Estimation: Problems & Solutions

Edps/Psych/Soc 587

Carolyn J. Anderson

Department of Educational Psychology

ILLINOIS

Outline

1. Introduction: Estimation of marginal models
2. Maximum Likelihood Estimation Methods
    - Likelihood Equations
    - Full Maximum Likelihood Estimation
    - Restricted Maximum Likelihood Estimation
3. Model Fitting Procedures: Algorithms
4. Estimation Problems
5. A brief example on Bayesian estimation in R

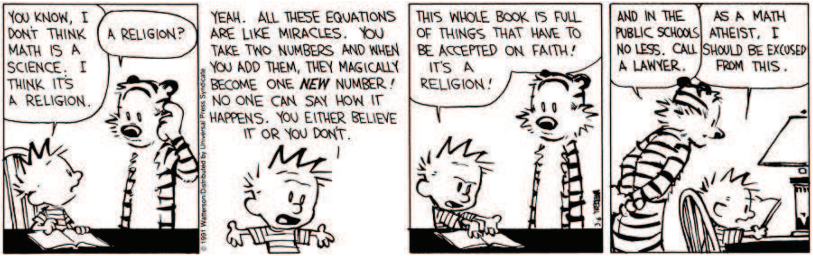
Reading

Snijders & Bosker, pp 60–61, pp 89–90

These notes are also based on

- Verbeke & Molenberghs, Chapter 5, Section 13.5, Section 21.5, Chapter 22.
- Longford (1993, *Random Coefficient Models*).
- Goldstein, H. (2003). *Multilevel Statistical Models*
- My experience with such stuff.

This is Not Magic


YOU KNOW, I DON'T THINK MATH IS A SCIENCE. I THINK IT'S A RELIGION.
A RELIGION?
YEAH. ALL THESE EQUATIONS ARE LIKE MIRACLES. YOU TAKE TWO NUMBERS AND WHEN YOU ADD THEM, THEY MAGICALLY BECOME ONE NEW NUMBER! NO ONE CAN SAY HOW IT HAPPENS. YOU EITHER BELIEVE IT OR YOU DON'T.
THIS WHOLE BOOK IS FULL OF THINGS THAT HAVE TO BE ACCEPTED ON FAITH! IT'S A RELIGION!
AND IN THE PUBLIC SCHOOLS NO LESS. CALL A LAWYER.
AS A MATH ATHEIST, I SHOULD BE EXCUSED FROM THIS.
© 1991 Watterson/Distributed by Universal Press Syndicate

Estimation of the Marginal Model

A hierarchical model:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij}$$

where $R_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Level 2:

$$\begin{array}{rcl} \beta_{0j} & = & \gamma_{00} + \gamma_{01} z_j + U_{0j} \\ \beta_{1j} & = & \gamma_{10} + \gamma_{11} z_j + U_{1j} \end{array}$$

where

$$\boldsymbol{U}_j \sim \mathcal{N}\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{array} \right) \right) \qquad i.i.d.$$

Corresponding Marginal Model

$$Y_{ij} \sim \mathcal{N}\left(\mu_{ij}, \text{var}(Y_{ij})\right)$$

where

$$\mu_{ij} = \gamma_{00} + \gamma_{10}x_{1ij} + \gamma_{01}z_j + \gamma_{11}x_{ij}z_j$$
$$\text{var}(Y_{ij}) = (\tau_0^2 + 2\tau_{10}x_{ij} + \tau_1^2 x_{ij}^2 + \sigma^2)$$

- The HLM implies the marginal model.
- The marginal model does not imply the HLM.
- The U_j's and R_{ij} are latent or unobserved variables and are not part of the marginal model.

The Logic

- What we observe: Data

  A sample of individuals from different groups and take measurements or make observations on Y_{ij}, x_{ij}, and z_j.

- Hypothesis: The HLM model.

  Implies the distribution Y_{ij}, the "marginal model."

- Using data, estimate the parameters of the marginal model:
  - Regression coefficients, the γ's.
  - Variance components, the τ's and σ^2.

Statistical Inference

- Are based on the marginal model.
- Regarding the γ's, τ's and σ^2.
- Not on the U_j's and R_{ij}. There are no explicit assumptions regarding the presence or existence of unobserved, random variables in the marginal model.
- Estimating the random, unobserved variables, the U_j's and R_{ij}, is the topic for later.

Methods and Algorithms

- Methods of estimation:
  - (Full) Maximum Likelihood (ML).
  - Restricted Maximum Likelihood (REML).
- Algorithms that implement the estimation method.
  - Newton-Raphson (NR).
  - Fisher Scoring.
  - Iterative Generalized Least squares (IGLS).
  - Expectation maximization (EM).
  - Bayesian
- Others.

Methods and Algorithms (continued)

The six possibilities we'll discuss,

	Computing Algorithm			
Estimation Method	Newton-Raphson	Fisher Scoring	Iterative GLS**	EM
MLE				
REML				

Given an estimation method, the results from different algorithms should be the same. ** Qualifications

** Asymptotically: Depends critically on normality assumption.

Methods and Algorithms (continued)

An estimation method yields the same results regardless of the algorithm used to implement it.

The algorithms differ with respect to

- Computational problems
- CPU time

Likelihood Equations:

- The marginal model derived from an HLM:

$$\boldsymbol{Y}_j \sim \mathcal{N}\left(\boldsymbol{X}_j\boldsymbol{\Gamma}, (\boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_J' + \sigma^2\boldsymbol{I})\right).$$

- We'll look at simple and familiar cases to explain principles. The principles for general and complex models are the same.
- We'll start with the univariate normal.

Likelihood Equations: Univariate Normal

- Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$.
- The probability density function $(p.d.f)$ of Y

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2} \frac{(y-\mu)^2}{\sigma^2} \right\}$$

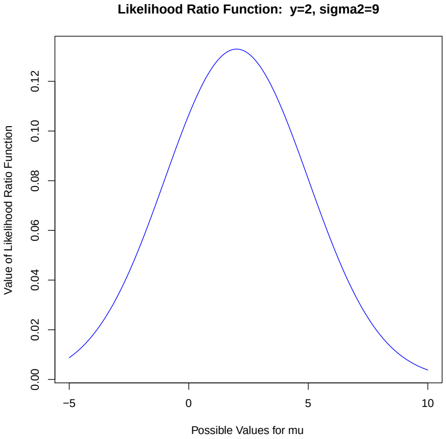
The likelihood of y given values of μ and σ^2.

- If we have one observation on Y, say y_1, and we know σ^2, the likelihood function of μ is

$$L(\mu|y_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2} \frac{(y_1-\mu)^2}{\sigma^2} \right\}$$

- The likelihood of various values for μ.
  How likely μ is given the data.

Univariate Normal: Example



Univariate Normal (continued)

A random sample from a normal population where $\sigma^2 = 9$ and μ is unknown:

$$y_1 = -1, \quad y_2 = 2, \quad y_3 = 3, \quad y_4 = 6, \quad y_5 = 10$$

Since observations are independent, the likelihood equation for μ given our data is

$$\begin{aligned} L(\mu|y_1,\ldots,y_5,\sigma^2) &= L(\mu|y_1,\sigma^2)L(\mu|y_2,\sigma^2)L(\mu|y_3,\sigma^2) \\ & L(\mu|y_4,\sigma^2)L(\mu|y_5,\sigma^2) \\ &= L(\mu|-1,9)L(\mu|2,9)L(\mu|3,9) \\ & L(\mu|6,9)L(\mu|10,9) \end{aligned}$$

Basically, an application of the multiplicative rule of probability.

