Random Effects

Edps/Psych/Soc 587

Carolyn J. Anderson

Department of Educational Psychology
ILLINOIS

Outline

- Introduction
- Empirical Bayes inference
- Henderson's mixed-model equations
- BLUP: Best Linear Unbiased Prediction
- Shrinkage
- The normality assumption for random effects
- SAS/MIXED and R

Snijders & Bosker: pp 161-172
Further reference: Verbeke & Molenberghs (Chapter 7) and therein.
(These notes are based primarily on the latter.)

Introduction

Why get estimate of the random effects, $\hat{\boldsymbol{U}}_j$?

- See how much groups (macro units) deviate from the average regression.
- Detect outlying groups.
- Predict group specific outcomes.

In education might be tempted to use these to

- Select a specific school for your child to attend (see Snijders & Bosker).
- Hold schools/teachers/students accountable.

Empirical Bayes Inference

- Bayes Theorem
- Empirical Bayes Estimates
- Example

Bayes Theorem

Based on probability theory, for two discrete variables, A and B, :

$$\text{joint probability} = \text{(conditional prob)(marginal prob)}$$

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes Theroem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{or} \quad P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayes Theorem: Continuous Variables

Replace probabilities (i.e., $P(.)$) by probability density functions, i.e., $f(.)$.

For two continuous variables, x and y,

$$f(x,y) = f(x|y)f(y) = f(y|x)f(x)$$

and Bayes Theroem:

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

or

$$f(y|x) = \frac{f(x|y)f(y)}{f(x)}$$

Bayes Theorem: Continuous Variables

We have vectors of continuous random variables:

- $\boldsymbol{y}_j$, observations/measures on the response variable from individuals in group j.
- $\boldsymbol{U}_j$, our random effects for group j.

The relationships that hold for two single variables also hold for sets of random variables

$$f(\boldsymbol{y}_j, \boldsymbol{U}_j) = f(\boldsymbol{y}_j|\boldsymbol{U}_j)f(\boldsymbol{U}_j) = f(\boldsymbol{U}_j|\boldsymbol{y}_j)f(\boldsymbol{y}_j)$$

and

$$f(\boldsymbol{U}_j|\boldsymbol{y}_j) = \frac{f(\boldsymbol{y}_j|\boldsymbol{U}_j)f(\boldsymbol{U}_j)}{f(\boldsymbol{y}_j)}$$

Empirical Bayes Estimates

$$f(\boldsymbol{U}_j|\boldsymbol{y}_j) = \frac{f(\boldsymbol{y}_j|\boldsymbol{U}_j)f(\boldsymbol{U}_j)}{f(\boldsymbol{y}_j)}$$

where

- $f(\boldsymbol{U}_j)$ is the "prior" distribution of the random effects.

  This distribution is $\mathcal{N}(\mathbf{0}, \boldsymbol{T})$.

- $f(\boldsymbol{y}_j)$ is the marginal distribution of the response variable,

$$\boldsymbol{Y}_j \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$$

Empirical Bayes Estimates (continued)

- $f(\boldsymbol{y}_j|\boldsymbol{U}_j)$ is the conditional distribution of the response variable given the random effects,

$$f(\boldsymbol{y}_j|\boldsymbol{U}_j) \quad \sim \quad \mathcal{N}((\boldsymbol{X}_j\boldsymbol{\Gamma} + \boldsymbol{Z}_j\boldsymbol{U}_j), \sigma^2\boldsymbol{I})$$

- $f(\boldsymbol{U}_j|\boldsymbol{y}_j)$ is the "posterior" distribution of the random effects; it's the distribution of the random effects conditional on the data (our observations on our response variable and our estimated parameters, $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{T}}$).

Empirical Bayes Estimates (continued)

$$
\begin{aligned}
f(\boldsymbol{U}_j|\boldsymbol{y}_j) &= \frac{f(\boldsymbol{y}_j|\boldsymbol{U}_j)f(\boldsymbol{U}_j)}{f(\boldsymbol{y}_j)} \\
&\propto f(\boldsymbol{y}_j|\boldsymbol{U}_j)f(\boldsymbol{U}_j)
\end{aligned}
$$

Replacing everything on the right hand side of this equation gives the distribution for $f(\boldsymbol{U}_j|\boldsymbol{y}_j)$, which is multivariate normal. (i.e., conjugate of normal is normal).

This is what we need to get estimates/predictions of $\boldsymbol{U}_j$.

Aside: In MCMC, we sample from

$$f(\boldsymbol{U}_j|\boldsymbol{y}_j) \propto f(\boldsymbol{y}_j|\boldsymbol{U}_j)f(\boldsymbol{U}_j)$$

Estimating U_j

One way, estimate the mean of the distribution of the "posterior" distribution, $f(\boldsymbol{U}_j|\boldsymbol{y}_j)$.

By the definition of the mean and algebra,

$$\begin{aligned}
\widehat{\boldsymbol{U}}_{j(\boldsymbol{\Gamma},\boldsymbol{T})} &= \mathsf{E}(\boldsymbol{U}_j|\boldsymbol{Y}_j = \boldsymbol{y}_j) \\
&= \int \boldsymbol{U}_j f(\boldsymbol{U}_j|\boldsymbol{y}_j) d\boldsymbol{U}_j \\
&= \boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma})
\end{aligned}$$

We can get a closed form (another very nice thing resulting for using normal distributions).

Estimating U_j (continued)

$$\hat{U}_{j(\boldsymbol{\Gamma},\boldsymbol{T})} = \boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma})$$

- The integration is over all possible values of $\boldsymbol{U}_j$ (like a sum).
- $\boldsymbol{V}_j = (\boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_j' + \sigma^2\boldsymbol{I})$.
- $(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma}) = (\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j)$.
- The estimate of $\boldsymbol{U}_j$ depends on $\boldsymbol{\Gamma}$ and $\boldsymbol{T}$ (model), as well as on $\boldsymbol{y}_j$ (data).

Estimating Covariance Matrix for U_j

- The covariance matrix for $\hat{\boldsymbol{U}}_{j(\boldsymbol{\Gamma},\boldsymbol{T})}$ is

$$\text{var}(\hat{\boldsymbol{U}}_j) = \boldsymbol{T}\boldsymbol{Z}_j' \left[\boldsymbol{V}_j^{-1} - \boldsymbol{V}_j^{-1}\boldsymbol{X}_j \left(\boldsymbol{X}_j \boldsymbol{V}_j^{-1} \boldsymbol{X}_j' \right)^{-1} \boldsymbol{X}_j' \boldsymbol{V}_j^{-1} \right] \boldsymbol{Z}_j \boldsymbol{T}.$$

  This covariance matrix isn't good for statistical inference about $(\hat{\boldsymbol{U}}_{j(\boldsymbol{\Gamma},\boldsymbol{T})} - \boldsymbol{U}_j)$, because it underestimates the variability of $(\hat{\boldsymbol{U}}_{j(\boldsymbol{\Gamma},\boldsymbol{T})} - \boldsymbol{U}_j)$... it ignores the variability of $\boldsymbol{U}_j$.

