Biostat 743 Lecture 10/12/17

Mark Fulginiti, Reviewed by: Jon Moyer

October 15, 2017

Outline

- Clustered Data
  - Notation
  - Example
- Marginal Models
- Generalized Estimating Equations (GEE)
  - Sandwich Estimator
- Generalized Linear Mixed Models

Clustered Data

- Set of observations in correlated sets (or clusters), e.g., individuals in a family, geographical unit, hierarchical data, teeth in mouth, repeated measures on units (longitudinal data)
- Why do we need to think of this data differently than that of independent data?
  - Observations in clusters breaks our assumption of independence
  - Clustered data gives less information than independent data

Notation

Y_{ij} = observations/outcome for unit (individual) i at observation j

$$j = 1,\ 2,\ ...,\ n_i$$

$$i = 1,\ 2,\ ...,\ n$$

$$N = \sum n_i$$

Example

Y_{ij} is the number of cases of some disease X in state i and district j.

$$Treatment\ \ Trt_{ij} = \begin{cases} 1, & \text{if } Trt \text{ is in i, j} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{ij} \sim Poisson(\lambda_{ij}, P_i)$$

$$log(\lambda_{ij}) = \beta_0 + \beta_1 Trt_{ij}\ + \beta_2 X$$

- X is a possible confounder; e.g., past incidence or socioeconomic status

- We want to make inference about β_1. What is the treatment effect?

- Need to adjust for clustering because Y_{ij} is not independent from $Y_{ij'}$
- One approach to adjusting for clustering: marginal models

Marginal Models

- Directly model $E(Y_{ij}|X_{ij})$
- Effects can be interpreted as population level average effects

Ingredients for Marginal Models

1. Model expected response linked to covarates using a GLM

$$E(Y_{ij}|X_{ij}) = \mu_{ij}$$
$$g(\mu_{ij}) = \eta(X_{ij}) = X_{ij}\beta$$

2. Marginal variance is a function of the marginal mean

$$Var(Y_{ij}|X_{ij}) = \phi v(\mu_{ij})$$

- where ϕ is a scale parameter and $v(\mu_{ij})$ a known function.
- e.g.,

(a) Gaussian/continuous outcome: $v(\mu_{ij}) = 1$, $\phi = \sigma^2$

(b) Poisson/binomial GLM: $v(\mu_{ij}) = \mu_{ij}$ *(poisson)*, $\phi = 1$ $(\phi > 1$ *in overdispersed model)*

3. Specify within unit correlation as a function of μ_{ij} and maybe other parameters

- e.g., $Cov(Y_{ij}, Y_{ij'}) = sd(Y_{i\ j}) * Corr(Y_{i\ j}, Y_{ij'}) * sd(Y_{ij'})$ where $Corr(Y_{ij}, Y_{ij'})$ is parameterized.

Examples of complete specification of marginal model

- **Normal outcome**:

$$1.\ \ \mu_{ij} = X_{ij}\beta$$
$$2.\ \ v(\mu_{ij}) = \phi$$
$$3.\ \ Corr(Y_{ij}, Y_{ij'}) = \alpha_{jj'}$$

- Unstructured correlation and $\alpha_{jj'}$ is different for every pair of (j, j').
- **Covariance structure of data**
- For n_i , $i = 1, 2, 3, ...,$

$$V = \phi \begin{pmatrix} R_{n_1} & 0 & 0 & ... \\ 0 & R_{n_2} & 0 & ... \\ 0 & 0 & R_{n_3} & ... \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \text{ where, say... } R = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & 1 & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & 1 \end{pmatrix}$$

$$\text{or, say... } R = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix} \text{ (compound symmetry)}$$

or, say... $R =$ exponential correlation structure, etc...

or, say... $R =$ autoregressive correlation structure, etc...

or, say... $R =$ unstructured correlation structure, etc...

or, say... $R =$ Toeplitz correlation structure, etc...

etc...

etc...

- **Poisson/count outcome**:

1. $ln(\mu_{ij}) = X_{ij}\beta$

2. $v(\mu_{ij}) = \phi\ \mu_{ij}$

overdispersed model if $\phi > 1$

3. $Corr(Y_{ij}, Y_{ij'}) = \alpha$

Generalized Estimating Equations (GEE)

- Quasilikelihood approach
- Marginal models can be fit using GEE

$$\mu(\beta) = \sum_{i=1}^{N} \left(\frac{\delta\mu_i}{\delta\beta}\right)^T V_i^{-1}(y_i - \mu_i) = 0, \quad where \quad \mu_i = g^{-1}(X_i\beta)$$

- In practice, V_i^{-1} is a scalar value; working covariance matrix - means approximate, i.e., know not exact truth

Key Points about GEE

1. β estimates are equivalent to ML estimates from GLM's when data are from an exponential family

2. Even though this is quasilikelihood, there still are nice asymptotic results: $\hat{\beta} \to \beta$ even if variance is unspecified, estimate $\hat{\beta}$ ignoring correlation structure and instead use...

3. *Sandwich Estimator* for variance to get our SE's to account for possible misspecification of variance

- **Model based covariance of estimator:**

$$V_{model} = \left[\sum_{i=1}^{n} \left(\frac{\delta\mu_i}{\delta\beta}\right)^T V(\mu_i)^{-1} \left(\frac{\delta\mu_i}{\delta\beta}\right)\right]^{-1}$$

- in practice, we use the sandwich estimator:

$$V_{sandwich} = V_{model} * \left[\sum_{i=1}^{n} \left(\frac{\delta\mu_i}{\delta\beta}\right)^T * [v(\mu_i)]^{-1} * var(Y_i) * [v(\mu_i)]^{-1} \left(\frac{\delta\mu_i}{\delta\beta}\right)\right] * V_{model}$$

- $var(Y_i)$ is the data driven estimate of variance

4. Interpretations of $\beta's$ are for the population as a whole.

- GEE setting: $g\left(E[Y_{ij}|X_{ij}]\right) = X_{ij}\beta$ where $g(.)$ is the link
- Correlations are considered "nuisance" parameters
- Allows us to "fix" SE's post-hoc

Generalized Linear Mixed Models (GLMM's)

- Just a different way of accounting for correlation structure
- Model the correlations directly as part of our mean model

$$g\left(E[Y_{ij}|X_{ij}, b_i]\right) = X_{ij}\beta + Z_{ij}b_i \text{ , where } X_{ij}\beta \text{ are fixed effects and } Z_{ij}b_i \text{ are random effects}$$

- Simple case: $Z_{ij} = 1 \implies Z_{ij}b_i = \alpha_i$, where the α_i are a set of random intercepts, $\alpha_i \sim N(0, \tau^2)$
- In a marginal model, when we don't have random effects, the $\beta's$ refer to overall population level effect, but when we use random intercepts, we have unit/subject specific interpretation or treatment effect e^{β}.