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

```
> 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
```

ignores repeated measures
incorrectly assumes iid

```
> 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

$$y_{ij} = \beta_{fid1} + \beta_{fid\ i} + \beta_{age\ j} \\ \dots \text{relative to } fid_1 = 1 \\ + \epsilon_{ij}$$

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

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 the type I error rate is equal to α
(i.e., it's valid to reject H_0 if $p < \alpha$)

Mixed model: Random Intercept

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

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

$\theta_i \stackrel{\text{iid}}{\sim} N(0, \tau^2)$, $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, $\theta_i \perp\!\!\!\perp \varepsilon_{ij}$

- μ = overall intercept (pop. mean when $x_{ij}=0$)
- θ_i = random effect: subject-specific difference from μ
- $\beta_0 i = \mu + \theta_i$ = participant i 's mean
- β : coefficients that do not vary between groups
- τ^2 = random effect variance; between-group variance
- σ^2 = residual variance, within group variability ("measurement error")

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

Mixed model: Random Intercept

↳ random + fixed = mixed

The following two models are equivalent:

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

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: $y_{ij} = \beta_{0i} + \mathbf{x}'_{ij}\beta + \epsilon_{ij} \quad \beta_{0i} \sim N(\mu, \tau^2)$
 $\epsilon_{ij} \sim N(0, \sigma^2)$

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

Assumptions:

- $\epsilon_{ij} \perp\!\!\!\perp \theta_i$ (where $\perp\!\!\!\perp$ = independent) for all i and j .

- θ_i are independent Normal for all i .

$$\theta_i \stackrel{iid}{\sim} N(0, \tau^2)$$

- $\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$ and identically

Properties of the Random Intercept Model

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

- Overall (average) trend:
$$\begin{aligned} E[y_{ij}] &= E[\mu + \theta_i + x'_{ij}\beta] \\ &= E[\mu] + E[\theta_i] + E[x'_{ij}\beta] \\ &= \mu + 0 + x'_{ij}\beta \end{aligned}$$
- Total variability around the overall trend:

$$\text{Var}(y_{ij}) = E[(y_{ij} - E[y_{ij}])^2] = E[\theta_i + \epsilon_{ij}]^2$$

- Conditional (group-specific) trend:
$$\begin{aligned} E[y_{ij} | \theta_i] &= \mu + \theta_i + x'_{ij}\beta \\ &= E[\theta_i] + 2E[\theta_i \epsilon_{ij}] + E[\epsilon_{ij}]^2 \end{aligned}$$

(i.e., individual-specific)

$$E[y_{ij} | \theta_i] = \mu + \theta_i + x'_{ij}\beta$$

- Conditional (within-group) residual variance:

$$\begin{aligned} \text{Var}(y_{ij} | \theta_i) &= E[(y_{ij} - \mu - \theta_i - x'_{ij}\beta)^2] \\ &= E[\epsilon_{ij}]^2 = \sigma^2 \end{aligned}$$

Pig Data Approach 3: Random Intercept Model

R-package : 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 -0.324
```

Compared to a model fitted with group dummy variables, the weeks slope estimate and SE are identical. (because week variable for all piglets, usually not the case)

