37457

Advanced Bayesian Methods

# Overview of Grouped Data Analysis

## Where Are We?

Topic 1 Probabilistic Graph Theory (COVERED)

Topic 2 Bayesian Statistical Inference (COVERED)

Topic 3 Bayesian Inference Engines (COVERED)

---

Topic 4 Advanced Bayesian Statistical Models and Analyses

First Part of Topic 4

# MODELS FOR GROUPED DATA

## Types of Grouped Data (Common Application Areas)

- Longitudinal data (medicine; public health).
- Multilevel data (education; sociology).
- Panel data (economics).
- Small area data (survey sampling).
- Item response data (psychology).

## New Concepts to Most of Class

Main modelling concepts:

random effects $\longrightarrow$ mixed models

## Longitudinal Studies

The defining characteristic of longitudinal studies is that subjects are measured repeatedly over time.

This is in contrast to cross-sectional studies where a single outcome is measured for each individual.

## New Concepts to Most of Class

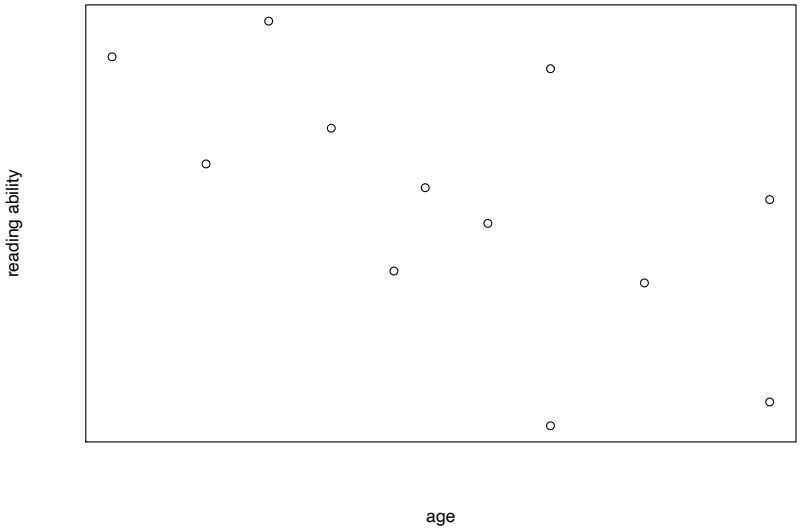
Main modelling concepts:

random effects $\longrightarrow$ mixed models

The Class 1 survey showed that:

**70%**

of our class have never seen these concepts before.



