MONASH University

MONASH BUSINESS SCHOOL

ETC3580: Advanced Statistical Modelling

Week 7: Mixed-effect models

Outline

Grouped data

Data come in groups, rather than iid:

- Survey of students, within classes, within schools
- Data on regions within states within countries
- Measurements on people over time
- Measuring different drugs on same people

Correlations between observations within the same group, so independence assumption inappropriate

Fixed and random effects

Fixed effect:

- coefficients we estimate from the data
- levels of categorical variable are non-random
- Parameters in LM and GLMs are fixed effects

Random effect:

- random variable within model
- levels of categorical variable drawn from random distribution
- estimate parameters of distribution of effect
- used to handle grouped data

Example: Estimating income by postcode

Data set consists of household incomes and postcodes.
Some postcodes have many observations, some only a couple of households.

Example: Estimating income by postcode

Data set consists of household incomes and postcodes.
Some postcodes have many observations, some only a couple of households.

Approach 1: take mean of each postcode.

- Fails with poorly sampled postcodes.

Approach 2: treat postcode as a random effect

- Shrinks individual estimates towards global mean
- Handles poorly sampled postcodes
- Closely related to hierarchical Bayesian modelling

Random effects are useful when …

- Lots of levels of a factor (categorical predictor)
- Relatively little data on some levels
- Uneven sampling across levels
- Not all levels sampled

Random effects are useful when . . .

- Lots of levels of a factor (categorical predictor)
- Relatively little data on some levels
- Uneven sampling across levels
- Not all levels sampled

Are these fixed or random?

- gender
- postcodes
- units (in student evaluation surveys)
- race

Random effects are useful when ...

- Lots of levels of a factor (categorical predictor)
- Relatively little data on some levels
- Uneven sampling across levels
- Not all levels sampled

Are these fixed or random?

- gender
- postcodes
- units (in student evaluation surveys)
- race

Somewhat controversial. Some authors say always use random effects.

Induced correlation

Suppose we have one random effect:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i = 1, \ldots, a$ and $j = 1, \ldots, n$,
$\alpha \sim N(0, \sigma_\alpha^2)$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

Induced correlation

Suppose we have one random effect:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i = 1, \ldots, a$ and $j = 1, \ldots, n$,
$\alpha \sim N(0, \sigma_\alpha^2)$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

Intra-class correlation

$$\text{Corr}(y_{ij}, y_{ik}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

General model

Error form:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \varepsilon$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ and $\gamma \sim N(\mathbf{0}, \sigma^2\boldsymbol{D})$.

General model

Error form:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \varepsilon$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ and $\gamma \sim N(\mathbf{0}, \sigma^2\boldsymbol{D})$.

Conditional distribution form:

$$\boldsymbol{y}|\gamma \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\gamma, \sigma^2\boldsymbol{I})$$

where $\gamma \sim N(\mathbf{0}, \sigma^2\boldsymbol{D})$.

General model

Error form:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \varepsilon$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ and $\gamma \sim N(\mathbf{0}, \sigma^2\boldsymbol{D})$.

Conditional distribution form:

$$\boldsymbol{y}|\gamma \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\gamma, \sigma^2\boldsymbol{I})$$

where $\gamma \sim N(\mathbf{0}, \sigma^2\boldsymbol{D})$.

Unconditional distribution form:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \sigma^2(\boldsymbol{I} + \boldsymbol{ZDZ}'))$$

Model specification

Formula	Meaning
`(1 | g)`	Random intercept with fixed mean
`(1 | g1) + (1 | g2)`	Random intercepts for both `g1` and `g2`
`x + (x | g)`	Correlated random intercept and slope
`x + (x || g)`	Uncorrelated random intercept and slope

Outline

Maximum likelihood estimation

Let $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{ZDZ}'$. Then

$$L = \frac{1}{(2\pi)^{n/2}|\sigma^2\boldsymbol{V}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)'\right\}$$

Maximum likelihood estimation

Let $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{ZDZ}'$. Then

$$L = \frac{1}{(2\pi)^{n/2}|\sigma^2\boldsymbol{V}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)'\right\}$$

So

$$\log L = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\sigma^2\boldsymbol{V}| - \tfrac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)'$$

Maximum likelihood estimation

Let $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$. Then

$$L = \frac{1}{(2\pi)^{n/2}|\sigma^2\boldsymbol{V}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)'\right\}$$

So

$$\log L = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\sigma^2\boldsymbol{V}| - \tfrac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)'$$

Optimize to find β, σ^2 and $\boldsymbol{D}$.

Maximum likelihood estimation

Let $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}'$. Then

$$L = \frac{1}{(2\pi)^{n/2}|\sigma^2\mathbf{V}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)'\right\}$$

So

$$\log L = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\sigma^2\mathbf{V}| - \tfrac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)'$$

Optimize to find β, σ^2 and $\boldsymbol{D}$.