Nepalese Children: Random Intercept Model

```
> 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

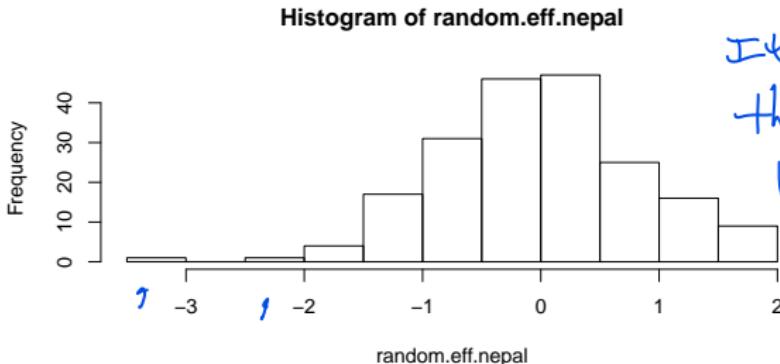
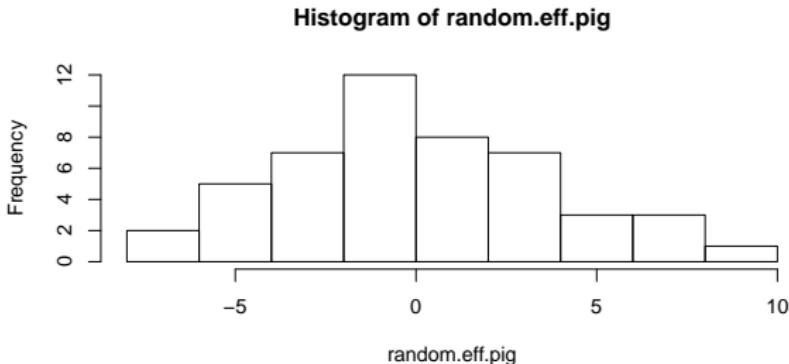
extracting $\hat{\theta}_i$ (BLUPs)

$\xrightarrow{\text{changes relative to fixed effects model}}$

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., different number of pts for different babies*
- We see a decrease in SE of slope of age with mixed model compared to fixed effects model.

Random Effect Histograms



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

$$\begin{matrix} \hat{\tau}_1 \\ \hat{\tau}_2 \end{matrix}$$

Note: this is just $\sqrt{0.78073}$

Fixed effects:

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

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**.

START
HERE
9/11/23

Covariance Structure

- APPROACH 1 (incorrect):
 - ignore repeated measurements
- APPROACH 2: (okay)
 - separate intercept (fixed effect) for each subject
- APPROACH 3: mixed model

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

conditional formulation

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

...

can be re-written as

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

$$\text{Let } \epsilon^* = [\epsilon_{11}^*, \epsilon_{12}^*, \dots, \epsilon_{1r}^*, \epsilon_{21}^*, \dots, \dots]'$$

What is $\text{Cov } \epsilon^*$, or equivalently, $\text{Cov } \underbrace{\mathbf{Y}}_{n \times n \times r}$?

Covariance Structure

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

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

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

Same subject different measurement

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

new subject, so new realization
of measurement error

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

Covariance Structure

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

	ε_{11}^*	ε_{12}^*	ε_{13}^*	ε_{21}^*	ε_{22}^*	ε_{23}^*
ε_{11}^*	$\sigma^2 + \tau^2$	τ^2	τ^2			
ε_{12}^*	τ^2	$\sigma^2 + \tau^2$	τ^2			
ε_{13}^*	τ^2	τ^2	$\sigma^2 + \tau^2$			
ε_{21}^*	0	0	0	$\sigma^2 + \tau^2$	τ^2	τ^2
ε_{22}^*	0	0	0	τ^2	$\sigma^2 + \tau^2$	τ^2
ε_{23}^*	0	0	0	τ^2	τ^2	$\sigma^2 + \tau^2$

$\sim \mathcal{O}$

Random Intercept Model in Matrix Form

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

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

$\underbrace{\mathbf{Z}\boldsymbol{\theta}}$ $\underbrace{\mathbf{X}\boldsymbol{\beta}}$ $\underbrace{\boldsymbol{\epsilon}}$

(Subtract)

where \mathbf{y} is $N \times 1$ vector of response.

\mathbf{Z} is $N \times n$ design matrix of indicator variables for each group.

$\boldsymbol{\theta}$ is $n \times 1$ vector of random intercepts.

\mathbf{X} is $N \times p$ design matrix of fixed effects (including overall intercept).

$\boldsymbol{\beta}$ is $p \times 1$ vector of fixed effects.

$\boldsymbol{\epsilon}$ is $N \times 1$ 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 $\leq \text{ICC} = \rho$

Note that the within-group covariance is

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

So the correlation between observations **within** the same group is

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

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**, $\rho = \text{reproducibility}$.

ICC, cont.

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

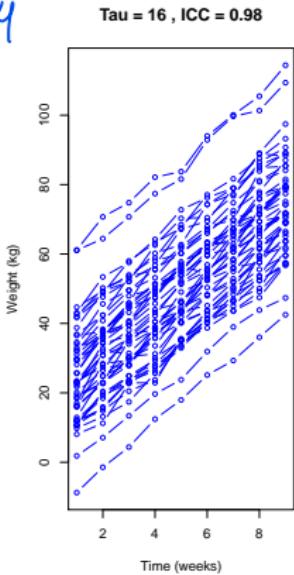
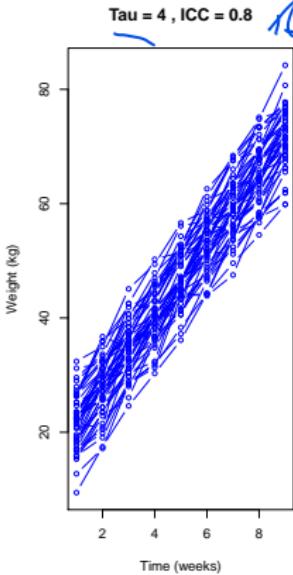
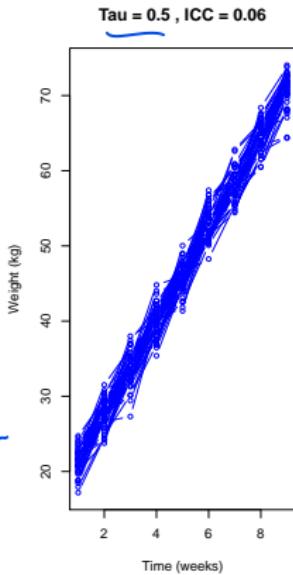
- $\rho \rightarrow 0$ when $\underline{\tau^2 \rightarrow 0}$ (i.e. same intercept).