reading ability
age


reading ability
age


reading ability
age


reading ability
age
reading ability
age
reading ability
age

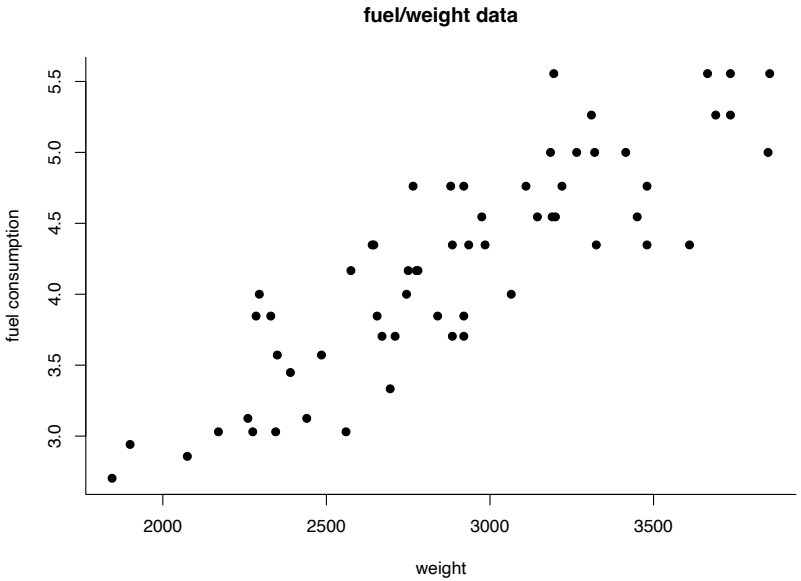
# Cross-sectional versus Longitudinal

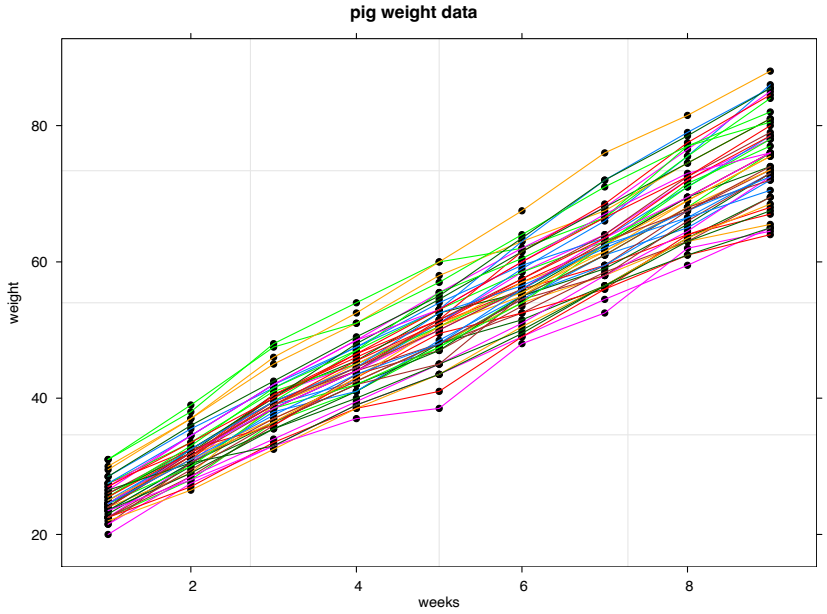
The next three slides show real data examples of

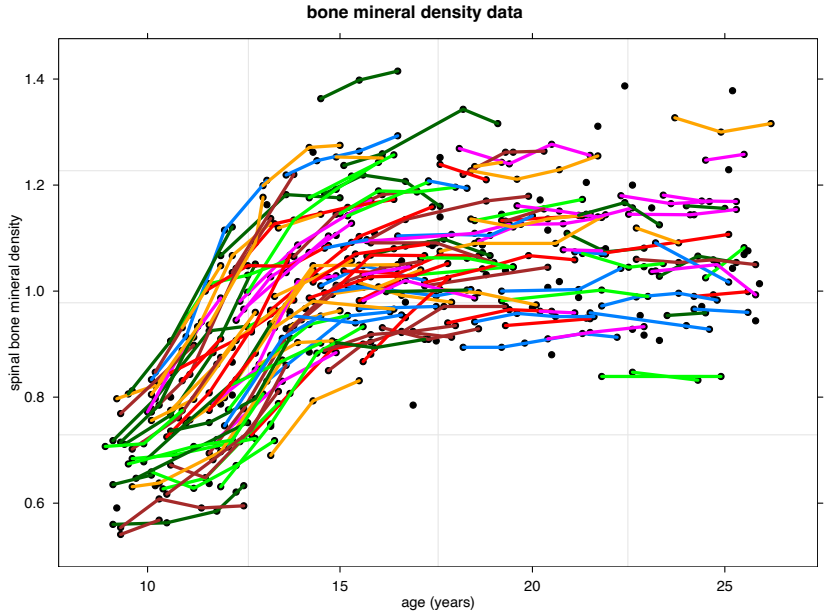
Cross-sectional data (1st slide)

and

Longitudinal data (2nd & 3rd slides)