Problems:

- biased parameters on boundaries
- non-zero derivatives at boundaries

Restricted Maximum Likelihood (REML)

- Designed to avoid MLE problems
- Find all independent linear combinations $\boldsymbol{k}$ of the response such that $\boldsymbol{k}'\boldsymbol{X} = 0$.
- Form matrix $\boldsymbol{K}$ with columns $\boldsymbol{k}$:

$$\boldsymbol{K}'\boldsymbol{y} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{K}'\boldsymbol{V}\boldsymbol{K})$$

- Maximize likelihood of $\boldsymbol{K}'\boldsymbol{y}$ (only $\boldsymbol{D}$ and σ), then find β.
- Less biased
- Implemented in `lme4::lmer()`

Estimates of random effects

$$\boldsymbol{y}|\gamma \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\gamma, \sigma^2\boldsymbol{I})$$

where $\gamma \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{D})$.

Estimates of random effects

$$\boldsymbol{y}|\gamma \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\gamma, \sigma^2\boldsymbol{I})$$

where $\gamma \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{D})$.

- γ is not estimated because it is random. But we might want to know something about the expected values.

$$\mathrm{E}(\gamma|\boldsymbol{y}) = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)$$

- Use `ranef(fit)`

Outline

Residuals

- More than one kind of fitted value, so more than one kind of residual.
- Default is to estimate ε which is most useful for model diagnostics.
- `plot` will plot residuals vs fitted values (good for spotting heteroskedasticity)
- Plotting residuals vs predictors helps in spotting nonlinearity as usual.
- qqnorm on residuals for normality check of residuals
- qqnorm on random effects for normality check on random effects

Outline

Likelihood ratio tests

- If you compare two nested models that differ only in their fixed effects, you cannot use REML. You must use MLE despite its problems.
- Assuming you use MLE, the χ^2 approximation can be seriously wrong.
- You can't test hypotheses of the form $H_0 : \sigma_\alpha^2 = 0$.
- *p*-values on fixed effects are too small, *p*-values on random effects are too large.
- `lme4` will not give you p-values
- The only reasonable approach at this stage is to use a **parametric bootstrap** or reframe as a Bayesian problem.

Bootstrap

1. Fit full model and null model to the data
2. Compute test statistic
3. Simulate pseudo-data from the null model
4. Fit both models to the pseudo-data and compute the test statistic.
5. Repeat steps 2–3 a large number of times.
6. Find proportion of times simulated test statistics are greater than actual test statistic.

Model selection

- AIC can be used provided we only compare models which differ on fixed effects, and we use full MLE (not REML)
- Comparing models with different random effects is hard due to no defined degrees of freedom.
- Probably best to go Bayesian.

Outline

Nested effects

Levels of one factor vary only within levels of another factor

- Workers within job locations
- Units within campus

Be careful: nested levels with the same labels are not the same thing.

Crossed effects

- Any non-nested effects are “crossed”.
- That is, every level of one factor can potentially interact with every level of another factor.
- Incomplete crossing occurs when not all combinations of factors exist in the data.

Multilevel models

- Models with nested (hierarchical) structure.
- Commonly used in psychology, education, and other social sciences where survey data is naturally clustered hierarchically.

Junior School Project (1988)

Variables: `student`, `class`, `school`, `gender`, `social`, `raven`, `math`, `english`, `year`

Nesting: `school:class:student`

Other variables crossed.

Outline

Longitudinal data

- Repeated measurements on each unit taken over time.
- Called "panel data" in econometrics. Called "longitudinal data" in every other discipline.
- Individuals treated as random effects
- Additional complexity of autocorrelation to address
- Differs from time series in having many units (e.g., people) but often not many observations per person.
- i.e., Longitudinal data has large N, small T; Time series data has small N, large T.

Longitudinal data

For unit (individual) i, $\boldsymbol{y}_i$ is a T-vector such that

$$\boldsymbol{y}_i | \gamma_i \sim N(\boldsymbol{X}_i \beta + \gamma_i, \sigma^2 \boldsymbol{\Lambda}_i)$$

- $\gamma_i \sim N(0, \sigma^2 D)$ is effect of ith unit
- $\boldsymbol{X}_i$ contains predictors for fixed effects
- $\boldsymbol{\Lambda}_i$ handles autocorrelations within units
- $\boldsymbol{y}_i \sim N(\boldsymbol{X}_i \beta, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\Sigma}_i = \sigma^2(\boldsymbol{\Lambda}_i + \boldsymbol{D})$
- Assume individuals are independent, and random effects and errors are uncorrelated.

Longitudinal data

Combining individuals (assuming independence):

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix}$$

$$\Sigma = \mathrm{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_N),$$

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \Sigma)$$

Longitudinal data

Combining individuals (assuming independence):

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix}$$

$$\Sigma = \text{diag}(\Sigma_1, \Sigma_2, \dots, \Sigma_N),$$

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \Sigma)$$

- Only additional complication is choosing correlation structure
- Other random effects can be added; then γ_i becomes a vector.