Univariate Normal (continued)

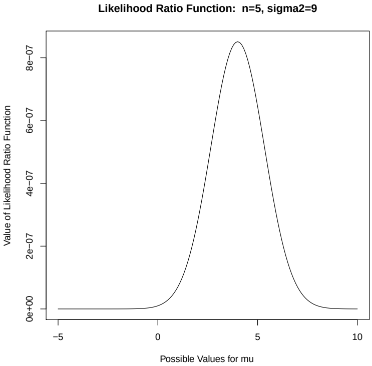
In general,

$$\begin{aligned} L(\mu|y_1,\ldots,y_n,\sigma^2) &= \prod_{i=1}^{n} L(\mu|y_i,\sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2} \frac{(y_i-\mu)^2}{\sigma^2} \right\} \end{aligned}$$

What does this looks like for our "data"?

Univariate Normal: Example 2

Data: $y_1 = -1, y_2 = 2, y_3 = 3, y_4 = 6, y_5 = 10$



What's your "best" guess for μ?

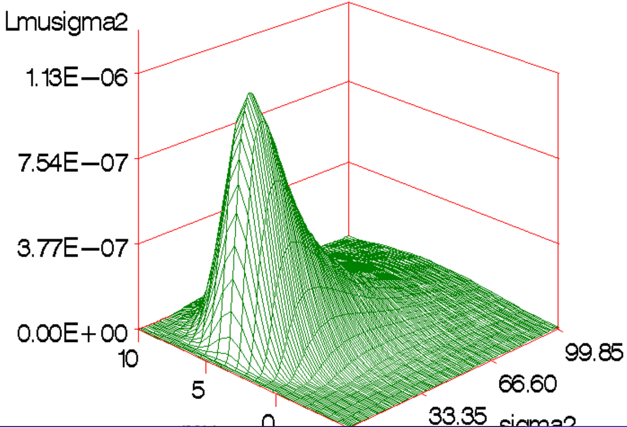
Univariate Normal: μ and σ^2

Data: $y_1 = -1, y_2 = 2, y_3 = 3, y_4 = 6, y_5 = 10$

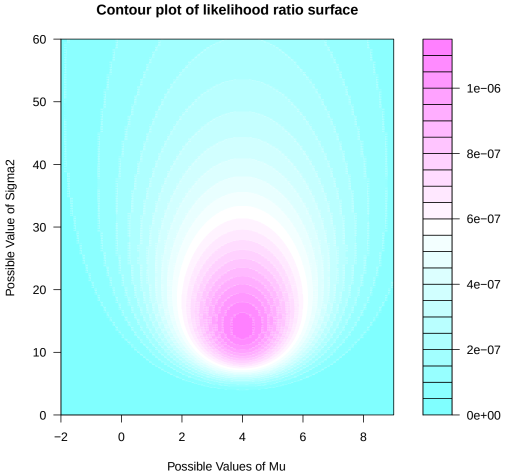
$$\begin{aligned} L(\mu, \sigma^2|y_1, \ldots, y_n,) &= L(\mu, \sigma^2|-1)L(\mu, \sigma^2|2)L(\mu, \sigma^2|3)L(\mu, \sigma^2|6) \\ &\quad L(\mu, \sigma^2|10) \\ &= \prod_{i=1}^{n} L(\mu, \sigma^2|y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Univariate Normal: μ and σ^2

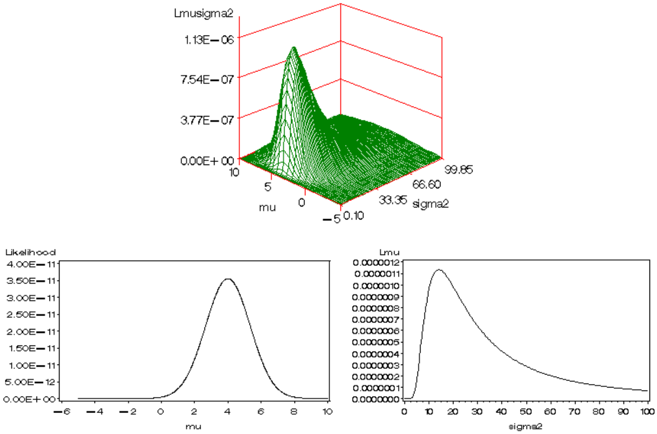
Data: $y_1 = -1, y_2 = 2, y_3 = 3, y_4 = 6, y_5 = 10$



Another View



Univariate Normal: μ and σ^2



Likelihood Function for Mean and Variance
Data: —1, 2, 3, 6, 10

Multivariate Normal: $p.d.f.$

The marginal model derived from an HLM is a multivariate normal distribution:

$$f(\boldsymbol{Y}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{\frac{-1}{2}(\boldsymbol{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})\right\}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_n^2 \end{pmatrix}.$$

Likelihood for Multivariate Normal

The likelihood equation for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given $\boldsymbol{y}$,

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{\frac{-1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}$$

A random sample of N vectors of variables from the same population; that is,

$$\boldsymbol{y}_j' = (y_{1j}, y_{2j}, \ldots, y_{nj}), \qquad \text{for } j = 1, \ldots, N$$

The likelihood equation for the parameters is

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N) = \prod_{j=1}^{N}(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{\frac{-1}{2}(\boldsymbol{y}_j-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu})\right\}$$

Likelihood Equations, Marginal Model

Now things are a bit more complex and simpler in that

- We have a different distribution for each of the N macro units; that is, $\boldsymbol{Y}_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $j = 1, \ldots, N$.
- We add our models for $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ into the multivariate likelihood equation,

$$\begin{aligned} \boldsymbol{\mu}_j &= \boldsymbol{X}_j \boldsymbol{\Gamma} \\ \boldsymbol{\Sigma}_j &= \boldsymbol{Z}_j \boldsymbol{T} \boldsymbol{Z}_j' + \sigma^2 \boldsymbol{I} \end{aligned}$$

Likelihood Equations for One Marco Unit

For one group/cluster/macro unit's population,

$$
\begin{aligned}
L(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j | \boldsymbol{y}_j) &= L(\boldsymbol{\Gamma}, \boldsymbol{T}, \sigma^2 | \boldsymbol{y}_j, \boldsymbol{X}_j, \boldsymbol{Z}_j) \\
&= (2\pi)^{-n_j/2} |(\boldsymbol{Z}_j \boldsymbol{T} \boldsymbol{Z}_j' + \sigma^2 \boldsymbol{I})|^{-1/2} \\
&\quad \times \exp\left\{ \frac{-1}{2} (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\Gamma})' (\boldsymbol{Z}_j \boldsymbol{T} \boldsymbol{Z}_j' + \sigma^2 \boldsymbol{I})^{-1} \right. \\
&\quad \left. (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\Gamma}) \right\}
\end{aligned}
$$

Since the observations between groups are independent, we take the product of the likelihood equations for the groups. . . ,

Likelihood Equations for Marginal Model

$$\begin{aligned} L(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) &= \\ L(\boldsymbol{\Gamma}, \boldsymbol{T}, \sigma^2 | \boldsymbol{y}_j, \boldsymbol{X}_j, \boldsymbol{Z}_j, j = 1, \ldots, N) &= \prod_{j=1}^{N} L(\boldsymbol{\Gamma}, \boldsymbol{T}, \sigma^2 | \boldsymbol{y}_j, \boldsymbol{X}_j, \boldsymbol{Z}_j) \end{aligned}$$

And

$$\begin{aligned} \prod_{j=1}^{N} L(\boldsymbol{\Gamma}, \boldsymbol{T}, \sigma^2 | \boldsymbol{y}_j, \boldsymbol{X}_j, \boldsymbol{Z}_j) &= \prod_{j=1}^{N} (2\pi)^{-n_j/2} |(\boldsymbol{Z}_j \boldsymbol{T} \boldsymbol{Z}_j' + \sigma^2 \boldsymbol{I})|^{-1/2} \\ & \quad \exp\left\{ \frac{-1}{2} (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\Gamma})' (\boldsymbol{Z}_j \boldsymbol{T} \boldsymbol{Z}_j' + \sigma^2 \boldsymbol{I})^{-1} \right. \\ & \quad \left. (\boldsymbol{y}_j - \boldsymbol{X}_j \boldsymbol{\Gamma}) \right\} \end{aligned}$$

Hang in there. . .