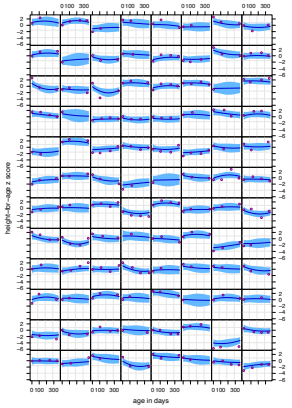



# SOME OTHER GROUPED DATA EXAMPLES

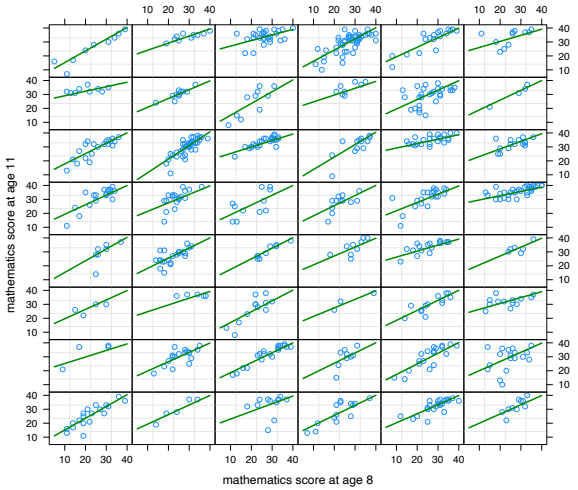
## "BIG DATA": 45 Thousand Infants

Size Versus age in first year of life:

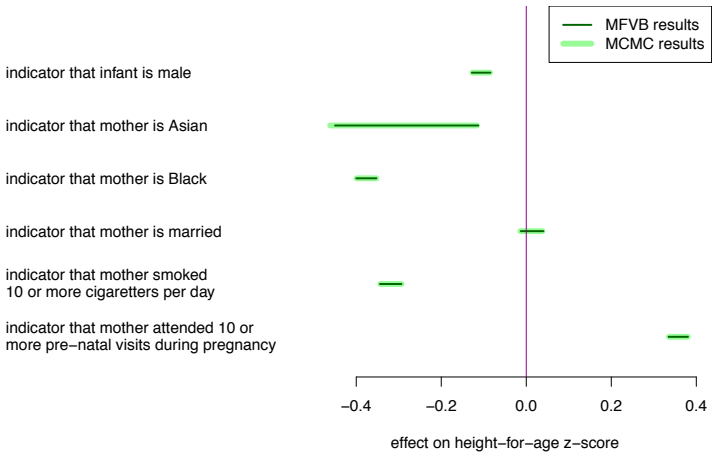



The above plot is only ONE FIVE HUNDREDTH of the full data.

## Grouping By Schools




## We Still Want Regression-Type Results for the Other Predictors (95% credible intervals shown in GREEN)



## Double Subscript Notation

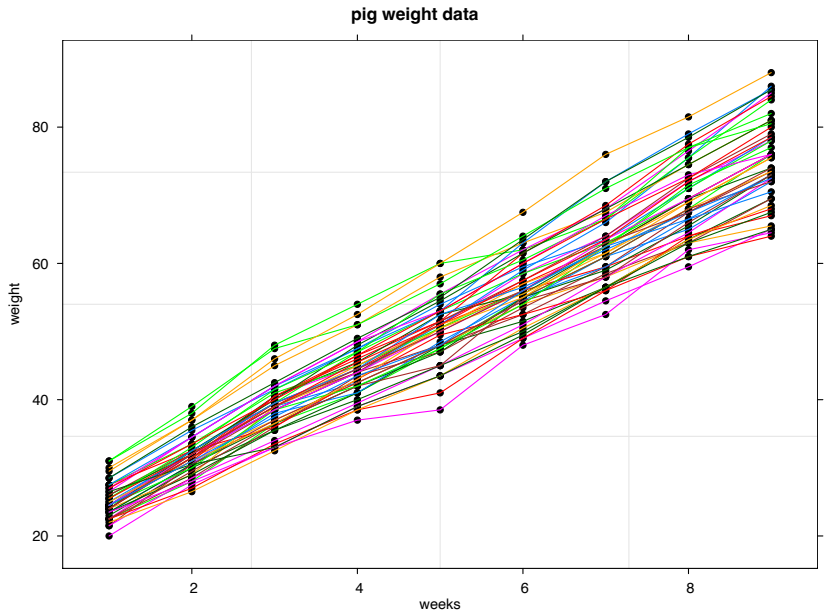
The letter $y$ is usually used for a generic response or outcome variable.