- $\rho \rightarrow 0$ when $\underline{\sigma^2 \rightarrow \infty}$ (i.e. growing measurement error).
- $\rho \rightarrow 1$ when $\underline{\tau^2 \rightarrow \infty}$ (i.e. large separation in intercepts).
- $\rho \rightarrow 1$ when $\underline{\sigma^2 \rightarrow 0}$ (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

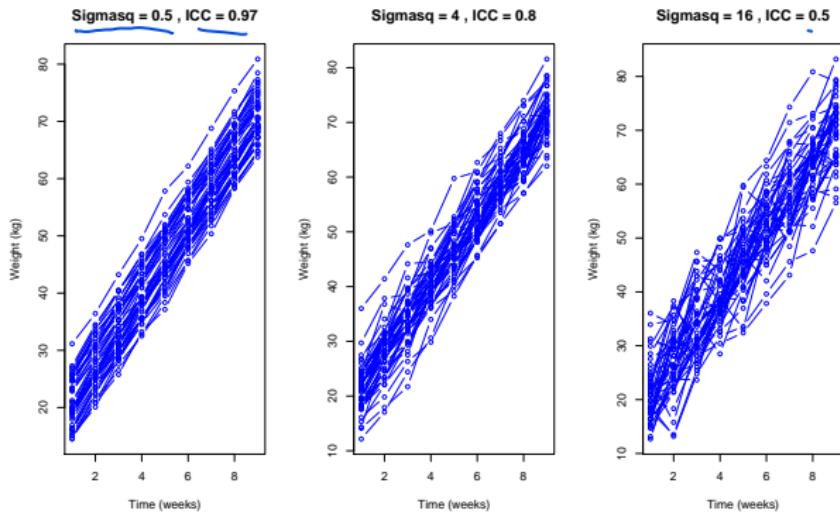
$$y_{ij} = 15 + \theta_i + 6.2 \times \text{weeks}_{ij} + \epsilon_{ij}, \quad \theta_i \sim N(0, \tau^2) \quad \epsilon_{ij} \sim N(0, 4)$$

σ^2_{θ}
 σ^2_{ϵ}



Simulated Pig Data Ex 2 – Measurement error

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



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:

random effects model

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

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

$$\begin{aligned} P(X, Y) &= \\ P(X|Y)P(Y) & \end{aligned}$$

$$\begin{aligned} \prod_{i,j} f(y_{ij}, \theta_i) &= \prod_{i,j} \underbrace{f(y_{ij}|\theta_i)}_{\text{Drop terms}} \times \prod_i g(\theta_i) \\ &\propto \exp \left[-\frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - \theta_i)^2 \right] \times \exp \left[-\frac{1}{2\tau^2} \boldsymbol{\theta}' \boldsymbol{\theta} \right] \\ &= \exp \left[-\frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - \theta_i)^2 - \frac{1}{2\tau^2} \boldsymbol{\theta}' \boldsymbol{\theta} \right] \\ &= \exp \left[-\frac{1}{2\sigma^2} \left[\sum_{i,j} (y_{ij} - \theta_i)^2 + \frac{\sigma^2}{\tau^2} \boldsymbol{\theta}' \boldsymbol{\theta} \right] \right] \end{aligned}$$

(drop terms
that are equal
to 0 when take
derivative
wrt θ)

Shrinkage and Random Effects

Then maximizing the log likelihood is equivalent to

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

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

Consider the matrix formulation

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

$$\begin{aligned} Z'y &= (Z'Z + \frac{\sigma^2}{\tau^2} I)\theta \\ (Z'Z + \frac{\sigma^2}{\tau^2} I)^{-1}Z'y &= \hat{\theta} \end{aligned}$$

where $\mathbf{Z} \in \mathbb{R}^{nr \times n}$ design matrix of indicator variables denoting the ij th 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:

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

Shrinkage and Random Effects, cont.

This is equivalent to

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

Note that (recall here, we have "where pop. mean = 0") a special case

- $\hat{\theta}_i \rightarrow 0$ when $\tau^2 \rightarrow 0$ (i.e. shrinks all random intercepts to zero).
- $\hat{\theta}_i \rightarrow \bar{y}_i$ when $\tau^2 \rightarrow \infty$ (i.e. no shrinkage = raw group mean estimates) (i.e., participant)
- $\hat{\theta}_i \rightarrow \bar{y}_i$ when $\sigma^2 \rightarrow 0$ (i.e. no shrinkage = raw group mean estimates).
- $\hat{\theta}_i \rightarrow \bar{y}_i$ when $r \rightarrow \infty$ (i.e. no shrinkage = raw group mean estimates) number of observations on 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

SKIP

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

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

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

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

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

EDF \rightarrow 48 (less shrinkage) when:

- $\sigma^2/\tau^2 \rightarrow 0$
- 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 \rightarrow 0 (more shrinkage) when:

- $\sigma^2/\tau^2 \rightarrow \infty$
- 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

