36-490: Undergraduate Research

Mixed Linear Models

Brian W. Junker
brian@stat.cmu.edu

Miscellaneous Business

- Your group working agreements – thanks!
- First set of presentations was scheduled for Feb 22 & 24.
  - Shall we do them Mar 1 & 3 instead?
- Homework – we haven't given any technical hw yet!
  - Would it be helpful?

Outline

- Multiple Linear Regression
- Adding a Second Random Effect
- Linear Mixed Models (LMM's) in Laird & Ware's (1982) Matrix Notation
- Facilities in R
- Computational Notes
- Predictors and Residuals
- Examples in R

Multiple Linear Regression

- A multiple linear regression model can be written as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_p + \varepsilon_i \\ &= X_i\beta + \varepsilon_i, i = 1, \ldots, n \end{aligned}$$

or in full matrix notation

$$y = X\beta + \varepsilon$$

where

- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are *fixed coefficients* to be estimated
- $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ is random and we can estimate its variance
- X is a fixed design matrix

In Multiple Linear Regression, What is Fixed and What is Random?

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

- X_{ij} are fixed and known
- β_j are fixed and unknown: ***fixed effects***
- $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ are iid random and unknown: a ***random effect***
- σ_ε^2 is the ***variance (component)*** of the random effect

Adding a Second Random Effect

- **Problem**: Batches of i=1, …, 3 widgets each come off of j=1..3 assembly lines.
  - There is too much variation between assembly lines to justify the "null" model

$$y_{ij} = \beta_0 + \varepsilon_{ij}$$

  - There is not enough variation between assembly lines to justify the one-way ANOVA model

$$y_{ij} = \beta_0 + \beta_j + \varepsilon_{ij}$$

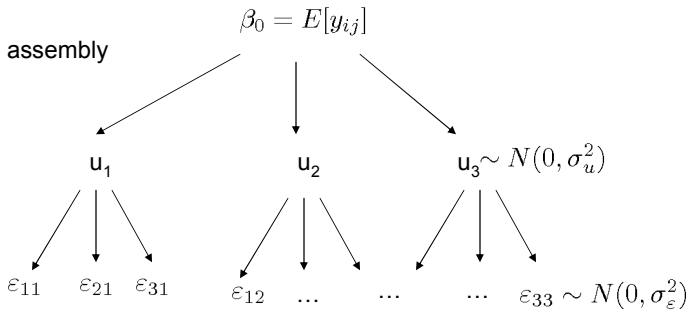
- **Solution**: introduce a per-assembly line error

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij}$$
$$u_j \sim N(0, \sigma_u^2),\ iid; \quad \varepsilon_j \sim N(0, \sigma_\varepsilon^2),\ iid$$

This corresponds to a two-stage "thought experiment":

Stage 1: Pick an assembly line

Stage 2: Pick a widget from that assembly line



Result: $y_{11} = \beta_0 + u_1 + \varepsilon_{11}, \ldots, y_{33} = \beta_0 + u_3 + \varepsilon_{33},$

Mixed Effects in Matrix Notation

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ \hline y_{12} \\ y_{22} \\ y_{32} \\ \hline y_{13} \\ y_{23} \\ y_{33} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \hline 1 \\ 1 \\ 1 \\ \hline 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \hline \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{32} \\ \hline \varepsilon_{13} \\ \varepsilon_{23} \\ \varepsilon_{33} \end{bmatrix}$$

- or just $Y = X\beta + Zu + \varepsilon$
  - β are the fixed effects
  - u are the random effects
  - X and Z are fixed, known design matrices

Laird & Ware (1982) Notation

- A general notation for linear mixed models (LMM's):

$$Y = X\beta + Zu + \varepsilon$$

  - β are fixed and unknown (***fixed effects***)
  - $X_{n\times p}$ and $Z_{n\times q}$ are fixed and known (***design matrices***)
  - $u \sim N(0, \Psi_{q\times q})$ are the ***random effects***
    - The diagonal elements ψ_{kk} of Ψ, and the residual variance are called ***variance components***
  - $\varepsilon \sim N(0, \sigma^2_\varepsilon I)$ are the usual ***residuals***