Maximum Likelihood Estimator of Γ

- If we know $\boldsymbol{\Sigma}_j$, then (with a bit of algebra & calculus):

$$\hat{\boldsymbol{\Gamma}} = \left(\sum_{j=1}^{N} \boldsymbol{X}_j' \boldsymbol{\Sigma}_j^{-1} \boldsymbol{X}_j \right)^{-1} \sum_{j=1}^{N} \boldsymbol{X}_j' \boldsymbol{\Sigma}_j^{-1} \boldsymbol{y}_j$$

- For univariate data (& independent observations), this is just the sample mean, $\bar{y} = (1/N) \sum_{j=1}^{N} y_j$.
- Since we don't know $\boldsymbol{\Sigma}_j$, we use an estimate of it in the above equation.

For those who want the derivation

Assume that we know $\boldsymbol{\Sigma}_j$ and take the log of the likelihood:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\Gamma}|\boldsymbol{y}_j, \boldsymbol{\Sigma}_j) &= \sum_j \left\{ \ln(2\pi)^{-n_j/2} + \ln(|\boldsymbol{\Sigma}_j|^{-1/2}) \right\} \\
&\quad + \sum_j \left\{ \frac{-1}{2} (\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma})' \, |\boldsymbol{\Sigma}_j|^{-1} \, (\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\Gamma}) \right\} \\
&= \sum_j \left\{ \ln(2\pi)^{-n_j/2} + \ln(|\boldsymbol{\Sigma}_j|^{-1/2}) - \frac{1}{2}(\boldsymbol{y}_j'|\boldsymbol{\Sigma}_j|^{-1}\boldsymbol{y}_j) \right\} \\
&\quad \underbrace{+ \sum_j \left\{ \boldsymbol{\Gamma}' \boldsymbol{X}_j' |\boldsymbol{\Sigma}_j|^{-1} \boldsymbol{y}_j - \frac{1}{2} \boldsymbol{\Gamma}' \boldsymbol{X}_j' |\boldsymbol{\Sigma}_j^{-1}| \boldsymbol{X}_j \boldsymbol{\Gamma} \right\}}
\end{aligned}$$

Take derivative of Kernel:

$$\frac{\partial(\text{Kernel})}{\partial\boldsymbol{\Gamma}} = \sum_j \left\{ \boldsymbol{X}_j' |\boldsymbol{\Sigma}_j|^{-1} \boldsymbol{y}_j - \boldsymbol{X}_j' |\boldsymbol{\Sigma}_j^{-1}| \boldsymbol{X}_j \boldsymbol{\Gamma} \right\}$$

Set equal to 0 and solve for $\boldsymbol{\Gamma}$:

$$\boldsymbol{\Gamma} = (\sum_j \boldsymbol{X}_j' |\boldsymbol{\Sigma}^{-1}| \boldsymbol{X}_j)^{-1} \sum_j \boldsymbol{X}_j' |\boldsymbol{\Sigma}_j|^{-1} \boldsymbol{y}_j$$

Maximum Likelihood Estimation

- The maximum likelihood estimates of the regression coefficients, $\hat{\mathbf{\Gamma}}$, and the variance components, $\hat{\boldsymbol{T}}$ (and σ^2), are those values that give us the largest value of the likelihood function.

- How do we do this?

- Univariate case: given a sample y_1, . . . , y_n, all from $\mathcal{N}(\mu, \sigma^2)$, the MLE of the mean μ and variance σ are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Maximum Likelihood Estimation

In the case of a marginal model derived from an HLM, there is no simple closed form equation(s) that will give us the MLE estimates of $\mathbf{\Gamma}$ and $\boldsymbol{T}$... This is where our computing algorithms come in.

Restricted MLE

- Restricted maximum likelihood (REML) estimation and maximum likelihood estimation (MLE) are similar with respect to estimating means but differ more with respect to to variance estimates.
- To get an idea of what REML is & how it differs from MLE, we'll consider
    - Variance estimation from a sample taken from a $\mathcal{N}(\mu, \sigma^2)$ population.
    - Residual variance estimation in linear regression.
    - REML estimation for marginal model.

Variance Estimation: Univariate $\mathcal{N}$

- Goal: Estimate the variance of a $\mathcal{N}(\mu, \sigma^2)$ based on a random sample $Y_1, Y_2, \ldots, Y_n$.
- MLE Solution(s):
  - When μ is known:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 \longrightarrow \text{unbiased estimator}$$

  - When μ is not known:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \longrightarrow \text{biased estimator}$$

Unbiased Variance Estimator

- The expected value of $\hat{\sigma}^2$ is

$$\mathsf{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

- The MLE estimate is too small, because we estimated μ.
- The unbiased estimator of σ^2 is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \text{and} \qquad \mathsf{E}(s^2) = \sigma^2.$$

- The unbiased estimator is an REML estimate.

General Procedure

To get an unbiased estimate of σ^2, it must not depend on the mean.

- n Observations $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.
- $(Y_1, Y_2, \ldots, Y_n)' = \boldsymbol{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \boldsymbol{I}_n)$
- $\mu \mathbf{1}_n = \mu \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}$
- Covariance matrix,

$$\sigma^2 \boldsymbol{I}_n = \sigma^2 \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix}$$

REML General Procedure (continued)

Let $\boldsymbol{A}$ equal an $(n \times (n-1))$ matrix with (n-1) linearly independent columns that are orthogonal to $\mathbf{1}_n$.

e.g.,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ -1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 \end{pmatrix}$$

Any matrix $\boldsymbol{A}$ that meets these two requirements will work.

REML General Procedure (continued)

Define the vector $\boldsymbol{W}$ as

$$\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{Y}$$

The vector $\boldsymbol{W}$ has $(n-1)$ elements that are "error contrasts."
e.g.,

$$\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{Y} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_1 - Y_2 \\ Y_1 - Y_3 \\ \vdots \\ Y_1 - Y_n \end{pmatrix}$$

REML General Procedure (continued)

$$\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{Y} = \begin{pmatrix} Y_1 - Y_2 \\ Y_1 - Y_3 \\ \vdots \\ Y_1 - Y_n \end{pmatrix}$$

The distribution of $\boldsymbol{W}$ is

$$\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{A}'\boldsymbol{A})$$

(remember the little extra linear algebra in notes on random intercept and slope models?)

So we now have a vector where we know what the mean of the variables equals.

REML General Procedure (continued)

Since we know the mean of $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{W}$ has a (multivariate) normal distribution, the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{\boldsymbol{Y}'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}}{n-1}$$

- which is just a complex way of writing s^2,
- This is the REML estimator of σ^2.
- The REML estimator is based on $(n-1)$ error contrasts; that is, it's based on what's left over after you get rid of (estimate) the mean.

Residual Variance Estimation

In Linear Regression

Goal: Estimate the variance of the residual σ^2 in standard linear regression,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\boldsymbol{X}$ is an $(n \times p)$ design matrix.
- $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of regression parameters.
- $\boldsymbol{\epsilon}$ is a $(n \times 1)$ vector of residuals (errors).
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ i.i.d.

