



ETC3580: Advanced Statistical Modelling

Week 7: Mixed-effect models

Outline

1 Random effects

2 Estimation

3 Diagnostics

4 Inference

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 $i = 1, \dots, a$ and $j = 1, \dots, n$,

$\alpha \sim N(0, \sigma_\alpha^2)$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

Induced correlation

Suppose we have one random effect:

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

where $i = 1, \dots, a$ and $j = 1, \dots, n$,

$\alpha \sim N(0, \sigma_\alpha^2)$ 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:

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

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

General model

Error form:

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

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

Conditional distribution form:

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

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

General model

Error form:

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

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

Conditional distribution form:

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

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

Unconditional distribution form:

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

Model specification

Formula	Meaning
$(1 \mid g)$	Random intercept with fixed mean
$(1 \mid g1) + (1 \mid g2)$	Random intercepts for both $g1$ and $g2$
$x + (x \mid g)$	Correlated random intercept and slope
$x + (x \parallel g)$	Uncorrelated random intercept and slope

Outline

1 Random effects

2 Estimation

3 Diagnostics

4 Inference

Maximum likelihood estimation

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

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

Maximum likelihood estimation

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

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

So

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

Maximum likelihood estimation

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

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

So

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

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

Maximum likelihood estimation

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

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

So

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

Optimize to find β , σ^2 and \mathbf{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 \mathbf{k} of the response such that $\mathbf{k}'\mathbf{X} = 0$.
- Form matrix \mathbf{K} with columns \mathbf{k} :

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

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

Estimates of random effects

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

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

Estimates of random effects

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

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

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

$$E(\gamma|\mathbf{y}) = \mathbf{DZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

- Use `ranef(fit)`

Outline

1 Random effects

2 Estimation

3 Diagnostics

4 Inference

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

1 Random effects

2 Estimation

3 Diagnostics

4 Inference

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.