- For statistical inference, the variance of $(\hat{\boldsymbol{U}}_{j(\boldsymbol{\Gamma},\boldsymbol{T})} - \boldsymbol{U}_j)$ is used,

$$\text{var}(\hat{\boldsymbol{U}}_j - \boldsymbol{U}_j) = \boldsymbol{T} - \text{var}(\hat{\boldsymbol{U}}_j).$$

  When $\boldsymbol{\Gamma}$ and $\boldsymbol{T}$ are unknown and we use estimates of them, the estimate of $\boldsymbol{U}_j$ is known as the Empirical Bayes (EB) estimate, $\hat{\boldsymbol{U}}_j$.

Example: HSB

A simple model.

$$(\text{math})_{ij} = \gamma_{00} + \gamma_{10}(\text{cSES})_{ij} + U_{0j} + R_{ij}$$

SAS/MIXED input:

```
PROC MIXED data=hsbcent noclprint covtest method=ML ic;
   CLASS id;
   MODEL mathach = cSES /solution;
   RANDOM intercept / subject=id type=un solution cl
      alpha=.05;
   ODS output SolutionR=RanUs;
```

SAS/MIXED Notes

- " solution" option in the RANDOM statement tells SAS to compute $\hat{\boldsymbol{U}}_j$'s.
- The "ODS" command ("Output Delivery System") replaces the "MAKE" command in earlier versions of SAS (MAKE can still be used).
  - "SolutionR" is an ODS table name that refers to the random effects solution vectors.
  - "RanUs" is the name of the SAS data set output that contains the contents of the random effect solution vectors.
- For other ODS table names, see SAS 9.4 PROC MIXED online documentation.

Getting Us for random Intercept Model

```
summary( model.1 ← lmer (mathach   cSES + (1 |id),
          data=hsb,REML=FALSE) )

ranU ← ranef(model.1)
df.1 <- as.data.frame(ranU)
head(df.1)
```

	grpvar	term	grp	condval	condsd
1	id	(Intercept)	1224	-2.6753498	0.84934807
2	id	(Intercept)	1288	0.7470468	1.1238806
3	id	(Intercept)	1296	-4.5904308	0.8411883
4	id	(Intercept)	1308	2.9791973	1.2341147
5	id	(Intercept)	1317	0.4962998	0.8411883
6	id	(Intercept)	1358	-1.2514746	1.0387323

Getting Us for random Intercept Model

Note:

- `condval` = Conditional value = $\hat{U}_{0j}$
- `condsd` = Standard deviation of conditional value

Getting Us for Complex Model

Note: To deal with "singular", xyses = cses/10

```
model.3 ← lmer(mathach ∼ 1 + xyses + female + meanses
       + (1 + xyses + female | id),
       data=hsb,REML=FALSE,
       control = lmerControl(optimizer ="Nelder_Mead"))

# The following is a very long data frame:  n++ × 3)
U ← as.data.frame(ranef(model.3))

# Pull out what we want
Uoj ← U[which(U$term=="(Intercept)"), ]
U1j ← U[which(U$term=="xyses"), ]
U2j ← U[which(U$term=="female"), ]
```

SAS/MIXED: Output

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	id	8.6071	1.0682	8.06	$< .0001$
Residual		37.0056	0.6245	59.26	$< .0001$

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr$>\|t\|$
Intercept	12.6494	0.2437	159	51.92	$< .0001$
cses	2.1912	0.1086	7024	20.17	$< .0001$

SAS: Solution for Random Effects

Solution for Random Effects

Effect	school	Estimate	Std Err Pred	DF	t Value	Pr> $\|t\|$
Intercept	1224	−2.6752	0.8782	7024	−3.05	.0023
Intercept	1288	0.7470	1.1429	7024	0.65	.5134
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Solution for Random Effects

Effect	school	Alpha	Lower	Upper
Intercept	1296	0.05	−6.2964	−2.8840
Intercept	1308	0.05	0.5280	5.4297
⋮	⋮	⋮	⋮	⋮

In R

- model.1 ← lmer (mathach ∼ cSES + (1 | id), data=hsb, REML=FALSE)
- summary(model.1)

Random effects:

Groups	Name	Variance	Std.Dev.
id	(Intercept)	8.612	2.935
Residual		37.005	6.083
Number of obs: 7185, groups: id, 160			

Fixed effects:

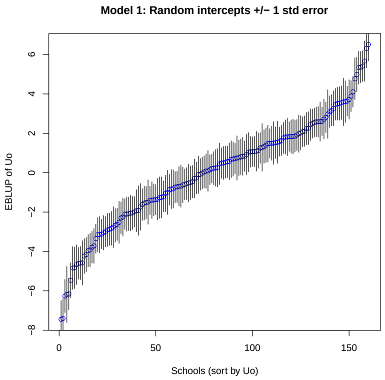
	Estimate	Std. Error	df	t value	$\Pr(>\|t\|)$
(Intercept)	12.65	0.24	157.72	51.90	0.00
cSES	2.19	0.11	7023.02	20.17	0.00

In R

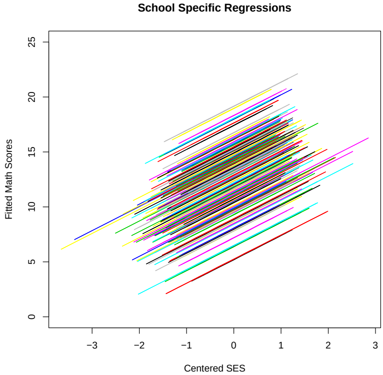
```
ranef(model.1)
```

$id	(Intercept)
1224	-2.67534979
1288	0.74704676
1296	-4.59043077
1308	2.97919727
1317	0.49629979
1358	-1.25147457
1374	-2.52157850
1433	6.30842737
1436	4.98842414
1461	3.72090916

Plotting U_{0j}



School Specific Regressions



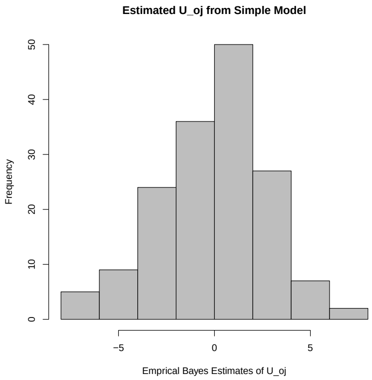
Solution for Random Effects

- "Estimate" are the empirical Bayes estimates.