REML in Linear Regression

- The MLE (which is also the ordinary least square estimate) of the regression coefficients is

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- We can estimate the residuals by

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- The MLE estimate of σ^2 is the variance of the estimated residuals:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

REML & Linear Regression (continued)

…in terms of linear algebra, the MLE estimate of σ^2 (the variance of the estimated residuals)

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n} = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y})'(\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y})\end{aligned}$$

This estimator of σ^2 is too small or “downward biased”.

$$\mathsf{E}(\hat{\sigma}^2) = \mathsf{E}\left(\frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n}\right) = \frac{n-p}{n}\sigma^2.$$

REML & Linear Regression (continued)

The unbiased estimator of the variance of the estimated residuals, which is the REML estimator, is

$$\begin{aligned} s^2 &= \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-p} \\ &= \frac{(\boldsymbol{Y} - \boldsymbol{X}\overbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}}^{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\overbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}}^{\boldsymbol{\beta}})}{n-p} \end{aligned}$$

REML & Linear Regression (continued)

The REML estimator s^2 can be obtained using the general method:

- Get a matrix $\boldsymbol{A}$ that is $(n \times (n-p))$ where the columns are linearly independent and orthogonal to the columns of the design matrix $\boldsymbol{X}$.
- Define a vector $\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{Y}$ of "error contrasts", so $\boldsymbol{W} \sim$ normal where the mean of $\boldsymbol{W}$ does not depend on the mean of $\boldsymbol{Y}$. The only unknown parameter of the distribution for $\boldsymbol{W}$ is σ^2.
- When you maximize the likelihood equations for $\boldsymbol{W}$, you get $s^2 = (\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}})/(n-p)$.

REML & Linear Mixed Model

- In the marginal models for HLMs, the mean depends on all the γ's. So, in estimating the τ's and σ^2, we want to take into account the loss of degrees of freedom in estimating the γ's.
- We can't just compute a variance estimate and figure out what the bias factor is and then correct our estimate so that it is no longer biased, which is all we really would have needed to do in the two previous sections.
- We have to use the error contrast method.

REML & Linear Mixed Model

Step 1: “Stack” all level 1 regressions,

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix} \boldsymbol{\Gamma} + \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{Z}_N \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \\ \vdots \\ \boldsymbol{U}_N \end{pmatrix} + \begin{pmatrix} \boldsymbol{R}_1 \\ \boldsymbol{R}_2 \\ \vdots \\ \boldsymbol{R}_N \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X\Gamma} + \boldsymbol{ZU} + \boldsymbol{R}$$

This $\boldsymbol{Y}$ is an $(n_+ \times 1)$ vector with distribution

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X\Gamma}, \boldsymbol{V})$$

Step 1 (continued)

where $\boldsymbol{V}$ is an $(n_+ \times n_+)$ covariance matrix for all the data,

$$\begin{aligned}\boldsymbol{V} &= \boldsymbol{Z}\boldsymbol{T}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}_{n_+} \\ &= \left(\begin{array}{cccc} (\boldsymbol{Z}_1\boldsymbol{T}\boldsymbol{Z}_1' + \sigma^2\boldsymbol{I}) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{Z}_2\boldsymbol{T}\boldsymbol{Z}_2' + \sigma^2\boldsymbol{I}) & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & (\boldsymbol{Z}_N\boldsymbol{T}\boldsymbol{Z}_N' + \sigma^2\boldsymbol{I}) \end{array}\right)\end{aligned}$$

Steps 2 & 3 (REML of Linear Mixed Model)

Step 2: Using any $(n_+ \times (n_+ - p))$ matrix with linearly independent columns and orthogonal to the columns of the fixed effects design matrix $\boldsymbol{X}$, define error contrasts,

$$\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{Y}$$

Note: $p =$ number of fixed effects (i.e., γ's).

We now have $\boldsymbol{W}$ which is an $(n_+ \times 1)$ vector such that

$$\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A}'\boldsymbol{V}\boldsymbol{A}).$$

Step 3 The error contrasts $\boldsymbol{W}$ do not depend on the means (i.e., $\boldsymbol{\Gamma}$), so use maximum likelihood estimation to find $\boldsymbol{T}$ and σ^2.

- These maximum likelihood estimates of $\boldsymbol{T}$ and σ^2 do not depend on the specific choice of $\boldsymbol{A}$.
- Estimation of fixed effects is a separate step.

Comparison of MLE & REML

By example using the NELS88 data.

- Specific data: Random samples of $N = 20$, 50, and 100 schools, as well as full $N = 1003$ schools.
- Use both (full) ML and REML estimation of
  - Model 1: Random intercept with only one fixed effect.
  - Model 2: Random intercept and slope with two fixed effects.
  - Model 3: Random intercept and slope with 6 fixed effects and a cross-level interaction.

Design of Mini-Study

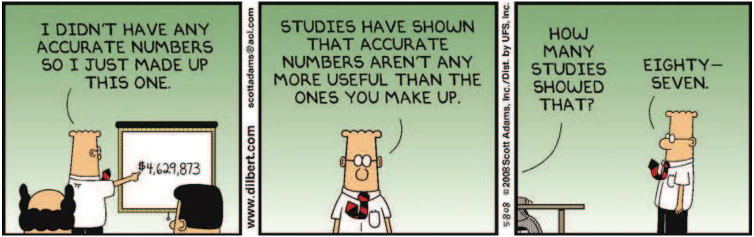
	$N = 20$		$N = 50$		$N = 100$		$N = 1003$	
Model	ML	REML	ML	REML	ML	REML	ML	REML
1. Random intercept								
2. Random intercept and slope								
3. Random intercept & slope with 6 fixed effects								

- I tried random samples of $N = 10$, but had lots of trouble fitting complex model.
- The random sample of $N = 50$ reported was my second attempt sample. Model 2 fit to my initial random sample of $N = 50$ would not converge for ML but it did for $REML$.

Why We Look at Toy Data

By "toy data", I mean here hypothetical, simulated, or sampled.

Thursday May 08, 2008


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

www.dilbert.com scottadams@aol.com

5.8.08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.

Mini-Study: Model 1

- Level 1:

$$(\mathsf{math})_{ij} = \beta_{0j} + R_{ij}$$

- Level 2:

$$\beta_{0j} = \gamma_{00} + U_{0j}$$

- Linear Mixed Model:

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

where

$$\begin{pmatrix} U_{0j} \\ R_{ij} \end{pmatrix} = \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right)$$

Results for Model 1

Fixed Effects

	$N=20$		$N=50$		$N=100$		$N=1003$	
	ML	REML	ML	REML	ML	REML	ML	REML
γ_{00}	50.74	50.74	50.91	50.91	50.54	50.54	50.80	50.80
	(1.08)	(1.11)	(.66)	(.66)	(.55)	(.55)	(.18)	(.18)

- $\hat{\gamma}_{00}$'s are very similar. They aren't exactly the same, differences show up in the third decimal point.
- The SE's for $\hat{\gamma}_{00}$ are smaller with ML, but differences get smaller with larger samples.

Results for Model 1

Random Effects

	$N = 20$		$N = 50$		$N = 100$		$N = 1003$	
	ML	REML	ML	REML	ML	REML	ML	REML
τ_0^2	19.62	20.83	17.27	17.72	26.11	26.41	26.55	26.58
	(7.34)	(7.93)	(4.29)	(4.43)	(4.26)	(4.32)	(1.35)	(1.37)
σ^2	83.32	83.33	77.55	77.55	77.77	77.78	76.62	76.62
	(5.50)	(5.50)	(3.48)	(3.48)	(2.43)	(2.43)	(.76)	(.76)

- $\hat{\tau}_0^2$'s are smaller when use ML.
- $\hat{\sigma}^2$'s are smaller when use ML (to 2^{nd} decimal point).
- The SE's for $\hat{\tau}_0^2$ and $\hat{\sigma}^2$ are smaller with ML, but the differences between the SE's from ML and REML get smaller with larger samples.