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 $\text{EDF} \rightarrow n$ 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 :

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

The joint density of the data and random effects is then

~~JKM~~

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

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

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

Shrinkage and Borrowing Information, cont.

The joint density is a multivariate Normal density:

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

where $\mathbf{A} = (\mathbf{Z}'\mathbf{Z} + \frac{\sigma^2}{\tau^2} \mathbf{I})$ 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:

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

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

• for $\theta_i \sim N(\mu, \tau^2)$, $\hat{\theta}_i = \frac{(\widehat{\sigma^2}/\tau^2)\mu + \sum_{j=1}^{r_i} y_{ij}}{r_i + \sigma^2/\tau^2}$.
 $\hat{\theta}_i$ borrows information from the population

Note that

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

i.e., τ^2 large relative to σ^2

Shrinkage and Borrowing Information, cont. ii

We can also express $\hat{\theta}_i$ as
$$\hat{\theta}_i = \frac{(1/\tau^2)\mu + (r_i/\sigma^2)\bar{y}_i}{1/\tau^2 + (r_i/\sigma^2)}$$

with annotations:
 $\underbrace{(1/\tau^2)\mu}_{\text{inverse of inter-subject variance}}$ + $\underbrace{(r_i/\sigma^2)\bar{y}_i}_{\text{within subject variance}}$.

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

1. standard estimate without penalization: \bar{y}_i .
2. overall mean μ .

with their corresponding **inverse-variances** as weights!

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

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

with subject average

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

Best Linear Unbiased Prediction

For the random intercept model

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

we wish to estimate the **unobserved random variable** θ_i .

We can also derive the estimators using the MVN distribution. Assume τ^2 and σ^2 are known. Then

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

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

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

Best Linear Unbiased Prediction: BLUPs

SKIP

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

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, $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_{ij}$, 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).

SLIP

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

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

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

However, the variances are smaller.

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

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

BLUPs: Matrix formulation

skip

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

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

Then

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

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

In practice, replace β , σ^2 , and τ^2 by their estimates.

X

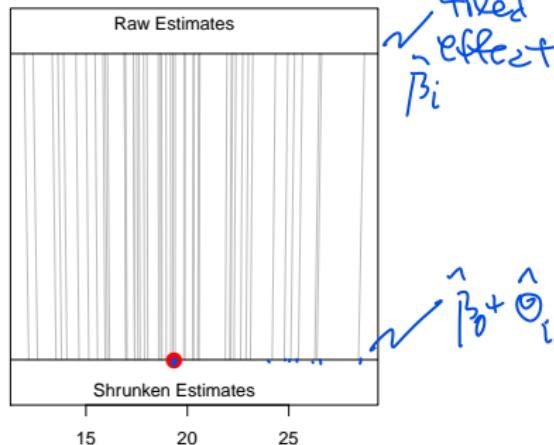
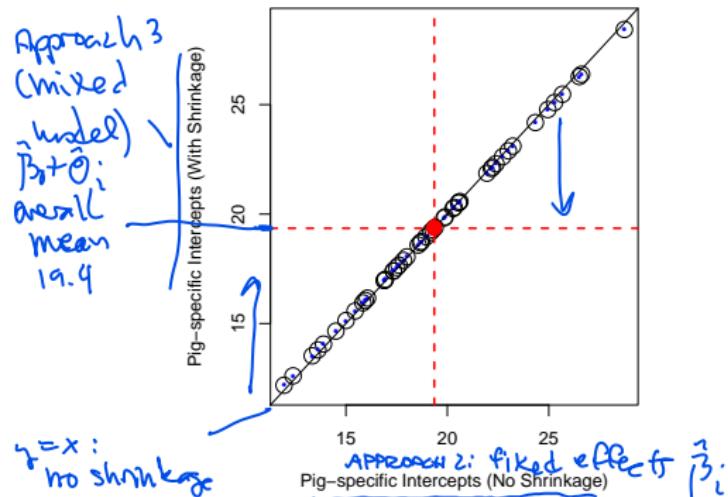
Shrinkage: Pig Data

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

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

"good"

APPROACH 2
fixed
effects
 β_i

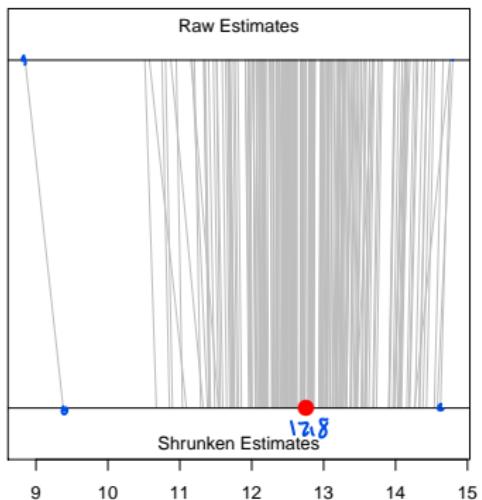
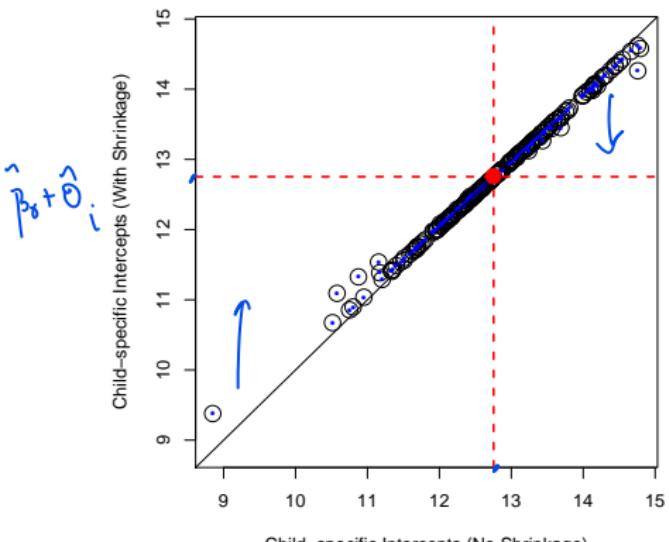


Shrinkage: Nepalese Data

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