Let

$$\begin{aligned} m &= \text{number of groups} \\ n_i &= \text{number of measurements for group } i \ (1 \leq i \leq m) \end{aligned}$$

Then

$$y_{ij} = \text{response for } j\text{th measurement on group } i.$$




## Whiteboard Interlude

This is to illustrate the ideas of double subscripting.

## Naïve Model for Pig Weights

The slopes look about the same.

But each pig seems to have his/her own intercept

$$\Longrightarrow \quad y_{ij} = \beta_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}$$

for $1 \leq i \leq 48$, with $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$.

But this model has 50 parameters!:

$$\beta_{01}, \beta_{02}, \ldots, \beta_{0,48}, \beta_1 \text{ and } \sigma^2_\varepsilon.$$

And only the last 2 are interpretable.

## Random Intercept Model

A better model is:

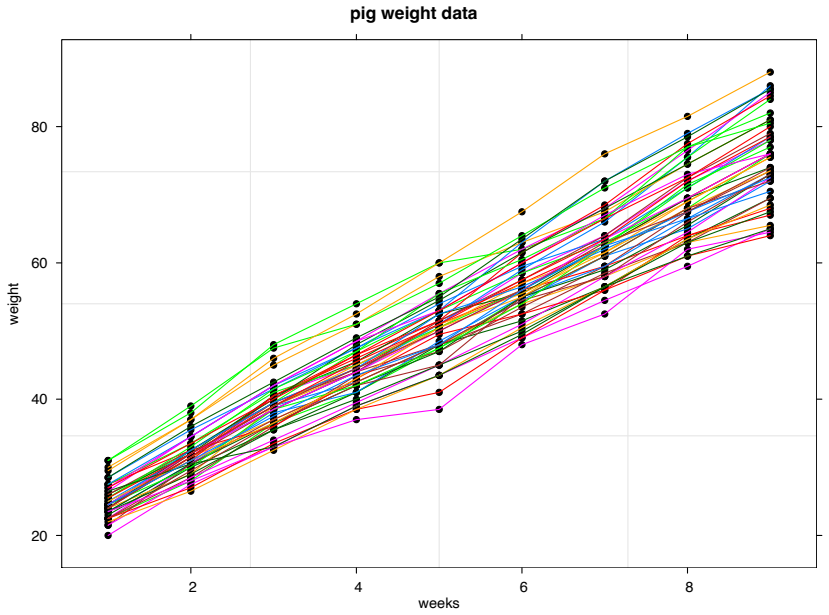
$$y_{ij} = u_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

where

$$u_i \overset{\text{ind.}}{\sim} N(0, \sigma_u^2)$$

are independent of the

$$\varepsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2).$$




## Random Effects

The $u_i$ are called

random effects

By design they are centred around zero and correspond to the $i$th pig's deviation from the 'average' intercept $\beta_0$.

## Mixed Model Terminology

$$y_{ij} = \underbrace{u_i}_{\text{random effect}} + \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{fixed effects}} + \varepsilon_{ij}$$

The right-hand side has a mixture of random effects and fixed effects and so is called a

(LINEAR) MIXED MODEL

# TECHNICAL ASIDE

## HOW MIXED MODELS INDUCE

## WITHIN GROUP

## CORRELATION STRUCTURE

We now describe the essential properties of the *random intercept model*

$$y_{ij} = u_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

$$u_i \overset{\text{ind.}}{\sim} N(0, \sigma_u^2), \qquad \varepsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$$

and the $u_i$ and $\varepsilon_{ij}$ are independent of each other.