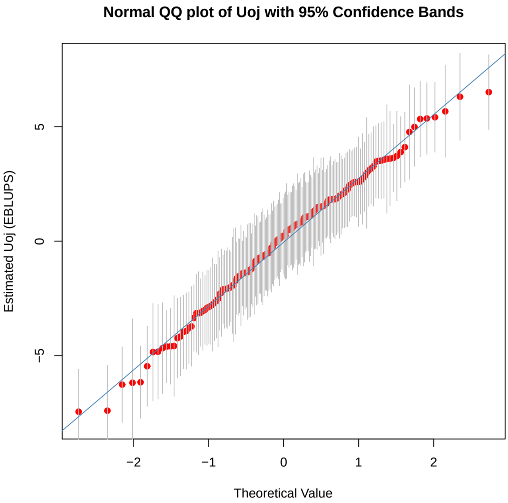
- The "Std Err Pred" are the standard errors of $(\hat{\boldsymbol{U}}_{j(\boldsymbol{\Gamma},\boldsymbol{T})} - \boldsymbol{U}_j)$, which are useful for statistical inference.
- The "t Value" is for testing $H_0 : U_j = 0$ versus $H_a : U_j \neq 0$.
- To understand what's in the SAS/MIXED manual, you need to know about "Henderson's Mixed Model" and "BLUP".
- lmer uses a version of Henderson's Mixed Models.

…But first let's look at the $\hat{U}_{oj}$'s graphically…

Empirical Distribution of $\hat{U}_{0j}$'s

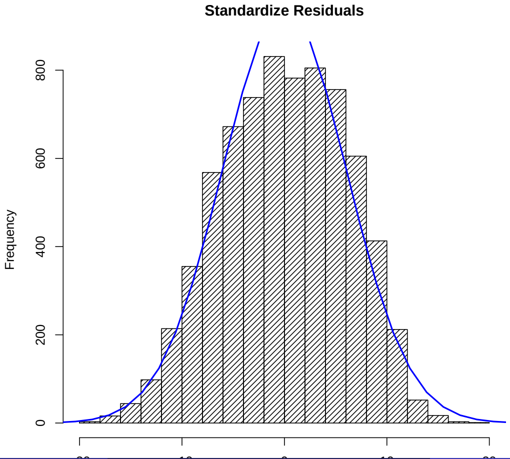


QQ plot with Hacked CIs



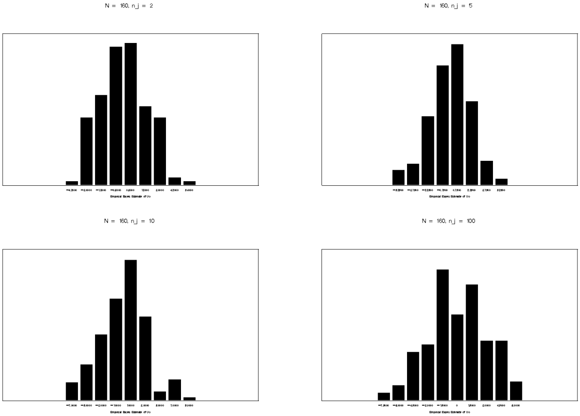
In R

For code, see course web-site.



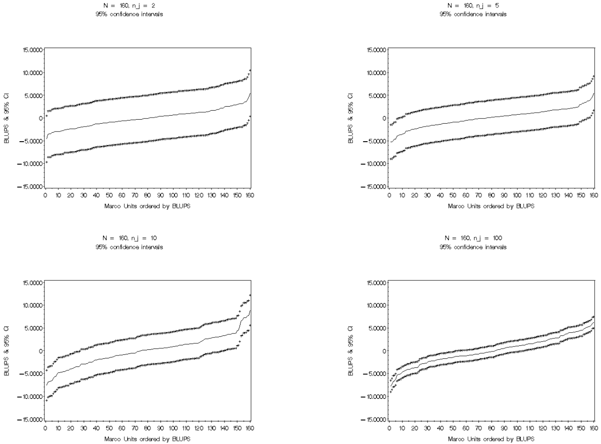
Effect of Micro Sample Size on $\hat{U}_j$:

$N = 160,\ n_j = 2, 5, 10, 100$



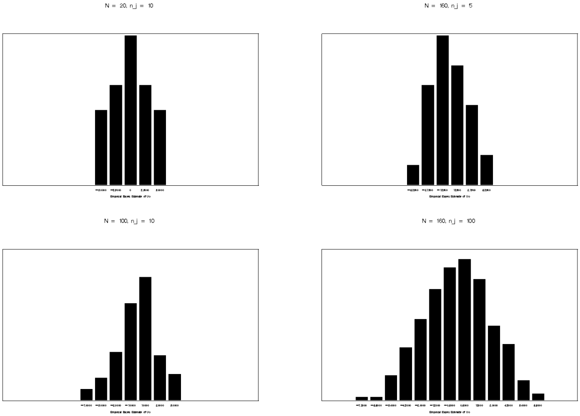
Effect of Micro Sample Size on $\hat{U}_j$:

$N = 160$, $n_j = 2, 5, 10, 100$



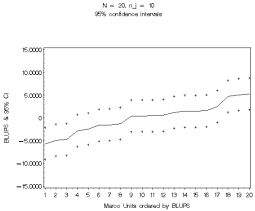
Effect of Macro Sample Size on $\hat{U}_j$:

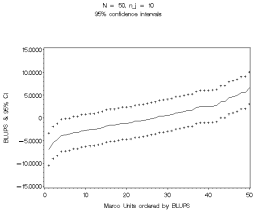
$n_j = 10,\ N = 20, 50, 100, 500$

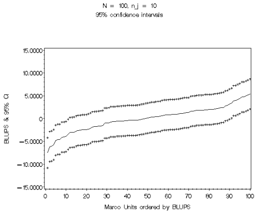


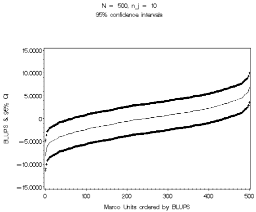
Effect of Macro Sample Size on $\hat{U}_j$:

$n_j = 10,\ N = 20, 50, 100, 500$









Effect of macro Sample Size on $\hat{\gamma}$:

$$n_j = 10,\ N = 20, 50, 100, 500$$

Marco sample size	Effect	Estimate	Standard Error
N=20	Intercept	12.5639	0.8847
	x	2.0568	0.4254
N=50	Intercept	11.9190	0.5726
	x	1.7124	0.2950
N=100	Intercept	12.1473	0.3691
	x	2.2072	0.2032
N=500	Intercept	11.9521	0.1645
	x	2.0185	0.0899