$$\hat{\tau}^2 = 0.78, \quad \hat{\sigma}^2 = 0.25, \quad \hat{\sigma}^2 / \hat{\tau}^2 = 0.32, \quad \text{ICC} = 0.76, \quad r_i \in \{1, \dots, 5\}$$

r_i smaller



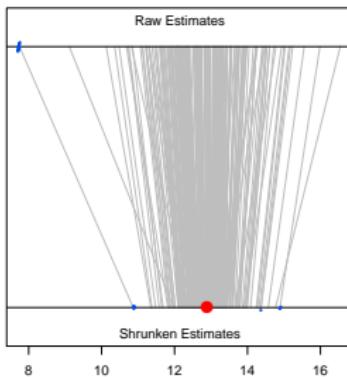
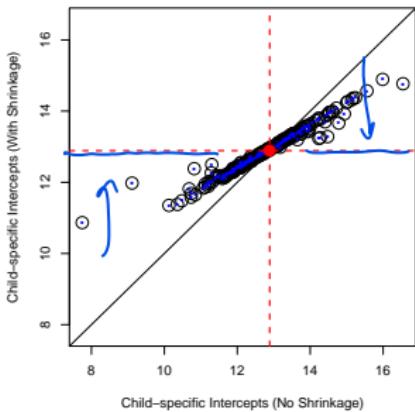
Shrinkage: Nepalese Data with Noise

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

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

where $\underline{\text{armc}^*} = \text{arm} + \underline{N(0, 2)}$. *added more movement error*

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



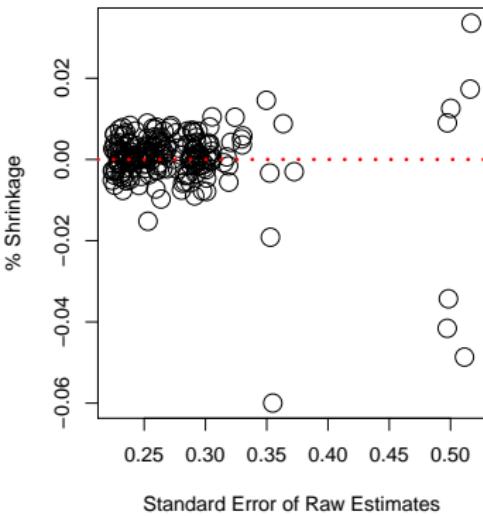
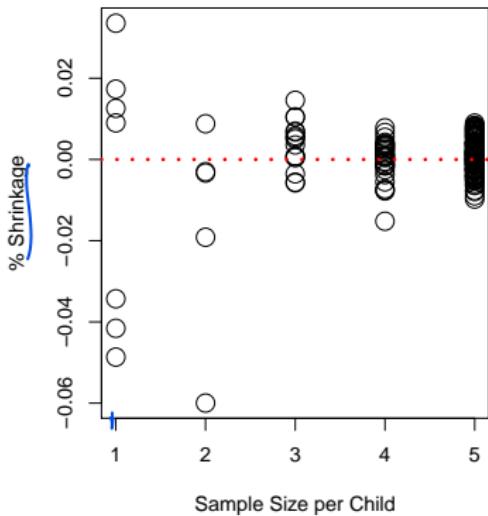
Shrinkage and Borrowing Information, cont. iii

The Nepalese children dataset contains missing data. Not all 200 children have complete 5 visits.

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

$$\frac{r_1}{\sigma^2}$$

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.

(amount of shrinkage)

Effects of the normality assumption on random effects depend on

1. group-specific sample size, $\sqrt{n_i}$
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 effects fixed effects

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

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

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

We know $Cov(\boldsymbol{\epsilon}^*) = \mathbf{Z}Cov(\boldsymbol{\theta})\mathbf{Z}' + Cov(\boldsymbol{\epsilon})$

$$= \tau^2 \mathbf{Z}\mathbf{Z}' + \sigma^2 \mathbf{I} \quad \left. \begin{array}{l} \text{(Equivalent to matrix} \\ \text{if } \approx \text{ SLIDE 24/25)} \end{array} \right\}$$

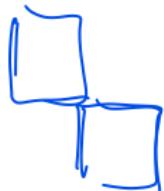
This is equivalent to integrating out the random effects. Then the marginal model is:

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

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

off-diagonal within subject: τ^2

off-block (between subjects): 0



Generalized Least Squares

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

$$\arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)' \underbrace{\mathbf{V}^{-1}}_{\text{no longer } \sigma^2 \mathbf{I}} (\mathbf{y} - \mathbf{X}\beta) \quad \text{for } \tilde{\mathbf{y}} \in \mathbb{R}^N$$

This is also called weighted least squares.

Note this is the kernel of the multivariate normal distribution.

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

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

$$f(\tilde{\mathbf{y}}) = (2\pi)^{-N/2} |\mathbf{V}|^{1/2} e^{-\frac{1}{2}(\tilde{\mathbf{y}} - \tilde{\mathbf{x}}'\beta)' \mathbf{V}^{-1} (\tilde{\mathbf{y}} - \tilde{\mathbf{x}}'\beta)}$$

"Kernel" of
multivariate
normal

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

$$N = \sum_{i=1}^n r_i, \quad r_i = \# \text{obs. on } i^{\text{th}} \text{ subject}$$

Parameter Estimation: Maximum Likelihood Approach

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

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

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

subst. tkt

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 ($\boldsymbol{\beta}$) 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)

$$l(\sigma^2, \tau^2) - \frac{1}{2} \log |\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}|$$

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 $\mathbf{X}_i \stackrel{iid}{\sim} N(0, \sigma^2)$, we have

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

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

Small samples: often prefer REML.

Likelihood ratio tests and AIC: use ML.

Default, a little more conservative

Parameter Estimation: MLE versus REML

library(lme4)