Mini-Study: Model 2

- Level 1:

$$(\text{math})_{ij} = \beta_{0j} + \beta_{1j}(\text{homework})_{ij} + R_{ij}$$

- Level 2:

$$\begin{array}{rcl} \beta_{0j} & = & \gamma_{00} + \gamma_{01}(\text{ratio})_j + U_{0j} \\ \beta_{1j} & = & \gamma_{10} + U_{1j} \end{array}$$

- Linear Mixed Model:

$$(\text{math})_{ij} = \gamma_{00} + \gamma_{10}(\text{homew})_{ij} + \gamma_{01}(\text{ratio})_j + U_{0j} + U_{1j}(\text{homew})_{ij} + R_{ij}$$

where

$$\begin{pmatrix} U_{0j} \\ U_{1j} \\ R_{ij} \end{pmatrix} = \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{10} & 0 \\ \tau_{10} & \tau_1^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right)$$

Mini-Study Results: Model 2

Fixed Effects

	$N = 20$		$N = 50$		$N = 100$		$N = 1003$	
	ML	REML	ML	REML	ML	REML	ML	REML
intercept	52.72	52.69	48.11	48.12	49.30	49.30	51.52	51.52
	(2.86)	(2.99)	(2.10)	(2.15)	(2.29)	(2.32)	(.61)	(.61)
homewk	1.61	1.61	1.74	1.74	1.50	1.50	1.47	1.47
	(.34)	(.35)	(.20)	(.20)	(.14)	(.14)	(.05)	(.05)
ratio	$-.30$	$-.30$	$-.04$	$-.04$	$-.09$	$-.09$	$-.25$	$-.25$
	(.15)	(.15)	(.11)	(.12)	(.13)	(.13)	(.03)	(.03)

- $\hat{\gamma}$'s pretty similar for given N.
- SE for $\hat{\gamma}$'s small with ML, but differences get smaller with larger N.
- γ_{01} for ratio n.s. until $N = 1003$.

Mini-Study Results: Model 2

Random Effects

	$N = 20$		$N = 50$		$N = 100$		$N = 1003$	
	ML	REML	ML	REML	ML	REML	ML	REML
τ_0^2	3.18	4.43	15.75	16.70	18.01	18.58	23.27	23.33
	(3.72)	(4.39)	(5.25)	(5.54)	(4.03)	(4.16)	(1.53)	(1.53)
τ_{10}	2.13	2.03	$-.87$	$-.97$	1.01	.99	$-.90$	$-.91$
	(1.46)	(1.60)	(1.13)	(1.18)	(.66)	(.73)	(.30)	(.31)
τ_1^2	.57	.72	.25	.31	.00	.02	.52	.52
	(.75)	(.82)	(.38)	(.40)	.	(.25)	.10)	(.10)
σ^2	76.16	76.16	71.52	71.49	73.74	73.73	71.74	71.74
	(5.13)	(5.13)	(3.29)	(3.29)	(2.31)	(2.35)	(.72)	(.72)

Random Effects: Model 2

- Variance estimates ($\hat{\tau}_0^2$, $\hat{\tau}_1^2$, $\hat{\sigma}^2$) are smaller with ML.
- SE's for the τ's and σ^2 are smaller with ML.
- All SE's get smaller with larger N.
- For $N = 100$, we have some trouble:
  - with ML, $\hat{\tau}_1^2 = 1.77E - 17$ and no standard error. Why don't we get this with REML?
  - In the SAS/LOG for both ML and REML, `‘‘Estimated` G `matrix is not positive definite.’’`
    (SAS G matrix is our $\boldsymbol{T}$ matrix)

Random Effects: Model 2, $N = 100$

Using the REML results...

$$\hat{\boldsymbol{T}} = \begin{pmatrix} 18.58 & .99 \\ .99 & .02 \end{pmatrix} \longrightarrow \widehat{\text{corr}} = \begin{pmatrix} 1.00 & 1.56 \\ 1.56 & 1.00 \end{pmatrix}$$

This result is fine for the marginal model for math scores, but it's not OK for HLM/Linear Mixed Model. This result is inconsistent with our interpretation of τ's.

Mini-Study Model 3

- Level 1:

$$(\text{math})_{ij} = \beta_{0j} + \beta_{1j}(\text{homework})_{ij} + \beta_{2j}(\text{cSES}_{ij}) + R_{ij}$$

- Level 2:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(\text{ratio})_j + \gamma_{02}(\text{public})_j + \gamma_{03}(\bar{\text{SES}})_j + U_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(\text{ratio})_j + U_{1j}\\ \beta_{2j} &= \gamma_{20}\end{aligned}$$

- Linear Mixed Model:

$$\begin{aligned}(\text{math})_{ij} &= \gamma_{00} + \gamma_{10}(\text{homew})_{ij} + \gamma_{20}(\text{cSES})_{ij}\\ &\quad +\gamma_{01}(\text{ratio})_j + \gamma_{03}(\bar{\text{SES}}_j)\\ &\quad +\gamma_{11}(\text{ratio})_j(\text{homew})_{ij}\\ &\quad +U_{0j} + U_{1j}(\text{homew})_{ij} + R_{ij}\end{aligned}$$

Model 3: Mini-Study Results

- With these data and the considerably more complex model, I got the same basic pattern of results.
- With more complex model and random sample $N = 100$ and ML, I was unable to get the model to converge.
- I was expecting to get more differences in terms of the values of estimated fixed than I actually got when I added more fixed effects.

Summary Comparison of REML & MLE

- Maximum Likelihood Principle:
  Both are based on this, so both yield estimators that are
  - Consistent
  - Efficient
  - Asymptotic normal
- Variance Components:
  The REML estimates are larger than the ML estimates.
- Fixed Effects
  - MLE provides estimates of fixed effects, while REML requires an extra step.
  - For univariate normal and standard linear regression, MLE of the mean and variance are independent.
  - The estimated mean is the same under MLE and REML; that is, MLE or REML could be used to estimate the variance and the estimated means wouldn't change.

Comparison of REML & MLE: Fixed Effects

- For the linear mixed model, the estimation of mean and variance are *not* independent, because

$$\hat{\mathbf{\Gamma}} = \left(\sum_{j=1}^{N} \boldsymbol{X}_j' \hat{\mathbf{\Sigma}}_j^{-1} \boldsymbol{X}_j \right)^{-1} \sum_{j=1}^{N} \boldsymbol{X}_j' \hat{\mathbf{\Sigma}}_j^{-1} \boldsymbol{y}_j$$

  where $\hat{\mathbf{\Sigma}}_j = (\boldsymbol{Z}_j \hat{\boldsymbol{T}} \boldsymbol{Z}_j' + \hat{\sigma}^2 \boldsymbol{I})$.

- The vector of fixed effects is not the same under MLE and REML (even though the REML estimations is only with respect to the variance components).

- With balanced designs, the REML estimators are the same as classical ANOVA-type ones and therefore they don't depend on assumption of normality.

Comparison of REML & ML: Std Errors

- Which is better (i.e., smaller standard errors) depends on number of marco units, number of fixed effects, and the actual value of the variance parameters.
- If N(number of fixed effects) is "large", then
  - MLE better when number of fixed effects ≤ 4.
  - REML better when number of fixed effects > 4.
- The greater the number of fixed effects, the greater the difference between REML and ML.

Comparison of REML & ML: Deviance

-2LogLike or "deviance."

These values will differ and only some conditional likelihood ratio tests are valid for REML whereas they are valid for MLE.. . . more on this in next section of notes (on statistical inference)

Model Fitting Procedures: Algorithms

- All are iterative.
- A "Parameter Space" is the possible values for the model parameters (i.e., γ's, τ's and σ^2.)

