BIOS 755: Generalized Estimating Equations (GEEs) or Marginal Models for Longitudinal Data

Alexander McLain

February 16, 2023

Longitudinal model for non-normal data

- Longitudinal model for normal data are heavily influenced by the multi-variate normal (MVN) distribution
- In fact, the MVN distribution makes most of what we did in our linear models section possible.
- The MVN distribution gives us a way to relate multiple variables through their covariances.
- With non-normal data this isn't as easy.
  - We would have to assume higher order relationships, i.e., How does $P(Y_{i2} = 1 | Y_{i1} = 1)$ vary by Y_{i3}
- What are we going to do?

Marginal Models

- One approach is to specify the marginal distribution at each time point:

$$Y_{ij} \quad \text{for } j = 1, \ldots, n_i$$

  along with some assumptions about the covariance structure of the observations.
- Marginal models avoid some of the distributional assumptions that are used with other methods (e.g., mixed models).
- They don't make a complete assumption on the full distribution, why they are "marginal".
- Marginal models are conditional on the covariates and the covariance structure only (i.e., no random effects needed), which is what gives them their population averaged interpretation.

Marginal Models

- The basic premise of marginal models is to make inferences about population averages.
  - What is happening to the average? vs What is happening to each subject?
- Marginal models will look at the impact of exposures on group A vs group B, instead of the impact of an exposure of a subject changing from group A to group B.
- For linear models all coefficients had the same interpretation, for GLM's this is no longer the case (we'll discuss this more later).
- Marginal models are primarily used to make inferences on the the impact covariates have on the population.

Assumptions of Marginal Models

- With marginal models we make the following assumptions:
  1. The marginal expectation of the response, $E(Y_{ij}) = \mu_{ij}$, depends on explanatory variables, X_{ij}, through a known link function

$$\eta_{ij} = g(\mu_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}$$

  2. The marginal variance of Y_{ij} depends on the marginal mean according to

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

     where $v(\mu_{ij})$ is a known 'variance function' and ϕ is a scale parameter that may need to be estimated. (You'll have limited impact on this portion)
  3. The covariance between Y_{ij} and Y_{ik} is a function of the means and additional correlation parameters α that will also need to be estimated. (Similar to covariance pattern models.)

Examples of Marginal Models

Continuous responses:

1. $\mu_{ij} = \eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}$, i.e., linear regression
2. $Var(Y_{ij}) = \phi$, i.e., homogeneous variance.
3. $Corr(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$ $(0 \leq \alpha \leq 1)$, i.e., autoregressive correlation.

Examples of Marginal Models

Binary responses:

1. $logit(\mu_{ij}) = \eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}$, i.e., logistic regression
2. $Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$, i.e., Bernoulli variance.
3. $Corr(Y_{ij}, Y_{ik}) = \alpha_{jk}$ $(0 \leq \alpha_{jk} \leq 1)$, i.e., unstructured correlation.

Examples of Marginal Models

Count responses:

1. $\log(\mu_{ij}) = \eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}$, i.e., Poisson regression
2. $Var(Y_{ij}) = \mu_{ij}\phi$, i.e., extra-Poisson variance.
3. $Corr(Y_{ij}, Y_{ik}) = \alpha$ $(0 \leq \alpha \leq 1)$, i.e., compound symmetry correlation.

Similarities with GLMs

- The assumptions of marginal models are similar to Generalized Linear Models (GLMs).
  - both have a systematic component
  - both have a formula
  - the variance of both is **usually** specified by a distribution (i.e., Bernoulli, Poisson, etc.)
- Marginal models add a covariance structure to the specification.
  - They're similar to a GLM with covariance.

Similarities with GLMs

- The assumptions of marginal models are similar to Generalized Linear Models (GLMs).
    - both have a systematic component
    - both have a formula
    - the variance of both is **usually** specified by a distribution (i.e., Bernoulli, Poisson, etc.)
- Marginal models add a covariance structure to the specification.
    - They're similar to a GLM with covariance.
- Marginal models don't specify a distribution, except through the relationship between the mean and the variance. (not likelihood based)
- Marginal models actually don't specify the entire joint distribution of the data.

Interpretations in a Marginal World

- The regression parameters β have 'population-averaged' interpretations:
  - describe the effect of covariates on the average responses
  - think of them as comparing the means in sub-populations
- In linear regression, what happens between the groups is the same as what would happen to an individual going from one group to the other. Here, that's not the case.
- The increase in probability of a heart attack between 40 year olds and 50 year olds is not the same as an individuals increase in the probability of a heart attack when aging from 40 to 50.