```
> 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

$$\begin{array}{l} \xrightarrow{\hat{\tau}^2} \\ \xrightarrow{\hat{\sigma}^2} \end{array}$$

Fixed effects:

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

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

```
> 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

smaller than REML

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

Note that the standard errors are larger for REML.

Fixed versus Random

Consider the model:

$$y_{ij} = \theta_i + \beta' x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$$

θ_i : 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, $\theta_i \stackrel{iid}{\sim} N(0, \tau^2)$.

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

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)

think of θ_i as fixed

all that's left is σ^2 \uparrow within-subject variance

Approach 2 Approach 3

Guidelines for choosing fixed vs random

- Are we interested in predicting subject effects?

/ RE leverage population information — lower prediction error

- If the experiment were repeated, would the same subjects (i.e., groups) be used?

→ FE

- Or are the subjects a random sample from a population of interest?

→ RE

- Are there enough subjects to estimate heterogeneity?

E.g., if only 2 subjects, do fixed effects

- Are there enough repeated measurements to estimate FE?

↳ if only have 2 obs/subject, consider RE

- Do some subjects have only 1 observation and/or is there different number of samples for each subject?

→ RE

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

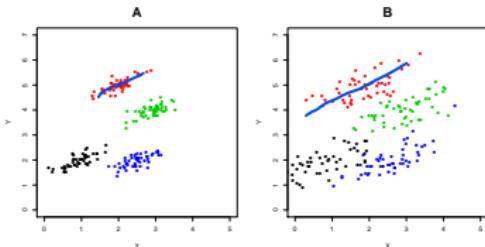
- Clustered / correlated / grouped / longitudinal / multi-level / hierarchical / nested data
- Random effect / (~~Bayesian~~) hierarchical / mixed / variance component model
- Between-group variability / heterogeneity / structured error
- Within-group correlation / intraclass correlation
- Within-group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

*borrow info from pop. to improve
subject-level predictions*

Quiz 3

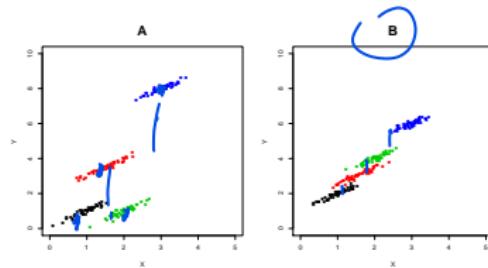
$$y_{ij} = \mu + \theta_i + \epsilon_{ij}, \quad \underline{\theta_i \sim N(0, \tau^2)} \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad \frac{\tau^2}{\tau^2 + \sigma^2}$$

Part I



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

Part II

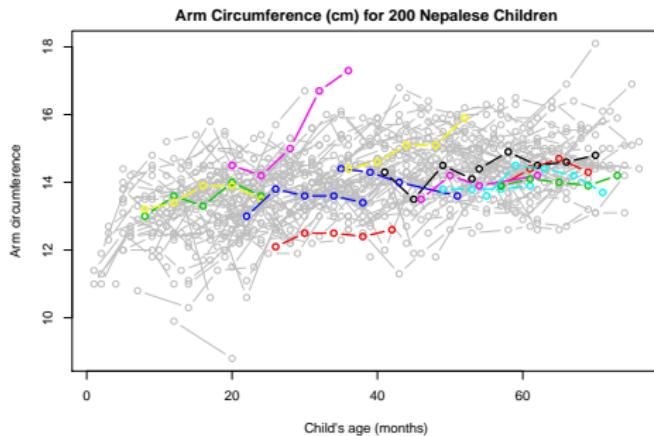


Which plot:

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

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.

$$arm_{ij} = \beta_0 + \theta_{0i} + (\beta_1 + \theta_{1i})ageC_{ij} + \varepsilon_{ij}$$

$\text{corr}(\theta_{0i}, \theta_{1i}) = \rho$
 $\text{cov}(\theta_{0i}, \theta_{1i}) = \rho\tau_1\tau_2$, $\text{cov}(\theta_{0i}, \varepsilon_{ij}) = 0$

$$\begin{bmatrix} \theta_{0i} \\ \theta_{1i} \end{bmatrix} \stackrel{\text{iid}}{\sim} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right), \quad \varepsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$\theta_{0i}, \theta_{1i} \perp \varepsilon_{ij}$

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_{0i} and θ_{1i} are **bivariate normal**.

- τ_1^2 describes between-children variation in baseline arm circumferences at at age three. variance of random
- τ_2^2 describes between-children variation in the linear effects of age. 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

lmer

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
```

