MONASH University

MONASH BUSINESS SCHOOL

ETC3580: Advanced Statistical Modelling

Week 8: Mixed Effect Models for non-Gaussian Responses

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):

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

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

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

Longitudinal data

Combining individuals (assuming independence):

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

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

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

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

Outline

Generalized Linear Mixed Models

- Combine GLMs with random effects
- y_i from exponential family distribution $f(y_i, \theta_i, \phi)$
- $\mathrm{E}(y_i) = \mu_i$
- Link function g: $g(\mu_i) = \boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\gamma$
- β are fixed effects; γ are random effects.
- $\gamma \sim N(\mathbf{0}, \boldsymbol{D})$ with density $h(\gamma|\boldsymbol{D})$

Generalized Linear Mixed Models

- Combine GLMs with random effects
- y_i from exponential family distribution $f(y_i, \theta_i, \phi)$
- $\mathrm{E}(y_i) = \mu_i$
- Link function g: $g(\mu_i) = \boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\gamma$
- β are fixed effects; γ are random effects.
- $\gamma \sim N(\mathbf{0}, \boldsymbol{D})$ with density $h(\gamma|\boldsymbol{D})$

Likelihood

$$L(\beta, \phi, \boldsymbol{D}) = \prod_{i=1}^{n} \int f(y_i|\beta, \phi, \gamma) h(\gamma|\boldsymbol{D})$$

- Can only solve integrals if f and h both normal

Numerical integration of likelihood

Likelihood

$$L(\beta, \phi, \mathbf{D}) = \prod_{i=1}^{n} \int f(y_i | \beta, \phi, \gamma) h(\gamma | \mathbf{D})$$

- Use numerical integration to approximate integrals
- More accurate than PQL
- Can be slow or impossible for complex models
- Inference will be problematic, as for MLE with LMMs

Penalized Quasi Likelihood

1. Transform fitted values:

$$\eta_i = g(\mu_i) = \mathbf{x}_i'\beta + \mathbf{z}_i'\gamma$$

2. Create pseudo-responses:

$$\tilde{y}_i^j = \hat{\eta}_i^j + (y_i - \hat{\mu}_i^j) \left.\frac{d\eta}{d\mu}\right|_{\hat{\eta}_i^j}$$

   where j is iteration in optimization algorithm

3. Find $\mathrm{Var}(\tilde{y}_i|\gamma)$
4. Use weighted linear mixed effects models

Penalized Quasi Likelihood

1. Transform fitted values:

$$\eta_i = g(\mu_i) = \mathbf{x}_i'\beta + \mathbf{z}_i'\gamma$$

2. Create pseudo-responses:

$$\tilde{y}_i^j = \hat{\eta}_i^j + (y_i - \hat{\mu}_i^j) \left.\frac{d\eta}{d\mu}\right|_{\hat{\eta}_i^j}$$

   where j is iteration in optimization algorithm

3. Find $\mathrm{Var}(\tilde{y}_i | \gamma)$
4. Use weighted linear mixed effects models

- fast
- approximate inference
- biased estimates, esp. for binary data or low counts
- even worse inference than regular LMM

Bayesian methods

- Much more accurate inference
- Allow for prior information and flexibility
- Usually take more computation
- Inferential form different
- Require additional software (either INLA or STAN).