  The possible values of (consistent with HLM)

  - γ's are Real numbers
  - τ_k^2 are non-negative real numbers
  - τ_{kl} $(k \neq l)$ are Real numbers
  - σ^2 are non-negative real numbers.

  "Boundary" of the parameter space.

  Sometimes the MLE's are outside the parameter space, in which case the estimates value is set equal to the boundary value (e.g., $\tau_k^2 = 0$).

Four Major Algorithms

- Newton-Raphson.
- Fisher Scoring.
- Iterative Generalized Least Squares (IGLS).
- Expectation-Maximization.

Newton-Raphson

- Iterative algorithm for solving non-linear equations.
- SAS default (ridge-stabilized version)
- It converges to a unique maximum of the likelihood function.
- It can be used even when the parameter space if not "open;" but in this case, there is no guarantee that you have a global maximum.

Newton-Raphson: How it works

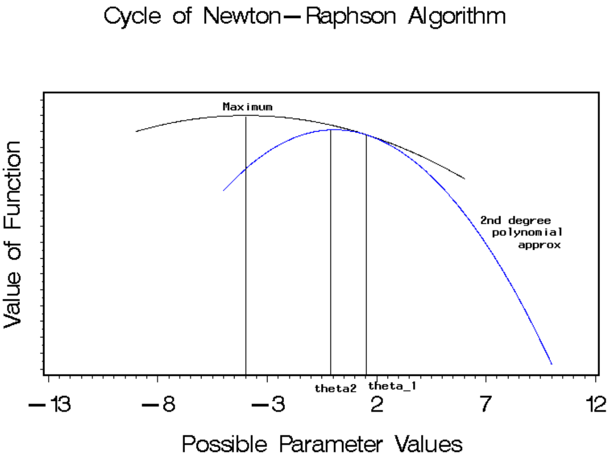
(1) Start with an initial guess of the solution (parameters).

(2) Approximates the function to be maximized in the neighborhood of the initial (current) guess by a second degree polynomial.

(3) Find the maximum of the polynomial to get better guesses for the parameters.

(4) Using new estimates, go back to step 2 and repeat until converge.

Newton-Raphson: Visual



Newton-Raphson: Some Math

Let $\boldsymbol{\theta}$ = vector of parameters, for example

$$\boldsymbol{\theta} = \left(\gamma_{00}, \gamma_{01}, \ldots, \tau_0^2, \ldots, \sigma^2\right)'.$$

- Get starting values for parameters.
- Up-date parameter estimates using

$$\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} - \boldsymbol{H}^{-1}\boldsymbol{\Delta}$$

- Check for convergence.

Newton-Raphson: Up-dating Equation

$$\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} - \boldsymbol{H}^{-1}\boldsymbol{\Delta}$$

- $\boldsymbol{\theta}_{\text{new}}$ are up-dated parameter estimates.
- $\boldsymbol{\theta}_{\text{old}}$ are the current parameter estimates.
- $\boldsymbol{\Delta}$ is "Score" vector.
  - Vector of first partial derivatives.
  - Computed using $\boldsymbol{\theta}_{\text{old}}$.
  - It should equal $\mathbf{0}$ when you have the MLEs.
- $\boldsymbol{H}$ is the "Hessian" matrix.
  - Matrix of second partial derivatives.
  - Computed using $\boldsymbol{\theta}_{\text{old}}$.
  - Sample Information matrix $= -\boldsymbol{H}$ (information matrix $\equiv -E(\boldsymbol{H})$).
  - After convergence, $-\boldsymbol{H}^{-1}$ contains estimates of the variances and covariances of parameter estimates.

Fisher Scoring

- (Bascially) Same as Newton-Raphson, but uses the Expected value of the Hessian matrix.
- Overall, preference for Newton-Raphson because it's easier to deal with sample $\boldsymbol{H}$.
- Sometimes works when Newton-Raphson doesn't, especially when fitting complex covariance structures.
- It's best that the final iterations are done by Newton-Raphson so that the final solution uses the sample Hessian matrix (i.e., the SE's are based on data rather than on expectations) — especially when there is missing data in longitudinal case (see Verbeke & Molenbergs, Chapter 21).
- If don't have random effects, then estimation of the regression models by Newton-Raphson, Fisher scoring and ordinary least squares are the same.

Iterative Generalized Least Squares

Our linear mixed model is

$$\boldsymbol{Y} = \boldsymbol{X\Gamma} + \boldsymbol{ZU} + \boldsymbol{R}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ y_{12} \\ \vdots \\ y_{n_j N} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix}, \qquad \boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \\ \vdots \\ \boldsymbol{U}_N \end{pmatrix}, \qquad \text{etc}$$

with our usual normality and independence assumptions for $\boldsymbol{U}_j$ and $\boldsymbol{R}$.

Iterative Generalized Least Squares (continued)

If we knew $\boldsymbol{T}$ and σ^2, then we could construct the covariance matrix for $\boldsymbol{Y}$,

$$\boldsymbol{V} = (\boldsymbol{Z}\boldsymbol{T}\boldsymbol{Z}' + \sigma^2\boldsymbol{I})$$

We could use Generalized Least Squares to estimate $\boldsymbol{\Gamma}$,

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

IGLS Algorithm

- Step 0: Using OLS regression to get initial estimate of $\boldsymbol{V}$.
- Step 1: Use generalized least squares (GLS) to estimate the fixed effects,
$$\hat{\boldsymbol{\Gamma}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$$
- Step 2: Compute "total" residuals $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Gamma}}$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}'$. According to our model,
$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{R}$$
So
$$\mathsf{E}(\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}') = \boldsymbol{V}.$$

IGLS Algorithm (continued)

- Step 3: Rearrange elements of $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}'$ into a vector and express these residuals as a linear model and use GLS to get new estimates of the elements of $\boldsymbol{T}$ and σ^2.
- Step 4: Check for convergence.
  - If not, go back to Step 1.
  - If yes, get estimates of standard errors of the regression parameters using the estimate of $\boldsymbol{V}$ from the least cycle (i.e., treat $\boldsymbol{V}$ as if it is known).

Little Example of IGLS: Step 1

Suppose

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij}$$

where $U_{0j} \sim \mathcal{N}(0, \tau_0^2)$ and $R_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Estimate the γ's using $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$, where

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_N,N} \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n_N,N} \end{pmatrix} \qquad \boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix}$$

and

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{\Sigma}_N \end{pmatrix} \qquad \text{where} \quad \boldsymbol{\Sigma}_j = \boldsymbol{Z}_j\boldsymbol{T}\boldsymbol{Z}_j' + \sigma^2\boldsymbol{I}_j$$

Little Example of IGLS: Step 2 & 3

- Compute vector of raw residuals,

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - (\boldsymbol{X}\hat{\boldsymbol{\Gamma}})$$

  That is

$$\tilde{y}_{ij} = y_{ij} - (\hat{\gamma}_{00} + \hat{\gamma}_{10}x_{ij})$$

- Compute cross-products, $\tilde{y}_{ij}\tilde{y}_{i'j'}$, for all i and j, which are the elements of $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}'$
- Expectations of elements of $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}'$:
  - Diagonals: $\mathsf{E}(\tilde{y}_{ij}\tilde{y}_{ij}) = \tau_0^2 + \sigma^2$
  - Same group, different individuals: $\mathsf{E}(\tilde{y}_{ij}\tilde{y}_{i'j}) = \tau_0^2$
  - Different groups: $\mathsf{E}(\tilde{y}_{ij}\tilde{y}_{i'j'}) = 0$

Little Example of IGLS: Step 3 (continued)

- Re-arrange elements of $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}'$ into a vector

$$\mathsf{E}\begin{pmatrix} \tilde{y}_{11}^2 \\ \tilde{y}_{21}\tilde{y}_{11} \\ \vdots \\ \tilde{y}_{12}\tilde{y}_{11} \\ \vdots \\ \tilde{y}_{n_N,N}^2 \end{pmatrix} = \begin{pmatrix} \tau_0^2+\sigma^2 \\ \tau_0^2 \\ \vdots \\ 0 \\ \vdots \\ \tau_0^2+\sigma^2 \end{pmatrix}$$

- We know have a linear model for τ and σ^2.