→ $\text{ageC} | \text{id} \rangle$:
it fits random intercept, random slope, and correlation between them (ρ)

Population distribution of random intercepts and slopes:

- Child-specific intercept: $\beta_0 + \theta_{0i}$ arm circumference of i^{th} child at 36 months
- Child-specific slope: $(\beta_1 + \theta_{1i}) \sim N(0.0325, 0.0043\dots)$

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. $0.0325 \pm 2 \cdot 0.00209$

- $\rho = \text{cor}(\beta_{0i}, \beta_{1i}) = 0.09$

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

→ 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: ≥ 2 difference, "substantially better"

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

$\log \text{likelihood}$

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

\hookrightarrow number of covariates plus
 \hookrightarrow number of variance components

where $\ell(\theta)$ is the log likelihood for all parameters θ 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. $\text{anova}(\text{reduced}, \text{full})$ LRT]

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

when specifying `lme`, set $\text{REML} \leftarrow \text{false}$

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^2_1 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 finite sample null distribution

Compare to model without random slope

fits a random slope and random intercept along their correlation
 (τ_1^2, τ_2^2, p)

```
> 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 (lower is better) → 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

Fixed effect model

An alternative framework would treat the intercepts as fixed.

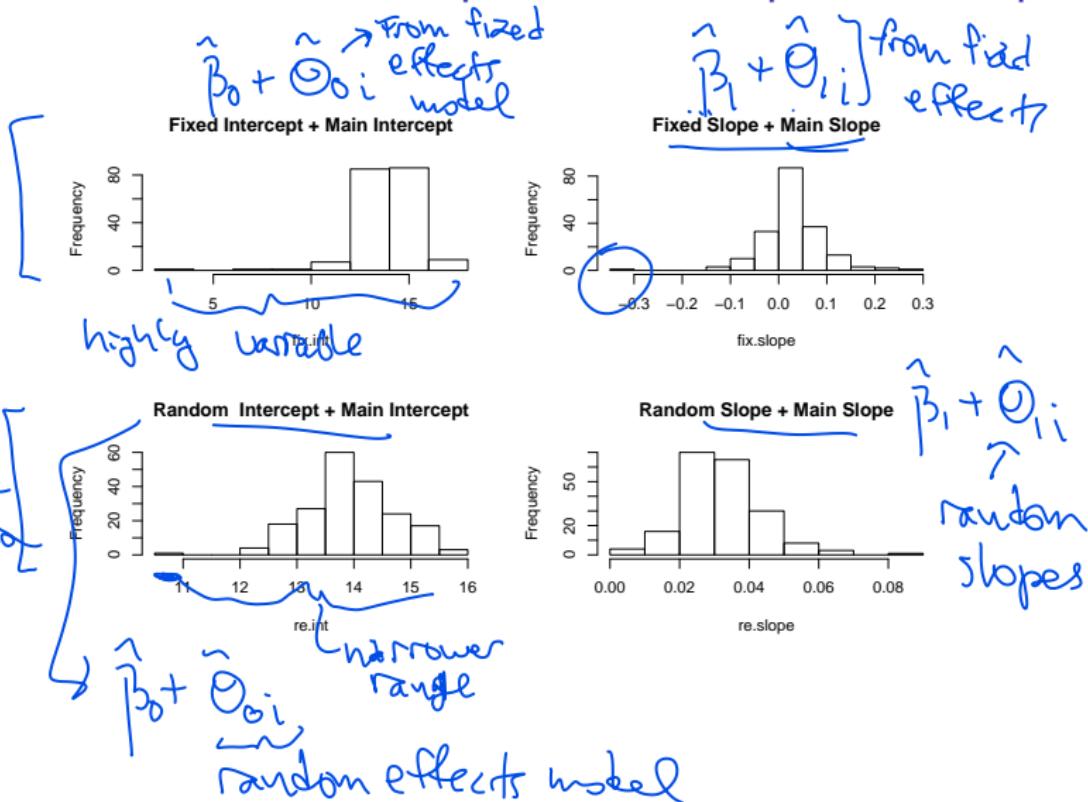
$$\text{arm}_{ij} = \beta_0 + \theta_{0i} + (\beta_1 + \theta_{1i}) \text{ageC}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$$

We can use the sum-to-zero contrasts:

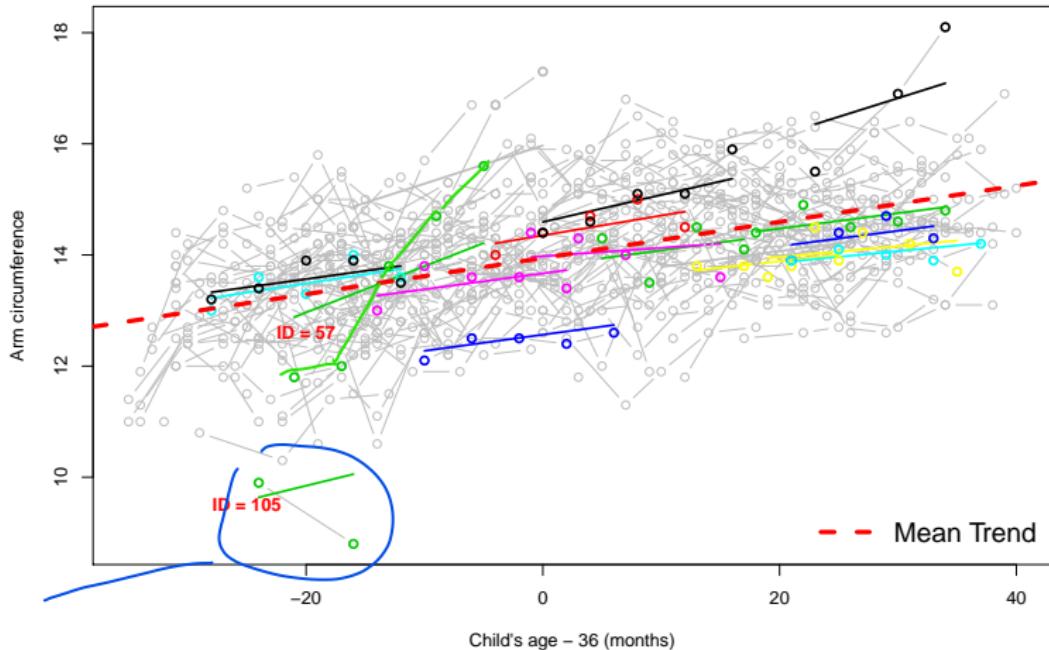
$$\sum_{i=1}^n \hat{\theta}_{0i} = 0 \quad \text{slopes sum to 0}$$

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



Mixed Model



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

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	

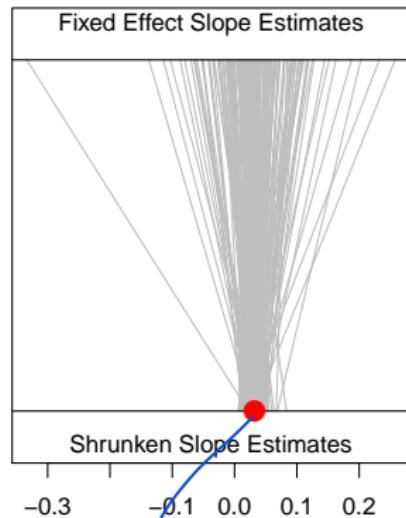
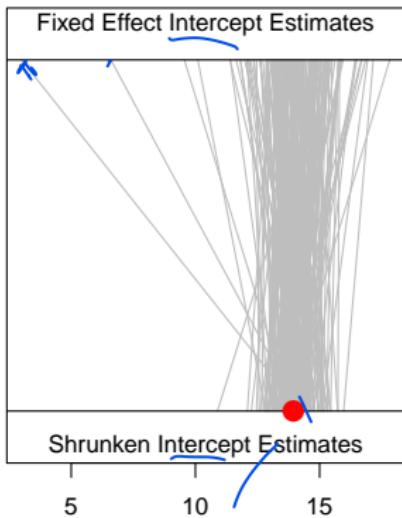
Fixed effects:

	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:
 - Intercept: 0.848 → 0.817.
 - AgeC: 0.021 → 0.017.

$\hat{\pi}^2$
 $\hat{\beta}^2$, in HW2, we will look at some more diagnostic plots

Shrinkage!



leverage
population
information

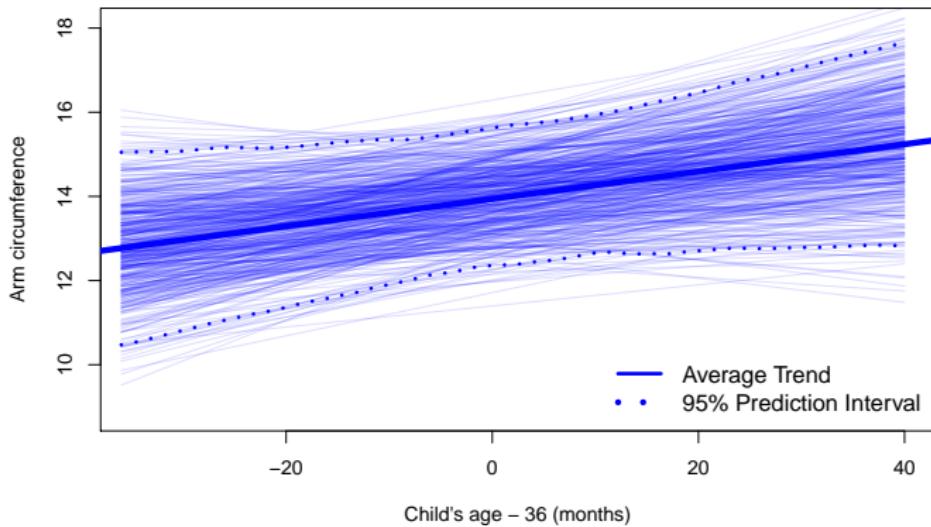
$$\hat{\beta}_0 + \hat{\theta}_{0i}$$

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

Simulating Out-of-Sample Growth Curves

$$y_{ij} = (13.94 + \hat{\beta}_0) + (0.0325 + \hat{\beta}_1) x_{ij} + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, 0.48^2)$$
$$\left\{ \begin{bmatrix} \theta_{0i} \\ \theta_{1i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.85^2 & 0.09 \times 0.85 \times 0.021 \\ 0.09 \times 0.85 \times 0.021 & 0.021^2 \end{bmatrix} \right) \right.$$

Predicted Growth Curves for 500 Children



Hierarchical Formulation