Effect of micro Sample Size on $\hat{\gamma}$:

$$N = 160,\ n_j = 2, 5, 10, 100$$

Micro sample size	Effect	Estimate	Standard Error
$n_j = 2$	Intercept	12.1021	0.4075
	x	1.9392	0.3691
$n_j = 5$	Intercept	12.0976	0.3398
	x	1.6875	0.2330
$n_j = 10$	Intercept	12.0318	0.4248
	x	2.0809	0.3646
$n_j = 100$	Intercept	11.9307	0.2447
	x	1.9794	0.0479

Summary: Effect of N and n_j on $\hat{U}_j$ & $\hat{\gamma}$

- Effect on shape of distribution of $\hat{U}_j$:
  - Increasing n_j (micro) doesn't have much of an effect.
  - Increasing N (macro) leads to more normal looking distribution.
- Effect on confidence limits for $\hat{U}_j$ (i.e. standard errors)
  - Increasing n_j (micro) leads to smaller standard errors of $\hat{U}_j$
  - Increasing N (macro) doesn't change the standard errors of $\hat{U}_j$.
- Effect on parameter estimates (i.e.,γ's): none
- Effect on standard errors of parameters: Increase N or n_j, standard errors get smaller.

  Why? Is there a differential effect?

Effect of N and n_j on s.e's of γ's

N	n_j	n_+		s.e.
20	10	300	Intercept	0.8847
			x	0.4254
160	2	320	Intercept	0.4075
			x	0.3691
50	10	500	Intercept	0.5726
			x	0.2950
160	5	800	Intercept	0.3398
			x	0.2330
100	10	1,000	Intercept	0.3691
			x	0.2032
160	10	1,600	Intercept	0.4248
			x	0.3646
500	10	5,000	Intercept	0.1645
			x	0.0899
160	100	16,000	Intercept	0.2447
			x	0.0479

Choosing sample size for given power

All methods need some educated guesses regarding effects and variance components.

There are 2 stand alone programs that can help:

- PINT (Snijders & Bosker)
- Optimal Design Plus Empirical Evidence: Documentation for the "Optimal Design" Software. By Bloom, Congdon, Hill, Martinez, Raudenbush.

Book that has lots of power with R, SAS and SPSS code to compute power for many designs:

Liu, X.S. *Statistical Power Analysis for the Social and Behavioral Sciences*.

Henderson's Mixed-Model Equations

The equation

$$\hat{\boldsymbol{U}}_{j(\boldsymbol{\Gamma},\boldsymbol{T})} = \boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma})$$

was also derived by Henderson who used a system of linear equations rather than Bayes Theorem.

He basically "stacks" the problem.

Henderson's Mixed-Model Equations

Staring with the linear mixed model,

$$\underbrace{\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\boldsymbol{\Gamma}}_{\boldsymbol{\Gamma}} + \underbrace{\begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{Z}_N \end{pmatrix}}_{\boldsymbol{Z}} \underbrace{\begin{pmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \\ \vdots \\ \boldsymbol{U}_N \end{pmatrix}}_{\boldsymbol{U}} + \underbrace{\begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{R}_2 \\ \vdots \\ \boldsymbol{R}_N \end{pmatrix}}_{\boldsymbol{R}}$$

i.e., $\boldsymbol{y} = \boldsymbol{X\Gamma} + \boldsymbol{ZU} + \boldsymbol{R}$.

Given estimates of variance components $\boldsymbol{T}$, estimates of $\boldsymbol{U}$ can be obtained by solving the "mixed-model" equations for $\boldsymbol{\Gamma}$ and $\boldsymbol{U}$,

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + \boldsymbol{T}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Gamma} \\ \boldsymbol{U} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \end{pmatrix}$$

Henderson's Mixed-Model Equations

Note that the within and between covariance matrices are block diagonal

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 \boldsymbol{I}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \sigma^2 \boldsymbol{I}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \sigma^2 \boldsymbol{I}_N \end{pmatrix} \quad \text{and} \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{T}_N \end{pmatrix}.$$

Solution to Henderson's Equations

$$\hat{\boldsymbol{\Gamma}} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}'\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$$

and

$$\hat{\boldsymbol{U}} = \boldsymbol{T}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\Gamma}})$$

which we've seen before.

- Using Henderson's approach is fine so long as you don't have large data sets (otherwise it's computationally difficult).
- In the SAS/MIXED documentation it is reported that Henderson's estimates are used....these are the same as the EB ones.
- In practice, the more direct approach is more efficient. (i.e., equation for $\hat{U}_j$ given at the beginning of this section and under EB and the equations previously given for. $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{T}}$ in the notes on estimation).

BLUP: Best Linear Unbiased Prediction

"Best" means a predictor or estimator is

- Unbiased.
- Has the smallest variance among all possible unbiased estimators (of a particular form).

If the variance components are known,

Then the Bayes predictions of $\boldsymbol{U}_j$ are the Best Linear Unbaised Predictors or "BLUP."

But the variance components are not known. . .

BLUP: Best Linear Unbiased Prediction

Since estimates of the variance components $\hat{\boldsymbol{T}}$ are used to estimate or predict $\boldsymbol{Y}_j$, i.e.,

$$\hat{\boldsymbol{Y}}_j = \boldsymbol{X}\hat{\boldsymbol{\Gamma}} + \boldsymbol{Z}\hat{\boldsymbol{U}}_j,$$

The EB estimates of $\boldsymbol{U}_j$ are Empricial Best Linear Unbaised Predictors or "EBLUP."

This is related to what follows and has or interpretation of the EB estimates of $\boldsymbol{U}_j$.

Shrinkage

1. Simple Model: Null/Empty
2. Example: HSB random intercept models with $(\text{cSES})_{ij}$.
3. Complex/General Model.
4. Example 2: More complex model.

Shrinkage: Null/Empty Model

$$\begin{aligned} Y_{ij} &= \beta_{0j} + R_{ij} \\ &= \gamma_{00} + U_{0j} + R_{ij} \end{aligned}$$

Using information from group j, the OLS estimates of β_{0j} is

$$\hat{\beta}_{0j} = (1/n_j) \sum_{i=1}^{n_j} Y_{ij} = \bar{Y}_{.j}$$

The group mean.

Shrinkage: Null Model (continued)

If we used information from all the groups, we could estimate β_{0j} as the mean over populations (i.e., γ_{00}); that is,

$$\hat{\gamma}_{00} = \left(\frac{1}{\sum_j n_j}\right) \sum_{j=1}^{M} \sum_{i=1}^{n_j} Y_{ij} = \sum_{j=1}^{M} \frac{n_j}{M} \bar{Y}_{.j} = \bar{Y}_{..}$$

So we can estimate β_{0j} using

- Group information.
- Information from all groups (population).
- A combination of information.

Optimal Estimator of β_{0j}

The optimal (linear) combination (BLUP) is the empirical Bayes estimator of β_{0j}.