For $j \neq j'$, $\text{Cov}(y_{ij}, y_{ij'})$ is the covariance between different measurements on the same group:

$$\begin{aligned}
\text{Cov}(y_{ij}, y_{ij'}) &= \text{Cov}(u_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \\
&\qquad u_i + \beta_0 + \beta_1 x_{ij'} + \varepsilon_{ij'}) \\
&= \text{Cov}(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ij'}) \\
&= \text{Cov}(u_i, u_i) + \text{Cov}(u_i, \varepsilon_{ij'}) + \text{Cov}(u_i, \varepsilon_{ij}) \\
&\quad + \text{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) \\
&= \sigma_u^2 + 0 + 0 + 0 \\
&= \sigma_u^2
\end{aligned}$$

For $j = j'$, $\text{Cov}(y_{ij}, y_{ij'}) = \text{Var}(y_{ij})$ is the variance of the $j$th measurement on group $i$, and has expression:

$$\begin{aligned}
\text{Cov}(y_{ij}, y_{ij}) &= \text{Var}(u_i + \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}) \\
&= \text{Var}(u_i + \varepsilon_{ij}) \\
&= \text{Var}(u_i) + \text{Var}(\varepsilon_{ij}) \\
&= \sigma_u^2 + \sigma_\varepsilon^2
\end{aligned}$$

For $i \neq i'$ we get

$$\text{Cov}(y_{ij}, y_{i'j'}) = 0.$$

This says that observations of different individuals are uncorrelated (e.g. Anne's blood pressure is not correlated with Bill's blood pressure).

Consider the sample sizes corresponding to the example:

$$m = 3, \qquad n_1 = 2, \qquad n_2 = 3, \qquad n_3 = 2.$$

The covariance matrix of

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{bmatrix}$$

is

$$\left[\begin{array}{cc|ccc|cc} \sigma_u^2+\sigma_\varepsilon^2 & \sigma_u^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2+\sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \sigma_u^2+\sigma_\varepsilon^2 & \sigma_u^2 & \sigma_u^2 & 0 & 0 \\ 0 & 0 & \sigma_u^2 & \sigma_u^2+\sigma_\varepsilon^2 & \sigma_u^2 & 0 & 0 \\ 0 & 0 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2+\sigma_\varepsilon^2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & \sigma_u^2+\sigma_\varepsilon^2 & \sigma_u^2 \\ 0 & 0 & 0 & 0 & 0 & \sigma_u^2 & \sigma_u^2+\sigma_\varepsilon^2 \end{array}\right]$$

The correlation matrix is then

$$\left[\begin{array}{cc|ccc|cc} 1 & \rho & 0 & 0 & 0 & 0 & 0 \\ \rho & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & \rho & \rho & 0 & 0 \\ 0 & 0 & \rho & 1 & \rho & 0 & 0 \\ 0 & 0 & \rho & \rho & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & \rho \\ 0 & 0 & 0 & 0 & 0 & \rho & 1 \end{array}\right] \quad \text{where} \quad \rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\varepsilon^2}.$$

Note the blocking structure corresponding to within-group correlation. The first block is for group 1, the second is for group 2 and the third is for group 3.

Remarks

1. The random intercept $u_i$ invokes correlation between measurements on same group.

2. A shortcoming of the random intercept model is that the correlation is the same for each group; e.g. Anne's $\rho$ is the same as Bill's $\rho$.

3. Another shortcoming is that the within-group correlation is constant over time; e.g. the correlation between Anne's blood pressure measurements 2 days apart is the same as those taken 10 days apart.

4. Fancier models are needed to overcome these shortcomings. For now we are just introducing linear mixed models using one of the simplest versions.

# BAYESIAN VERSION OF LINEAR MIXED MODELS

$$y_{ij}|\beta_0, \beta_1, u_i, \sigma^2_\varepsilon \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_{ij} + u_i, \sigma^2_\varepsilon)$$

$$u_i|\sigma^2_u \overset{\text{ind.}}{\sim} N(0, \sigma^2_u)$$