Consider the following model now including interactions:

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

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

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

$$\text{arm}_{ij} = \underbrace{\beta_{0i}}_{\dots} + \underbrace{\beta_{1i}}_{\dots} \text{ageC}_{ij} + \epsilon_{ij} \quad (5)$$

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

Equation (5) can also be written as:

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

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

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

(earlier called this β_0)

(earlier called the β_i)

Hierarchical Formulation: Back to Random Intercepts

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

$$\begin{aligned} \text{arm}_{ij} &= \beta_0 + \theta_{0i} + \beta_1 \text{ageC}_{ij} + \beta_2 \text{lit}_{ij} + \epsilon_{ij} \\ \theta_{0i} &\stackrel{iid}{\sim} N(0, \tau^2), \quad \epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

lit_{ij} : i^{th} child's mother's literacy

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

$$\theta_{0i} \sim N(\beta_0 + \beta_2 \text{lit}_i, \tau^2)$$

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

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 and sex

Level 1: $\underline{B}_{0i} = \mu_0 + \alpha_{01} \text{lit}_i + \alpha_{02} \text{sex}_i + \eta_{0i}$ random
 $\underline{B}_{1i} = \mu_1 + \alpha_{11} \text{lit}_i + \alpha_{12} \text{sex}_i + \eta_{1i}$ slope
 $\rightarrow \text{age}_{1ij}, \text{weight}_{1ij}$ capture interactions with age_{1ij}

Level 2: $\underline{\text{arm}}_{1ij} = \underline{B}_{0i} + \underline{B}_{1i} \text{age}_{1ij} + \beta_2 \text{weight}_{1ij} + \varepsilon_{1ij}$

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:

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

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

Level 2:

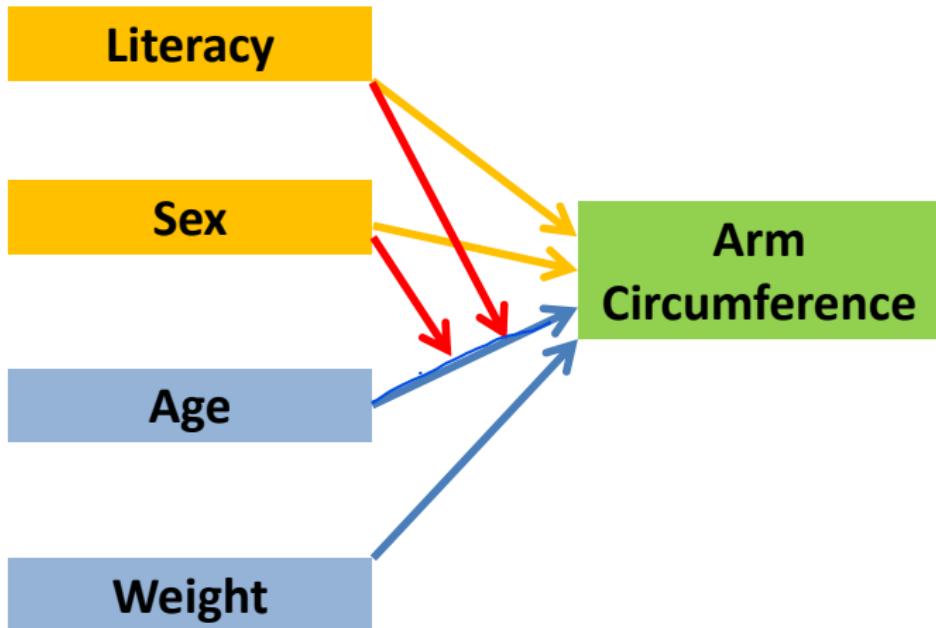
$$arm_{ij} = \beta_{0i} + \beta_{1i} ageC_{ij} + \beta_2 weightC_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$$

By substituting Level 1 regressions into Level 2:

$$\begin{aligned} arm_{ij} &= \mu_0 + \alpha_{01}lit_i + \alpha_{02}sex_i + \theta_{0i} \\ &\quad + (\mu_1 + \alpha_{11}lit_i + \alpha_{12}sex_i + \theta_{1i}) ageC_{ij} + \beta_2 weightC_{ij} + \epsilon_{ij} \\ &= \mu_0 + \alpha_{01}lit_i + \alpha_{02}sex_i + \theta_{0i} \\ &\quad + \mu_1 ageC_{ij} + \alpha_{11}lit_i \times ageC_{ij} + \alpha_{12}sex_i \times ageC_{ij} + \theta_{1i} ageC_{ij} \\ &\quad + \beta_2 weightC_{ij} + \epsilon_{ij}, \\ &\quad [\theta_{0i}, \theta_{1i}]' \stackrel{iid}{\sim} N(\mathbf{0}, \Sigma), \quad \epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

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

Cross-level Interactions



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
```

including more covariates,
we expect these
to decrease

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
<u>sex2:ageC</u>	0.002838	0.005493	0.52
<u>lit:ageC</u>	0.007405	0.015026	0.49

- 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.