Example: Growth Curve Models

- Toxins accumulate in muscles over time as the muscles are used. There is an overall shape to the growth (say, linear) but due to individual variation, no one individual follows the overall trend exactly.
- Toxin level = general trend line + individual variation:

$$y_{ij} = \underbrace{(\beta_0 + \beta_1 t_j)}_{\text{general curve}} + \underbrace{(u_{0i} + u_{1i} t_j)}_{\text{individual variation}} + \varepsilon_{ij} = \underbrace{(\beta_0 + u_{0i})}_{\text{individual slope}} + \underbrace{(\beta_1 + u_{1i})}_{\text{individual intercept}} t_j + \varepsilon_{ij}$$

Express as a Matrix/Laird-Ware Model:

$$y_{ij} = (\beta_0 + \beta_1 t_j) + (u_{0i} + u_{1i} t_j) + \varepsilon_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i}) t_j + \varepsilon_{ij}$$

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1J} \\ \hline \vdots \\ \hline y_{I1} \\ \vdots \\ y_{IJ} \end{bmatrix} = X \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_J \\ \hline \vdots & \vdots \\ \hline 1 & t_1 \\ \vdots & \vdots \\ 1 & t_J \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 & t_1 & & \\ \vdots & \vdots & & \\ 1 & t_J & & \\ \hline & \vdots & \vdots & \\ \hline & & 1 & t_1 \\ & & \vdots & \vdots \\ & & 1 & t_J \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ \hline \vdots \\ \hline u_{1I} \\ u_{2I} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \\ \hline \vdots \\ \hline \varepsilon_{I1} \\ \vdots \\ \varepsilon_{IJ} \end{bmatrix}$$

$$Y = X\beta + Zu + \varepsilon$$

Facilities in R (& SAS)

- R (and Splus) provide two commonly-used packages for LMM analysis:
  - `library(nlme)` provides
    - `lme()` for working with $y_i = X_i\beta + Z_i u + \varepsilon_i$
    - `nlme()` for working with $y_i = f(X_i, Z_i; \beta, u_i) + \varepsilon_i$

    These can be hard to learn but are very flexible.
  - `library(lme4)` provides `lmer()`, for both LMM's and "generalized" LMM's (e.g. mixed effects logistic regression).
- For larger problems, SAS provides PROC MIXED and PROC NLMIXED.

Computational Notes

Consider the general LMM

$$Y = X\beta + Zu + \varepsilon$$

If $Cov(u, \varepsilon) = 0$, then

$$\begin{aligned} Y &\sim N(X\beta, V(\omega)) \\ V(\omega) &= \sigma^2 I + Z\Psi(\omega)Z^T \end{aligned}$$

where $u \sim N(0, \Psi(\omega))$ and $\varepsilon \sim N(0, \sigma^2 I)$. Then $-2\log(\textit{likelihood})$ is

$$(Y - X\beta)^T V^{-1}(\omega)(Y - X\beta) + \log|V(\omega)| \qquad (*)$$

- To find MLE's we can iterate between minimizing in ω given β, and minimizing in β given ω.
- This can be slow, so an approximate method called "REML" is often used by default.
- Most statistical testing, variable selection, etc., is based on MLE, not REML.

Predictors and Residuals

- For the multiple regression model

$$y_i = X_i\beta + \varepsilon_i$$

the predictor is

$$\hat{y}_i = X_i\hat{\beta}$$

and the residual is

$$r_i = y_i - \hat{y}_i$$

and so a standardized ("Pearson") residual is

$$s_i = \frac{r_i}{\widehat{SE}(r_i)}$$

Predictors and Residuals

- For the GMM,

$$y_i = X_i\beta + Z_i u_i + \varepsilon_i$$

there can be at least two predictors

$$\hat{y}_i^0 = X_i\hat{\beta} \quad \text{or} \quad \hat{y}_i^1 = X_i\hat{\beta} + Z_i\hat{u}_i$$

and so two residuals

$$r_i^0 = y_i - \hat{y}_i^0 \quad \text{or} \quad r_i^1 = y_i - \hat{y}_i^1$$

and so two std' ized ("Pearson") residuals

$$s_i^0 = \frac{r_i^0}{\widehat{SE}(r_i^0)} \quad \text{or} \quad s_i^1 = \frac{r_i^1}{\widehat{SE}(r_i^1)}$$

- Make sure you know which residual you are looking at!

R Examples

- A measurement error model for batches of dyes.
- A growth curve model for spinal bone marrow density.
- These examples only scratch the surface. See also
  - Venables & Ripley, Chapter 10
  - John Fox's notes on GMM's and hierarchical linear models (I will make available on course website)
  - Look for examples of lme and lmer on the www…
  - Ask us!