Little Example of IGLS: Step 3 (continued)

- Linear model for τ and σ^2

$$\begin{pmatrix} \tilde{y}_{11}^2 \\ \tilde{y}_{21}\tilde{y}_{11} \\ \vdots \\ \tilde{y}_{12}\tilde{y}_{11} \\ \vdots \\ \tilde{y}_{n_N,N}^2 \end{pmatrix} = \underbrace{\tau_0^2 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 1 \end{pmatrix}}_{\text{block diagonals \&}} + \underbrace{\sigma^2 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \end{pmatrix}}_{\text{diagonals of } \boldsymbol{V}}$$

- Use GLS to estimate this model.

Comments Regarding IGLS

- Can be used to get REML's.
- Implemented in MLwin.
- The normality assumption for U_j's and R_{ij}'s permits expressing the variances and covariances of the Y_{ij}'s as a linear function of the τ's and σ^2.
- The parameter estimates are the MLE's if the normality assumption holds.
- If normality does not hold, then
  - Parameters estimates are consistent, but not efficient.
  - Estimated standard errors are not consistent.

Expectation-Maximization, EM

- Although this is the "old" method of fitting variance components models, it is still useful, especially when
  - Have complicated likelihood functions.
  - Get good starting values for Newton-Raphson and/or Fisher scoring (direct likelihood maximization methods).
  - Data are not missing at random (pertains to longitudinal data).
- Incomplete data.
  - Either missing or latent.
  - In HLM context, we have latent or unobserved random variables — the U_j's and R_{ij}'s.
- Complete data.

  The observed data and the unobserved values of the random effects.

Expectation-Maximization, EM (continued)

Repeat two steps until convergence achieved:

(1) Expectation or "E–Step":

Given current values of parameters, compute the expected value of the missing data, so that you now have complete data.

(2) Maximization or "The M–Step":

Standard maximization.

Drawback: EM tends to be slow to converge.

Estimation Problems

They are caused by estimating variance components (iterative procedures needed because we have these).

Problems encountered:

- Lack of convergence
- Boundary values.
- $\hat{\boldsymbol{T}}$ not positive definite.
- Hessian not positive definite.

Things to look for and ways to deal with them. . .

According to Singer & Willet, the source of the estimation problems is data are unbalanced data.

Small Variance Components

- Variances components depend on the scale of the explanatory variables.

- We can transform explanatory variables so that variance estimates will be larger. e.g.,

  Level 1:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} x_{ij} + R_{ij} \\ &= \beta_{0j} + 10\beta_{1j} \left(\frac{x_{ij}}{10}\right) + R_{ij} \\ &= \beta_{0j} + \beta^*_{1j} x^*_{ij} + R_{ij} \end{aligned}$$

  where $R_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Small Variance Components

Level 2:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} z_j + U_{0j}\\ \beta^*_{1j} &= 10(\gamma_{10} + \gamma_{11} z_j + U_{1j})\\ &= \gamma^*_{10} + \gamma^*_{11} z_j + U^*_{1j}\end{aligned}$$

If the covariance matrix for $\boldsymbol{U} = (U_{0j}, U_{1j})'$ equals

$$\boldsymbol{T} = \begin{pmatrix} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{pmatrix},$$

Then the covariance matrix for $(U_{0j}, U^*_{1j})'$ is

$$\boldsymbol{T}^* = \begin{pmatrix} \tau_0^2 & 10\tau_{10} \\ 10\tau_{10} & 100\tau_1^2 \end{pmatrix} = \begin{pmatrix} \tau_0^2 & \tau^*_{10} \\ \tau^*_{10} & \tau_1^{2*} \end{pmatrix}.$$

- The estimated variance τ_1^{2*} is 100 times larger when you use $x_{ij}/10$ as the explanatory variable instead of x_{ij}.
- Using transformed explanatory variable x^*_{ij} moves the solution away from the "boundary" of the parameter space.

Model Miss-specification & Zero Variances

The HLM is more restrictive than the marginal model in terms of what parameters can equal, which can cause problems.

- Zero variances
- Negative variances
- Correlations greater than 1.

Example 5.3 from Snijders & Bosker.

In this example they fit a very complex model with

- 2 micro level variables.
- 3 macro level variables.
- Random intercept and 2 random slopes.

To get any estimates (i.e., so they get a solution that converges), Snijder & Bosker fixed $\tau_{12} = 0$ and estimated that $\hat{\tau}_2^2 = \hat{\tau}_{20} = 0$.

Without fixing $\tau_{12} = 0$, no convergence.

Zero Variances: SAS/MIXED

To replicate their results in SAS/MIXED, first we don't restrict τ_2^2, τ_{12} and τ_{20} to equal zero:

```
PROC MIXED data=center noclprint covtest method=ML;
CLASS schoolNR;
MODEL langPOST= oIQ_verb oses ogrpIQ ogrpsize mixedgra
     oIQ_verb*ogrpIQ oIQ_verb*ogrpsize oIQ_verb*mixedgra
     oses*ogrpIQ oses*ogrpsize oses*mixedgra /solution;
RANDOM int oIQ_verb oses / type=un sub=schoolNR;
```

Results from SAS

WARNING: Did not converge.
Covariance Parameter Values
At Last Iteration

Cov Parm	Subject	Estimate
UN(1,1)	schoolNR	7.0404
UN(2,1)	schoolNR	-0.3180
UN(2,2)	schoolNR	0.08614
UN(3,1)	schoolNR	-0.02363
UN(3,2)	schoolNR	-0.03093
UN(3,3)	schoolNR	0
Residual		39.9664

Zero Variances: SAS/MIXED

PARMS (7) (-.3) (.1) (0) (0) (0) (40) / eqcons=4, 5, 6;

- PARMS allows you to specify starting values.
- The order to give starting values shown in the previous slide, i.e.,

$$\tau_0^2, \quad \tau_{01}, \quad \tau_1^2, \quad \tau_{02}, \quad \tau_{12}, \quad \tau_2^2, \quad \sigma^2$$

- Can use either EQCONS= or HOLD= options;
    - EQCONS= Equality Constraints
    - HOLD= set values are starting values.

MIXED Results with PARMS

```
  Parameter Search
  CovP1    CovP2    CovP3    CovP4    CovP5    CovP6    CovP7       LogLike
   7.00    -0.30     0.10      0        0        0      40.00    -7545.9166

-2 Log Like
15091.8333
```

Iteration History

```
  Iteration    Evaluations        -2 Log Like       Criterion
      1             2          15089.76773914    0.00004382
      2             1          15089.51198275    0.00000067
      3             1          15089.50829394    0.00000000
```

MIXED Results with PARMS

Covariance Parameter Estimates

		Standard	Z	
Cov Parm	Estimate	Error	Value	Pr Z
UN(1,1)	7.4859	1.2511	5.98	$< .0001$
UN(2,1)	-0.6447	0.2563	-2.51	0.0119
UN(2,2)	0.1154	0.08246	1.40	0.0809
UN(3,1)	0	.	.	.
UN(3,2)	0	.	.	.
UN(3,3)	0	.	.	.
Residual	39.3518	1.2268	32.08	$< .0001$

Convergence criteria met.

OK then?

MIXED Results with PARMS

Not OK.