It is a weighted average.

$$\begin{aligned}
\hat{\beta}_{0j}^{\text{EB}} &= \lambda_j \hat{\beta}_{0j} + (1-\lambda_j)\hat{\gamma}_{00} \\
&= \left(\frac{\tau_o^2}{\tau_0^2 + \sigma^2/n_j}\right)\hat{\beta}_{0j} + \left(1 - \frac{\tau_o^2}{\tau_0^2 + \sigma^2/n_j}\right)\hat{\gamma}_{00} \\
&= \left(\frac{\tau_o^2}{\tau_0^2 + \sigma^2/n_j}\right)\bar{Y}_{.j} + \left(1 - \frac{\tau_o^2}{\tau_0^2 + \sigma^2/n_j}\right)\bar{Y}_{..} \\
&= \left(1 - \frac{\sigma^2/n_j}{\tau_0^2 + \sigma^2/n_j}\right)\bar{Y}_{.j} + \left(\frac{\sigma^2/n_j}{\tau_0^2 + \sigma^2/n_j}\right)\bar{Y}_{..}
\end{aligned}$$

Optimal Estimator of β_{0j} (continued)

$$\hat{\beta}_{0j}^{\text{EB}} = \lambda_j \hat{\beta}_{0j} + (1-\lambda_j)\hat{\gamma}_{00} = \left(1 - \frac{\sigma^2/n_j}{\tau_0^2 + \sigma^2/n_j}\right)\bar{Y}_{.j} + \left(\frac{\sigma^2/n_j}{\tau_0^2 + \sigma^2/n_j}\right)\bar{Y}_{..}$$

- The weights are both less than 1, so the EB estimate will be closer to the overall mean than the OLS estimator of β_{0j}.
- Consider the extreme cases: $\tau_0^2 = 0$ and $\sigma^2 = 0$.
- This phenomenon is known as Shrinkage.
- The model fitted values are "shrunken" toward the prior average (prior mean of the random effects is ~ 12.

HSB Example of Shrinkage

Random intercept model with $(\text{cSES})_{ij}$.

Consider school 8367 where $n_j = 14$.

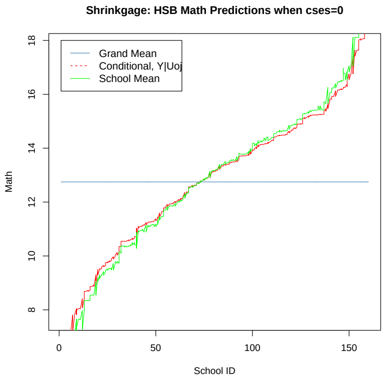
- The weight for the group data is

$$\begin{aligned} \hat{\tau}_o^2/(\hat{\tau}_o^2 + \hat{\sigma}^2/n_j) &= 8.6071/(8.6071 + 37.0056/n_j) \\ &= .76 \end{aligned}$$

- The weight for overall average regression is

$$1 - \hat{\tau}_o^2/(\hat{\tau}_o^2 + \hat{\sigma}^2/n_j) = .23.$$

HSB Example of Shrinkage (continued)



HSB Example of Shrinkage (continued)

- The estimates of math from model conditioning on U_{oj} are "shrunken" toward the prior average (i.e., Y_{++}) where the prior mean of the random effects is ~ 12.
- The variance of the estimates id U_{0j} is less than (or equal) to the data.

	Simple Model		Complex model	
	Model est.	var($\hat{U}_{oj}$)	Model est	var($\hat{U}_{pj}$)
$\hat{\tau}_0^2 = 8.612$		7.853	3.185	2.215
$\hat{\tau}_1^2 =$			59.719	16.22
$\hat{\tau}_2^2 =$			0.911	0.246

Shrinkage for More General/Complex

- Shrinkage also occurs in more complex models.
- Instead of developing this in terms of $\hat{\beta}_{0j}^{\text{EB}}$ we can do it in terms of predicted values of Y_{ij}...

$$\hat{Y}_{ij} = \hat{\beta}_{0j}.$$

- and Because I had to show it to myself ...

Shrinkage for More General/Complex

$$\begin{aligned}
\hat{\boldsymbol{Y}}_j &= \boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + \boldsymbol{Z}_j\hat{\boldsymbol{U}}_j \\
&= \boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + \boldsymbol{Z}_j(\boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1})(\boldsymbol{y}_j - \boldsymbol{X}_j\hat{\boldsymbol{\Gamma}}) \\
&= \boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} - \boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1}\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + \boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1}\boldsymbol{y}_j \\
&= (\boldsymbol{I}_{n_j} - \boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1})\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + \boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_j'\boldsymbol{V}_j^{-1}\boldsymbol{y}_j \\
&= (\boldsymbol{I}_{n_j} - (\boldsymbol{V}_j - \sigma^2\boldsymbol{I}_{n_j})\boldsymbol{V}_j^{-1})\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + (\boldsymbol{V}_j - \sigma^2\boldsymbol{I}_{n_j})\boldsymbol{V}_j^{-1}\boldsymbol{y}_j \\
&= (\boldsymbol{I}_{n_j} - \boldsymbol{I}_{n_j} + \sigma^2\boldsymbol{V}_j^{-1})\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + (\boldsymbol{I}_{n_j} - \sigma^2\boldsymbol{I}_{n_j}\boldsymbol{V}_j^{-1})\boldsymbol{y}_j \\
&= (\sigma^2\boldsymbol{V}_j^{-1})\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + (\boldsymbol{I}_{n_j} - \sigma^2\boldsymbol{V}_j^{-1})\boldsymbol{y}_j
\end{aligned}$$

English translation

$$\hat{\boldsymbol{Y}}_j = (\sigma^2 \boldsymbol{V}_j^{-1})\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + (\boldsymbol{I}_{n_j} - \sigma^2 \boldsymbol{V}_j^{-1})\boldsymbol{y}_j$$

- Predictions of Y_{ij} are weighted combinations of
  - The overall/average population regression (i.e., $\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}}$), and
  - The data from group j (i.e., $\boldsymbol{y}_j$).
- Recall that that covariance matrix for $\boldsymbol{Y}_j$ is

$$\boldsymbol{V}_j = \boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_j' + \sigma^2\boldsymbol{I}_{n_j}$$

Weights for Extreme Case: $T = 0$

- $\boldsymbol{V}_j = \sigma^2 \boldsymbol{I}_{n_j}$ and $\boldsymbol{V}_j^{-1} = (1/\sigma^2)\boldsymbol{I}_{n_j}$
- Weight for the overall average regression is

$$(\sigma^2 \boldsymbol{V}_j^{-1}) = (\sigma^2(1/\sigma^2)\boldsymbol{I}_{n_j}) = \boldsymbol{I}_{n_j},$$

