Generalized Linear Mixed Models

Recall: For continuous response data we have discussed two related approaches to regression analysis. One approach is based on specification of means and covariances, a second approach constructs a complete likelihood for the response vectors.

Correlated Continuous Response Data

⋆ SEMI-PARAMETRIC APPROACH:

- Model: General Linear Model
  - $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i \boldsymbol{\beta}$
  - $\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{\Sigma}_i$

- Estimation: Weighted Least Squares
  - solve $\sum_i \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) = \boldsymbol{0}$
  - $\widehat{\boldsymbol{\beta}} = (\sum_i \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i)^{-1} \sum_i \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Y}_i$
  - $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{-1}$
  - simple moment estimation for $\widehat{\boldsymbol{\Sigma}}_i$

Correlated Continuous Response Data

$\star$ PARAMETRIC APPROACH:

- **Model:** Linear Mixed Model
  - $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$
  - $\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{\Sigma}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i$
  - mean/covariance *induced* from:
  - conditional mean $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$
  - heterogeneity model $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$

- **Estimation:** Maximum Likelihood & REML
  - solve $\sum_i \boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{0}$
  - $\text{cov}(\widehat{\boldsymbol{\beta}}) = (\sum_i \boldsymbol{X}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i)^{-1}$
  - ML/REML equations for $\boldsymbol{\alpha}$, variance components of the matrices $\boldsymbol{D}$ and $\boldsymbol{R}$.

$\star\star\star$ Additional features:

(1) Likelihood ratio tests

(2) Empirical Bayes estimates of $\boldsymbol{b}_i$

(3) Complete probability model (ie. we could simulate $\boldsymbol{Y}_i$).

Generalized Linear Mixed Models

Parallel approaches exist for generalized linear models:

Correlated Discrete Response Data

$\star$ SEMI-PARAMETRIC APPROACH:
(Generalized Estimating Equations)

- $\boxed{\text{Model:}}$ Marginal Model
  - $E(\boldsymbol{Y}_{ij} \mid \boldsymbol{X}_i) = \mu_{ij}$
    $g(\mu_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}$
  - $\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{\Sigma}_i = \boldsymbol{V}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{V}_i^{1/2}$
    $\boldsymbol{R}_i(\boldsymbol{\alpha}) =$ "working correlation"
- $\boxed{\text{Estimation:}}$ Estimating Eqns / Sandwich Variance
  - solve $\sum_i \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}^T \right) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}$
  - $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$
  - moment estimation of $\boldsymbol{\alpha}$
  - $\boldsymbol{B}$ estimated empirically

Correlated Discrete Response Data

⋆ PARAMETRIC APPROACH:
(Generalized Linear Mixed Models)

- $\boxed{\textsf{Model:}}$ Conditional Model + Heterogeneity

  ◦ $[\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i] \sim$ exponential family

  ◦ $E(\boldsymbol{Y}_{ij} \mid \boldsymbol{X}_i, \boldsymbol{b}_i) = \mu_{ij}^b$

  $$g(\mu_{ij}^b) = \boldsymbol{X}_{ij}\boldsymbol{\beta}^* + \boldsymbol{Z}_{ij}\boldsymbol{b}_i$$

  ◦ <u>heterogeneity</u> model $\boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D})$

  ◦ Conditional independence:
  $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ independent given $\boldsymbol{b}_i$.

Correlated Discrete Response Data

$\star$ PARAMETRIC APPROACH: (GLMMs)

• Estimation: Maximum Likelihood & Bayes

$\Rightarrow$ The likelihood of the observed data, $\boldsymbol{Y}_i$, is obtained by integrating over the random effects ($\boldsymbol{b}_i$) distribution. In general, this integration can not be done analytically.

$\circ$ $P(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \int_b P(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i) P(\boldsymbol{b}_i \mid \boldsymbol{X}_i) db_i$

$\circ$ Maximize $\log \mathcal{L}$ numerically

+ Quadrature Methods (ie. Gauss-Hermite)
+ EM (expectation-maximization)
+ Monte-Carlo methods

$\circ$ Approximate ML methods

+ MQL (Zeger, Liang and Albert, 1988)
+ PQL (Breslow & Clayton, 1993)

$\circ$ MCMC Approaches (Bayes)

+ Gibbs sampling (Zeger & Karim, 1991)
+ General MCMC

⋆ Linear Mixed Models

- General definition
  - $\boldsymbol{Y}_i = \boldsymbol{X}_i + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{e}_i$
  - $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$
- Clustered data models
- Serial covariance models (Diggle, 1988)
- Interpretation of variance components
- Estimation for LMM
  - Maximum Likelihood (ML)
  - Restricted Maximum Likelihood (REML)
- Inference for the LMM
  - Likelihood ratio tests

