An introduction to Multilevel Modelling

Mark Tranmer
CCSR & Mitchell Centre for Social Network Analysis
Social Statistics
University of Manchester

The importance of data structure

• We already know how to model data using regression models. Why do we need additional techniques?

• Real data tend to violate the assumptions of:
  • Independence
  • Homogeneity of residual variance

• Often in reality, data is hierarchically structured
Example

- Population defined as being made up of a number of levels or classifications
- E.g. pupils in schools in areas

Extension of conventional models

- We might be interested in the social context in which a person lives.
- Sources of variability from different levels: Individual and context are distinct sources of variability
- Conventional models cannot accomplish this
Structure of the data

- Hierarchical structures are generated by
  - Data collection mechanism
  - ‘Natural’ structures within the population

The independence assumption
A: data collection

- Survey data rarely comes from a Simple Random Sample (SRS)
- Surveys often have multi-stage designs
- Cost advantages
- Often necessary when there is no suitable frame for households (or individuals).
- Outcome: clustered data
- i.e. the data collection process generates observations that are not independent. e.g. clustered by geography / household, etc.
Data collection

• Clustering is sometimes regarded as a nuisance, not of primary interest

• If data has been collected using a two stage design, carrying out an individual level analysis is equivalent to assuming it is a simple random sample

• But independence assumption unrealistic: e.g. expect positive correlation between exam scores of pupils from the same school

• This would underestimate standard errors

The independence assumption.
B: Structures in the population

• Even if we have collected data in an unclustered way there is still ‘natural’ clustering in the population

• Model based approach - build a model that represents the population from which the data was selected

• Impact of clustering is no longer just a nuisance, but of substantive interest
Examples of natural clustering

- People in households in areas
- Pupils within classes within schools: a pupil’s performance will not only depend on their characteristics, but also on the class they are in and the school from which the class is drawn
- Patients within wards within hospitals.

Why does clustered data matter?

- Standard analysis assumes independence and estimates standard errors of model parameters accordingly
- If observations within clusters positively correlated, this will underestimate standard errors.
- Result: variables will appear significant when in fact they are not!
- So we need to take account of clustering.
Methods for hierarchical structures

- Traditional approaches to analysing clustered data treat clustering as a nuisance that must be accounted for.
- Parameters estimated in the usual way but standard error estimates are adjusted for impact of clustering.
- Multilevel modelling: takes account of hierarchical structure and regards structure of substantive interest.

Hierarchical structures

(level 1)

- school 1
  - pupil 1
  - pupil 2
  - pupil 3
- school 2
  - pupil 1
  - pupil 2
  - pupil 3
  - pupil 4

(level 2)
Hierarchical structures: area (3); household (2); individual (1)

General framework: notation

- Level one unit: \( i \) (e.g. individual) microlevel
- Level two unit: \( j \) (e.g. area, school) macrolevel
- There are \( i = 1,...,n_j \) level one units within the \( j^{th} \) level two unit, and \( j = 1,...,J \) level two units
- (continuous) response variable \( y_{ij} \)
- explanatory variables: \( x_{0ij}, x_{1ij},...,x_{ pij} \)
Different approaches to analysing hierarchical data

• Aggregate analysis
• Disaggregate analysis
  A. Ignore the problem
  B. Fixed effects models
  C. Multilevel models

Variance components model

\[ y_{ij} = \beta_0 + u_j + e_{ij} \]

Fixed Part: \( \beta_0 \) is the overall mean

Random Part:
  \( u_j \) - group level residual  
  \( e_{ij} \) - individual level residual

\( u_j \sim N(0, \sigma_u^2), \text{cov}(u_j, u_j')=0 \)  
\( e_{ij} \sim N(0, \sigma_e^2), \text{cov}(e_{ij}, e_{ij}')=0 \)

Also \( \text{cov}(e_{ij}, u_j)=0 \)

\( \sigma_u^2, \sigma_e^2 \) variance components (group and individual).
Intra-cluster correlation coefficient

- \( \rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \)
- \( \sigma_u^2 + \sigma_e^2 \) is the total variation
- \( \sigma_u^2 \) is the variation at level 2
- \( \sigma_e^2 \) is the variation at level 1
- possible values \( 0 \leq \rho \leq 1 \)

Fixed effects vs random effects (multilevel) models

**Fixed effects:**
Better if number of groups \( (m) \) is small.
\( m-1 \) parameters to estimate group effects
If group sample sizes small group effects poorly estimated
cannot make inferences to other groups
cannot estimate effects of area level explanatory variables
Fixed effects vs random effects (multilevel) models

Random effects:
Better if $m$ large.

one parameter, $\sigma_u^2$, to estimate for group effects

Estimation performs well even if group sample sizes small

Inference to wider population of groups

Can estimate effects of area level explanatory variables

Random intercepts model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

Fixed Part: $\beta_0$, $\beta_1$

Random Part: $u_j$, $e_{ij}$
Model assumptions

$u_j$ - group level residual $e_{ij}$ - individual level residual
$u_j \sim N(0, \sigma^2_u)$, $\text{cov}(u_j, u_j') = 0$  
$e_{ij} \sim N(0, \sigma^2_e)$, $\text{cov}(e_{ij}, e_{ij}') = 0$
Also $\text{cov}(e_{ij}, u_j) = 0$

$\sigma^2_u, \sigma^2_e$ variance components (group and individual).

Visualization

[Image: Visualization of the random intercept model]
Random intercepts and slopes model: aim

\[ y_{ij} = \beta_0 + \beta_1 x_{1ij} + e_{ij} \]

Both intercept and slope now group dependent.
Individual and group level variances and covariances

- $\text{var}(e_{ij}) = \sigma_e^2$  individual level variance
- $\text{var}(u_{0j}) = \sigma_{u0}^2$  group level intercept change variance
- $\text{var}(u_{1j}) = \sigma_{10}^2$  group level slope change variance
- $\text{cov}(u_{0j}, u_{1j}) = \sigma_{u01}$  group level intercept and slope change covariance