- Weight for group j data is

$$(\boldsymbol{I}_{n_j} - \sigma^2 \boldsymbol{V}_j^{-1}) = (\boldsymbol{I}_{n_j} - \boldsymbol{I}_{n_j}) = \boldsymbol{0}$$

- Predicted value of the response variable is

$$\hat{\boldsymbol{Y}}_j = (\sigma^2 \boldsymbol{V}_j^{-1})\boldsymbol{X}_j\hat{\boldsymbol{\Gamma}} + (\boldsymbol{I}_{n_j} - \sigma^2 \boldsymbol{V}_j^{-1})\boldsymbol{y}_j = \boldsymbol{X}_j\hat{\boldsymbol{\Gamma}}$$

Weights for Extreme Case: $\sigma^2 = 0$

- Then $\boldsymbol{V}_j = \boldsymbol{Z}_j \boldsymbol{T} \boldsymbol{Z}_j' + \sigma^2 \boldsymbol{I}_{n_j} = \boldsymbol{Z}_j \boldsymbol{T} \boldsymbol{Z}_j'$
- Weight for the overall average regression is

$$(\sigma^2 \boldsymbol{V}_j^{-1}) = \boldsymbol{0},$$

- Weight for the group j data is

$$(\boldsymbol{I}_{n_j} - \sigma^2 \boldsymbol{V}_j^{-1}) = \boldsymbol{I}_{n_j}$$

- Predicted value of the response variable is

$$\hat{\boldsymbol{Y}}_j = (\sigma^2 \boldsymbol{V}_j^{-1}) \boldsymbol{X}_j \hat{\boldsymbol{\Gamma}} + (\boldsymbol{I}_{n_j} - \sigma^2 \boldsymbol{V}_j^{-1}) \boldsymbol{y}_j = \boldsymbol{y}_j$$

Comments on Complex Case

- Generally, the covariance matrix for the data is "in between" the two extremes and the predictions are "shrunken" toward toward the priori average regression $(\boldsymbol{X}_j\boldsymbol{\Gamma})$.
- This also implies that for any linear combination (say vector $\boldsymbol{L}$),

$$\text{var}(\boldsymbol{L}'\hat{\boldsymbol{U}}_j) \leq \text{var}(\boldsymbol{L}'\boldsymbol{U}_j)$$

Shrinkage Example w/ Complex Model

HSB: The linear mixed model is

$$
\begin{aligned}
(\text{math})_{ij} &= \gamma_{00} + \gamma_{10}\text{cses}_{ij} + \gamma_{20}(\text{female})_{ij} + \gamma_{30}(\text{minority})_{ij} \\
&\quad + \gamma_{01}(\text{sector})_j + \gamma_{02}(\text{size})_j + \gamma_{03}(\overline{\text{SES}})_j \\
&\quad + \gamma_{11}(\text{sector})_j\text{cses}_{ij} + \gamma_{22}(\text{size})_j(\text{female})_{ij} \\
&\quad + \gamma_{23}(\overline{\text{SES}})_j(\text{female})_{ij} + \gamma_{31}(\text{sector})_j(\text{minority})_{ij} \\
&\quad + U_{0j} + U_{1j}(\text{female})_{ij} + U_{2j}(\text{minority})_{ij} + R_{ij}
\end{aligned}
$$

Shrinkage Example w/ Complex Model

The Mixed Procedure

Model Information	
Data Set	WORK.HSBCENT
Dependent Variable	mathach
Covariance Structure	Unstructured
Subject Effect	id
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z	
τ_0^2	UN(1,1)	id	2.2548	0.5011	4.50	$< .0001$
τ_{12}	UN(2,1)	id	-0.9594	0.4375	-2.19	0.0283
τ_2^2	UN(2,2)	id	0.7119	0.5142	1.38	0.0831
τ_{13}	UN(3,1)	id	-0.2327	0.5053	-0.46	0.6452
τ_{23}	UN(3,2)	id	0.2784	0.4685	0.59	0.5524
τ_3^2	UN(3,3)	id	0.9761	0.6943	1.41	0.0799
σ^2	Residual		35.3860	0.6057	58.42	$< .0001$

Covariance Parameter Estimates

Estimated from MIXED and computed variances and covariances (PROC CORR) of the EB $\hat{\boldsymbol{U}}_j$.:

Parameter	PROC MIXED		PROC CORR	
intercept, intercept	τ_0^2	2.2548	var(U_{0j})	1.2829
female, intercept	τ_{01}	-0.9594	cov(U_{0j}, U_{1j})	-0.4490
female, female	τ_1^2	0.7119	var(U_{1j})	0.1971
minority, intercept	τ_{02}	-0.2327	cov(U_{0j}, U_{2j})	0.0309
minority, female	τ_{12}	0.2784	cov(U_{1j}, U_{2j})	0.0425
minority,minority	τ_2^2	0.9761	var(U_{2j})	0.1561
Residual	σ^2	35.3860	—	

The covariances from PROC CORR are smaller (indicates shrinkage).

Weight Matrices: School 8367

Lower triangle Overall average $\hat{\sigma}^2 \hat{\boldsymbol{V}}_j^{-1}$ and upper Group $\boldsymbol{I} - \hat{\sigma}^2 \hat{\boldsymbol{V}}_j^{-1}$

.95	*	.04	.02	.03	.02	.04	.04	.03	.02	.02	.04	.02	.04	.02	.02
-.02	.98	*	.01	.02	.01	.02	.02	.02	.01	.01	.02	.01	.02	.01	.01
-.03	-.02	.94	*	.05	.02	.03	.03	.05	.02	.02	.03	.02	.03	.02	.02
-.02	-.01	-.02	.98	*	.01	.02	.02	.02	.01	.01	.02	.01	.02	.01	.01
-.04	-.02	-.03	-.02	.95	*	.04	.04	.03	.02	.02	.04	.02	.04	.02	.02
-.04	-.02	-.03	-.02	-.04	.95	*	.04	.03	.02	.02	.04	.02	.04	.02	.02
-.03	-.02	-.05	-.02	-.03	-.03	.94	*	.05	.02	.02	.03	.02	.03	.02	.02
-.02	-.01	-.02	-.01	-.02	-.02	-.02	.98	*	.01	.01	.02	.01	.02	.01	.01
-.02	-.01	-.02	-.01	-.02	-.02	-.02	-.01	.98	*	.01	.02	.01	.02	.01	.01
-.04	-.02	-.03	-.02	-.04	-.04	-.03	-.02	-. 02 .	95	*	.04	.02	.04	.02	.02
-.02	-.01	-.02	-.01	-.02	-.02	-.02	-.01	.-01	-.02	.98	*	.01	.02	.01	.01
-.04	-.02	-.03	-.02	-.04	-.04	-.03	-.02	.-02	-.04	-.02	.95	*	.04	.02	.02
-.02	-.01	-.02	-.01	-.02	-.02	-.02	-.01	.-01	-.02	-.01	-.02	.98	*	.01	.01
-.02	-.01	-.02	-.01	-.02	-.02	-.02	-.01	.-01	-.02	-.01	-.02	-.01 .	98	*	.01

Overall being given more weight.