    Hypotheses regarding variance components

    Hypotheses regarding regression parameter
  - Wald tests
  - F tests
- Empirical Bayes estimates
  - Estimates of $E[\boldsymbol{b}_i \mid \boldsymbol{Y}_i]$
- Evaluation of covariance assumptions
  - Compare fitted and observed covariance
  - Likelihood ratio, AIC, BIC
  - Evaluate impact on $\widehat{\boldsymbol{\beta}}$, $s.e.(\widehat{\boldsymbol{\beta}})$
- Fitting LMMs

- S+ function `lme()`
- SAS procedure `MIXED`

- Analysis of Residuals
  - Population residuals
  - Cluster residuals

$\boxed{\star}$ Discrete Response Data – GEE

- Impact of ignoring correlation
  - Between- and Within- cluster covariates
  - Sandwich variance, $\text{var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\ \ \boldsymbol{A}^{-1}$

Marginal mean

  - $\mu_{ij} = E[Y_{ij} \mid \boldsymbol{X}_{ij}]$

Correlation model

  - $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i = \boldsymbol{V}_i^{1/2}\boldsymbol{R}(\boldsymbol{\alpha})\ \ \boldsymbol{V}_i^{1/2}$
  - "Working Correlation"

- Semi-parametric model (only mean and covariance)
- Asymptotic properties of $\widehat{\boldsymbol{\beta}}$
  - $\widehat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$ even under cov misspecification
  - $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{H}_N)$
  - Sandwich is consisent estimate of $\lim N \cdot \boldsymbol{H}_N$
- Estimation
  - Estimating function, $\boldsymbol{U}(\boldsymbol{\beta})$
  - Simple moment estimates for $\boldsymbol{\alpha}$
- Efficiency and working correlation models
  - IEE versus WEE

- Attempt to approximate $\text{cov}(\boldsymbol{Y}_i)$
- Inference
  - Wald tests
  - Score tests
- Caveats with time-dependent covariates
- GEE extensions
  - GEE with second covariance parameter EE
  - Odds ratio dependence models for binary data
  - ALR
  - GEE2

    Optimal for $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$

    ML for QEF

$\boxed{\star}$ Discrete Response Data – GLMM

- Model definition
  - Conditional distribution:

$$E[\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i]$$

  - Population heterogeneity model:

$$\boldsymbol{b}_i \mid \boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$$

- Regression parameter interpretation
  - Conditional expectation
  - "control" all covariates, including $\boldsymbol{b}_i$
- Covariance parameter interpretation

- Estimation
  - Maximum likelihood for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$

    Numerical Integration

    EM

    Monte Carlo
  - Empirical Bayes estimates
- Marginal and Conditional Regressions
  - Induced marginal means:

$$E[Y_{ij} \mid \boldsymbol{X}_i] = E\left(E[Y_{ij} \mid \boldsymbol{X}_i, \boldsymbol{b}_i]\right)$$

  - Attenuation?

Inference

Likelihood ratio

Wald, Score tests

$\boxed{\star}$ **Other Topics...**

- Categorical Data Likelihood Methods
- Missing Data / Drop-out
- Transition Models
- Bayes / MCMC Methods
- Non-linear Models (PK/PD)

The BIG Picture

- Generalized linear models
  - Models for the mean response
  - Univariate response / independent
- Multinomial models
  - Models for the mean response (transformed)
  - Univariate response / independent
- Overdispersed GLMs
  - Models for the mean response
  - Models for the variance
  - Univariate response / independent
- General Linear Model for Correlated Data
  - Models for the mean response (continuous)
  - Models for the covariance
  - Vector response / dependent within
- Linear Mixed Model
  - Models for the mean response (continuous)
  - Models for the covariance (hierarchical)
  - Vector response / dependent within
- Marginal GLM / GEE
  - Models for the mean response (discrete,continuous)
  - Models for the correlation
  - Vector response / dependent within
- GLMM

- Models for the conditional mean
  response (discrete,continuous)
- Models for the heterogeneity (hierarchical)
- Vector response / dependent within

The BIG Picture

	SEMI-PARAMETRIC	PARAMETRIC
Overdispersion	Quasilikelihood Est. Eq. $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\ \ \boldsymbol{A}^{-1}$	beta-binomial poisson-gamma likelihood
Continuous Resp. / linear model	WLS Est. Eq. $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\ \ \boldsymbol{A}^{-1}$	multiv. normal LMM likelihood
Discrete Response / GLM	GEE Est. Eq. $\text{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\ \ \boldsymbol{A}^{-1}$	multiv. dist. GLMM likelihood