Checking the SAS LOG window. . .

```
NOTE: Convergence criteria met.
NOTE: Estimated G matrix is not positive definite.
```

For a much better model. . . see Table 5.4

Negative “Variances”

The ML estimates may be outside the parameter space for an HLM.

e.g., Remember that “bad” model using the HSB data?

$$\begin{aligned}(\text{math})_{ij} &= \gamma_{00} + \gamma_{10}(\text{cSES})_{ij} + \gamma_{01}(\overline{\text{SES}})_j \\ &\quad +U_{0j} + U_{1j}(\text{cSES})_{ij} + U_{2j}(\overline{\text{SES}})_j + R_{ij}\end{aligned}$$

Negative “Variances”

The implied marginal model

$$(\text{math})_{ij} \sim \mathcal{N}(\mu_j, v_j)$$

where

$$\mu_j = \gamma_{00} + \gamma_{10}(\text{cSES})_{ij} + \gamma_{01}(\overline{\text{SES}})_j$$

and

$$\begin{aligned} v_j &= \tau_0^2 + 2\tau_{01}(\text{cSES})_{ij} + 2\tau_{02}(\overline{\text{SES}})_j \\ &\quad +2\tau_{12}(\text{cSES})_{ij}(\overline{\text{SES}})_j \\ &\quad +\tau_1^2(\text{cSES})_{ij}^2 + \tau_2^2(\overline{\text{SES}})_j^2 + \sigma^2 \end{aligned}$$

- The τ's and σ^2 can be positive or negative just so long as v_j is positive; however,
- By default, SAS/MIXED restricts τ_k^2's and σ^2 to be non-negative.
- If this restriction is removed, then the model converges.

Bad Model for HSB: Input

Input:

```
PROC MIXED data=hsb noclprint covtest method=reml
     nobound ;
CLASS id;
MODEL mathach = cSES meanSES / solution;
RANDOM intercept cSES meanSES
     / subject=id type=un;
RUN;
```

“nobound” removes the restriction that τ_k^2 and σ^2 be non-negative.

Bad Model for HSB: Output

Fixed Effects		estimate	SE	Random Effects		estimate	SE
intercept	γ_{00}	12.64	.15	intercept	τ_0^2	3.08	.52
cSES	γ_{01}	2.19	.13		τ_{10}	$-.35$	.26
meanses	γ_{02}	5.83	.31	cSES	τ_1^2	.69	.28
					τ_{20}	$-.13$	.40
					τ_{21}	$-.72$	.56
					τ_2^2	-2.22	1.26
					σ^2	36.71	.63

This is a "bad" HLM — we get a good HLM if we remove U_{2j}.

Correlations Greater Than 1

We have seen examples of this:

- Random Sample from NELS88 for $N = 100$ (random slope & intercept model):

  Using the REML results...

$$\hat{\boldsymbol{T}} = \begin{pmatrix} 18.58 & .99 \\ .99 & .02 \end{pmatrix} \longrightarrow \widehat{\text{corr}} = \begin{pmatrix} 1.00 & 1.56 \\ 1.56 & 1.00 \end{pmatrix}$$

- Computer lab/homework... be on the look out for this.

Fisher Scoring in SAS

PROC MIXED . . . SCORING=<*number*> ;

From SAS/MIXED documentation:

- SCORING=<*number*> requests that Fisher scoring be used in association with the estimation method up to iteration *number*, which is by default 0. When you use the SCORING= options and PROC MIXED converges without stopping the scoring algorithm, PROC MIXED uses the expected Hessian matrix to compute approximate standard errors for the covariance parameters instead of the observed Hessian. The output from the ASYCOV and ASYCORR options is similarly adjusted.

Summary: Estimation

- Maximum likelihood equations and principle.
- Methods:
  - MLE — biased but it is important for inference.
  - REML — unbiased and sometimes work when MLE doesn't.
- Algorithms:
  - Newton-Raphson
  - Fisher scoring
  - Iterated Generalize Least Squares (IGLS)
  - EM
  - Bayesian
- Problems & Solutions....

Summary: Problems & Solutions

Numerical problems arise from the fact that we're estimating the variance components.

These problems include:

- Failure to converge.
- Estimates that do not conform to our HLM.

Possible ways to fix numerical problems:

- Transform explanatory variables.
- Allow negative "variances".
- Use REML instead of ML.
- Use Fisher scoring, IGLS or EM instead of Newton-Raphson.
- Correct model specification!
- Use Bayesian esian estimation

Bayesian Estimation

by example using the NELS data with 23 schools.

See html document on web-site: “587Work.html”.

If you want to run the R script, you will need to install STAN and the brms package. Instructions for this can be found on the web-site. Follow the instructions exactly.

The Basics

From Bayes theorem

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &\propto p(y|\theta)p(\theta) \end{aligned}$$

- y is data.
- θ is unknown parameter(s).
- $p(y|\theta)$ is sample model, the data model, or the likelihood function.
- $p(\theta)$ is the prior distribution of the parameter θ.
- $p(y)$ is the probability of data or evidence.
- $p(\theta|y)$ is the posterior distribution of the parameter given data.

Random Effects Models are Naturally Bayesian

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + R_{ij}$$

and

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + U_{0j}\\ \beta_{1j} &= \gamma_{10} + U_{1j}\end{aligned}$$

or with assumptions on random variables,

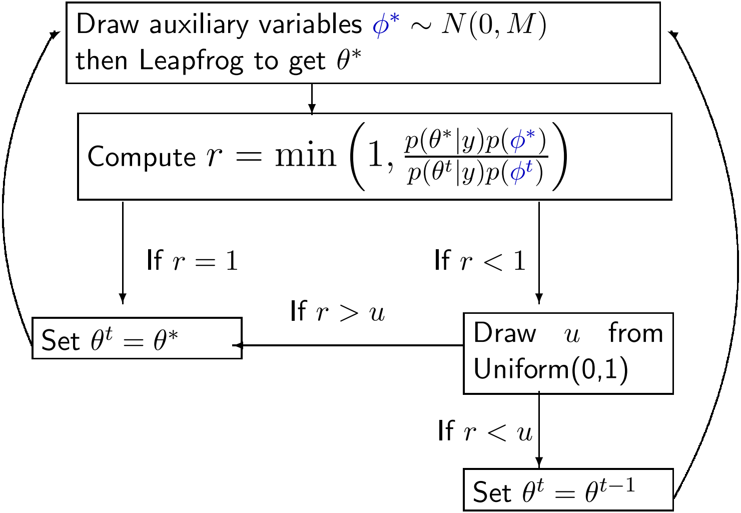
$$\begin{aligned}\beta_{0j} &\sim N((\gamma_{00}), \tau_0^2)\\ \beta_{1j} &\sim N((\gamma_{10}), \tau_1^2)\end{aligned}$$

How?

- Analytic – derive formula and "plug-n-chug"
- Grid method
- Monte Carlo Markov Chain — these are all special cases of Metropolis-Hastings Algorithm.

  The major ones:
  - Gibbs sampling (JAGS or just another gibbs sampler)
  - Metropolis
  - Hamletonian Sampling and STAN (brms is wrapper function for rstan)

Metroplois-Hastings Algorthm



Draw auxiliary variables $\phi^* \sim N(0, M)$ then Leapfrog to get θ^*

Compute $r = \min\left(1, \frac{p(\theta^*|y)p(\phi^*)}{p(\theta^t|y)p(\phi^t)}\right)$

If $r = 1$

If $r < 1$

If $r > u$

Set $\theta^t = \theta^*$

Draw u from Uniform(0,1)

If $r < u$

Set $\theta^t = \theta^{t-1}$

Example Simmulation

https://chi-feng.github.io/mcmc-demo/app.html