Normality Assumption: Random Effects

Using $\hat{\boldsymbol{U}}_j$ is to examine the assumption of normality for the random effects.

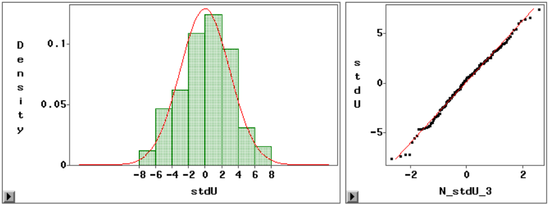
- Problem Even when the linear mixed model is correctly specified, the distribution of the U_j's are all different unless all groups have the same $\boldsymbol{X}_j$ and $\boldsymbol{Z}_j$.
- Solution: Standardize the $\hat{U}_j$'s,

$$\hat{U}_j^* = \frac{\hat{U}_j}{\widehat{\text{s.e.}}_j}$$

And then examine for normality. . .

HSB: stdU $= \hat{U}^*_{0j}/\widehat{\text{S.E.}}_{.j}$ w/ (cSES$_{ij}$)

stdU



Normal QQ Ref Lines		
Line	Intercept	Slope
(red line)	0.0548	3.0805

Parametric Density Estimation					
Curve	Distribution	Method	Mean/Theta	Sigma	Mode
(red line)	Normal	Sample	0.0548	3.0655	0.0548

Moments			
N	160.0000	Sum Wgts	160.0000
Mean	0.0548	Sum	8.7651
Std Dev	3.0655	Variance	9.3975
Skewness	-0.1566	Kurtosis	-0.2021
USS	1494.6881	CSS	1494.2079
CV	5595.9268	Std Mean	0.2424

Quantiles			
100% Max	7.5429	99.0%	6.6874
75% Q3	2.1280	97.5%	6.2364
50% Med	0.2594	95.0%	4.9050
25% Q1	-1.9023	90.0%	3.7711
0% Min	-7.6652	10.0%	-4.3160
Range	15.2081	5.0%	-4.6564
Q3-Q1	4.0303	2.5%	-6.5503
Mode	.	1.0%	-7.3809

Normality: Problem of Shrinkage

Problem: Due to shrinkage, $\hat{U}_j$'s show less variability than is actually present in the population of random effects.

Plots of $\hat{U}_j$ and $\hat{U}_j^*$ do not necessarily reflect the actual distribution of the random effects.

The EB estimates of U_j are very dependent on their assumed prior distribution.

Impact of Non-normality

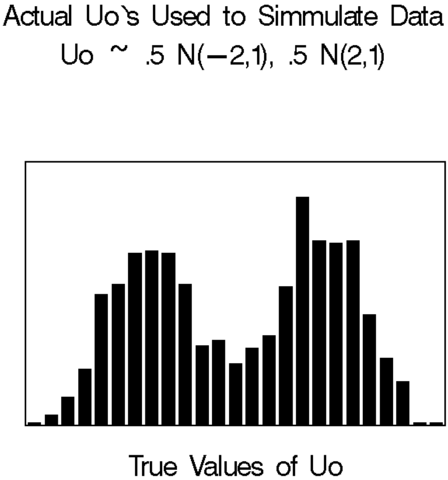
How sensitive are $\hat{U}_j$ to the normality assumption?

Study by Verbeke & Molenberghs, which I ran (out of curiosity and to show you too):

1. Simulate samples from a population with 1000 marco units where each macro unit had 5 observations (i.e., $n_j = 5$) where the random effects followed a mixture of two normal distributions,

$$U_{0j} \sim \left[\frac{1}{2}\mathcal{N}(-2,1) + \frac{1}{2}\mathcal{N}(2,1)\right].$$

Simulated Distribution for U_{0j}



Impact of Non-normality

2 Using the simulated U_{0j}'s, simulate y_{ij}'s according to

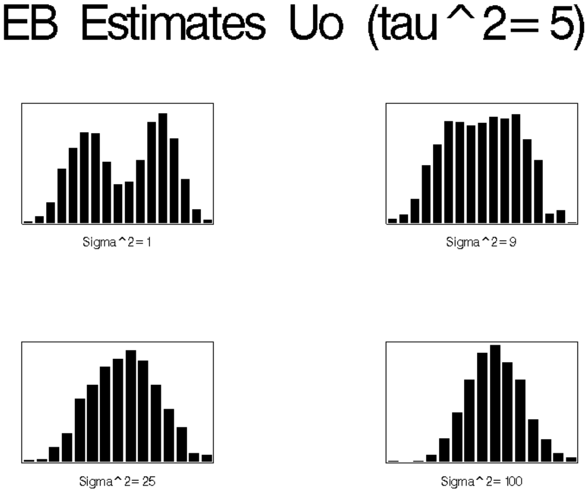
$$y_{ij} = U_{0j} + R_{ij}$$

with $R_{ij} \sim \mathcal{N}(0, \sigma^2)$ where I used $\sigma^2 = 1$, 9, 25 and 100.

3 Fit random intercept model to the simulated data.

4 Examine the resulting distribution of the $\hat{U}_{0j}$'s.

Distributions for $\hat{U}_{0j}$ (Note: $\hat{\tau}_0^2 \sim 5$)



Impact of Non-Normality

Both σ^2 and τ_0^2 influence the shape of the distribution of $\hat{U}_{oj}$:

- If σ^2 is large relative to τ_0^2,

$$\rho_I = \frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

  is small and it's difficult to detect sub-group in the random effects. That is sub-groups in the population are difficult to recognize.

- If a model includes random slopes, it will be easier to detect subgroups in the random effects.

Impact of Non-Normality: In class study

R function `simdata` (on course web-site) to show impact of non-normal U_{oj} (in particular, mixture of two normal distributions).

Try Changing:

- Distribution of U_{oj} (i.e., change mixing weights, "`cut`").
- Values of σ^2 and τ^2 (i.e., ICC).
- Distance between the two distributions.
- N (number of clusters/groups), n_j (number per group).

What happens to

- Distribution of data, Y_{ij}?
- Distribution of $\hat{U}_{0j}$ estimated from lmer?
- Parameter estimates for fixed effects?
- Estimated se for the fixed effects (model based & robust)?
- Results of hypothesis tests for fixed effects?
- Estimated variances?
- Anything else?

Non-Normality & Marginal Model

Based on a number of simulation studies (by others), the wrong distributional assumption for $\boldsymbol{U}_j$,

- Has little effect on the $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{T}}$. (yeah)
- Effects the estimated standard errors of $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{T}}$. (boo)
- Estimated standard errors for $\hat{\boldsymbol{\Gamma}}$ are generally pretty close to the robust ones. (yeah)
- Estimated standard errors for $\hat{\boldsymbol{T}}$ can be really bad. The uncorrected s.e.'s could be 5 times too large or 3 times too small.

  But we generally don't use s.e.'s to test whether $\tau_k^2 = 0$, so this result is not too critical for valid tests of such hypotheses...

  How would/could you test this hypothesis?

Checking the Normality Assumption

- The EB's estimates of $\boldsymbol{U}_j$ depend heavily on the distribution assumption for them.
- Best way to check for non-normality?
  - Compare $\hat{\boldsymbol{U}}_j$ obtained assuming normality to model with those obtained from relaxing the normality assumption.
  - Alternative distributional assumption for $\boldsymbol{U}_j$'s is a mixture of a number of (multivariate) normal's,

$$\boldsymbol{U}_j \sim \sum_{r=1}^{g} p_r \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{T})$$

    where $\sum_k p_r = 1$.
  - Maybe a Gamma distribution (e.g., skewed distribution $\longrightarrow$ reaction times). . . But this would be for Y_{ij}.

Alternative Distribution for U_js

A mixture of (multivariate) normals would be found if

- There is unobserved heterogeneity in the population. . . p_r represents a cluster of the total population.
- You have not included a categorical variable that's important.
- Such a mixture of normals implies a range of possible non-normal distributions for $\boldsymbol{U}_j$'s.

For examples, see Verbeke & Molenberghs (page 90).

Take Away Points

- The estimates of U_j are from analytic derivation using Bayes Theorem (if using MLE or REML).
- Estimates of U_js are a function of y_{ij}, $\hat{\gamma}$s, $\hat{\tau}$s and $\hat{\sigma}^2$.
- Increase n_j increases the precision of the estimate of U_j.
- Increase N has minimal impact on precision of the estimate of U_js.
- See Page 89.
- Choosing sample size for given power (resources given but not illustrated).

Take Away Points

- Shrinkage (a property of Bayesian estimates).
- We assume the U_js are normally distributed; however,
    - If $\hat{U}_j$s are approximately normal, then the normality assumption maybe OK. It's tenable, but definitely not "proven".
    - If $\hat{U}_j$s are not approximately normal, then the normality assumption is violated.
    - If normality assumption is violated, try another distribution (perhaps mixture of normals).

SAS PROC MIXED Options

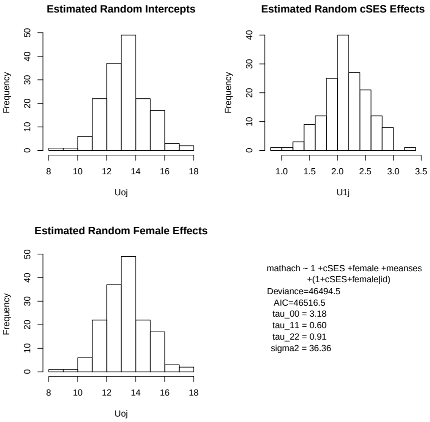
Options to produce predictions and estimates of the random effects, as well as a little data manipulation:

```
PROC MIXED data=hsbcent noclprint covtest method=ML ic;
CLASS id;
MODEL mathach = cSES
/solution outpred=HSBpred outpredm=hsbpm;
RANDOM intercept / subject=id type=un solution cl;
ODS output SolutionR=RanUs;
```

Data Steps

```
DATA tmp99;
   SET hsbpred;
   predij=pred;
   residij=residual;
   lowij=lower;
   upij= upper;
   stdij=StdErrPred;
   KEEP predij residij lowij upij stdij id;
PROC SORT data=tmp99;
   by id;
PROC SORT data=hsbpm;
   by id;
```

How to Do this in R



R

Example using HSB data:

```
model.3 ← lmer(mathach ∼ 1 + xycSES + female
     + meanses + (1 + xycSES + female | id),
      data=hsb, REML=FALSE)

U ← as.data.frame(ranef(model.3))
Uoj ← U[which(U$term=="(Intercept)"),]
U1j ← U[which(U$term=="xyses"),]
U2j ← U[which(U$term=="female"),]

par(mfrow=c(2,2))
hist(Uoj$condval, main="Estimated Random
Intercepts",xlab="Intercept (male)")
hist(U1j$condval, main="Estimated Random cSES
Effects",xlab="cSES/10")
hist(U2j$condval, main="Estimated Random Female
Effects",xlab="female")
```

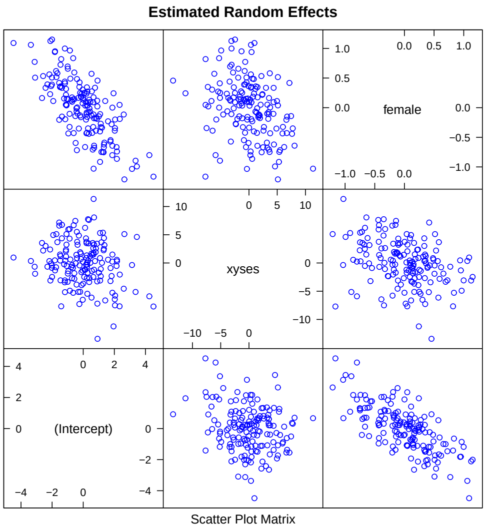
R

```
# and (add some information in the empty plot)
x ← seq(0:10)
y ← seq(0:10)
plot(x,y,type="n",axes=FALSE,ylab="",xlab="")
text(6.1,10,"mathach   1 +cSES +female +meanses")
text(6.4,9,"+(1+cSES+female|id)")
text(3.3,8,"Deviance=46494.5")
text(3,7,"AIC=46516.5")
text(3,6,"tau_00 = 3.18”)
text(3.6,5,"Variance for cses...")
text(3,4,"tau_11 = 1.49”)
text(3,3,"tau_22 = 0.91”)
text(3.1,2,"sigma2 = 36.36")
```

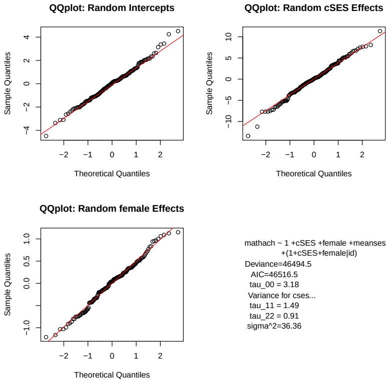
More Graphics

See R on course web-site for how to do the following two figures. . . .

Scatter Plots of $\hat{U}_j$s



QQ plots of random effects



Sign of Problem: QQ plots of random effects