Estimating Marginal Models

- Unfortunately, with discrete response data there is no analogue of the multivariate normal distribution.
- In the absence of a 'convenient' likelihood function for discrete data, there is no unified likelihood-based approach for marginal models.
- Recall: In linear models for normal responses, specifying the means and the covariance matrix fully determines the distribution of the data and the likelihood.
- This is not the case with discrete response data.

Generalized Estimating Equations

- Since there is no 'convenient' or natural specification of the joint multivariate distribution of $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in}$ for marginal models when the responses are non-normal, we use an alternative to maximum likelihood (ML) estimation.
- Liang and Zeger (1986) and Zeger, Liang and Albert (1988) proposed such a method based on the concept of 'estimating equations.' This work comes from:
  - Wedderburn (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method, *Biometrika*, 61: 439–447
  - McCullagh (1983) Quasi-Likelihood Functions, *The Annals of Statistics*, 11: 59–67
- This provides a general and unified approach for analyzing discrete and continuous responses with marginal models.

Generalized Estimating Equations

- The essential idea was to generalize the usual univariate (i.e., cross-sectional) likelihood equations by introducing the covariance matrix of the vector of responses.
- For linear models, generalized least squares (GLS) can be considered a special case of this 'estimating equations' approach.
- For non-linear models, this approach is called 'generalized estimating equations' (or GEE).

Fitting Marginal Models

- Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$ be a vector of correlated responses for the ith subject $(i = 1, \ldots, N)$.
- Then, an estimate of $\boldsymbol{\beta}$ can be obtained as the solution to the following generalized estimating equations

$$\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0} \tag{1}$$

  where $\boldsymbol{D}_i = \partial \mu_i / \partial \boldsymbol{\beta}$
- V_i is a ‘working’ covariance matrix, i.e. $\boldsymbol{V}_i \approx Cov(Y_i)$, which is a function of ϕ and α.

Fitting Marginal Models

- Generalized estimating equations depend on $\boldsymbol{\beta}$, ϕ and α.
- Because the generalized estimating equations depend on both $\boldsymbol{\beta}$ and α, an iterative two-stage estimation procedure is required:
    1. Given current estimates of α and ϕ, an estimate of $\boldsymbol{\beta}$ is obtained as the solution to (1) on the previous slide.
    2. Given current estimate of $\boldsymbol{\beta}$ estimates of α and ϕ are obtained based on the residuals,

$$r_{ij} = Y_{ij} - \hat{\mu}_{ij}$$

Properties of GEE estimators

Assuming α and ϕ are consistent:

- $\hat{\boldsymbol{\beta}}$ is a consistent estimate of $\boldsymbol{\beta}$ (with high probability $\hat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}$ for large n).
- In large sample, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution.
- $Cov(\boldsymbol{\beta}) = \boldsymbol{F}^{-1}\boldsymbol{G}\boldsymbol{F}^{-1}$ where

$$
\begin{aligned}
F &= \sum_{i=1}^{n} \boldsymbol{D}_i^{-1}\boldsymbol{V}_i\boldsymbol{D}_i^{-1} \\
G &= \sum_{i=1}^{n} \boldsymbol{D}_i\boldsymbol{V}_i^{-1}Cov(\boldsymbol{Y}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i
\end{aligned}
$$

This is referred to as the “empirical” or “sandwich” variance estimator.

Summary of GEE estimators

- $\hat{\beta}$ is consistent even if the covariance of $\boldsymbol{Y}_i$ has been misspecified (robust).
- The variance of $\hat{\beta}$ can be estimated by $\boldsymbol{F}^{-1}$ or $\boldsymbol{F}^{-1}\boldsymbol{G}\boldsymbol{F}^{-1}$.
  - $\boldsymbol{F}^{-1}$ is the 'model-based' estimator.
  - $\boldsymbol{F}^{-1}\boldsymbol{G}\boldsymbol{F}^{-1}$ is the 'empirical' or 'sandwich' estimator.
- The standard errors of $\hat{\beta}$, as measured by $\boldsymbol{F}^{-1}\boldsymbol{G}\boldsymbol{F}^{-1}$ are asymptotically valid even when the correlation structure is incorrect.
  - Why model the correlation then?

F versus *G*

Both model based and sandwich based estimators are useful in different situations:

- **Sandwich based** is best to use when
  - sample size is relatively large (several hundred subjects or more)
  - when the assumed model for the covariances is questionable.
- **Model based** is best to use when
  - sample size is smaller
  - small number of clusters.
- Model based needs the correlation/covariance to be modeled correctly.

GEE limitations

- There is no likelihood function since the GEE does not specify completely the joint distribution; thus some do not consider it a model but just a method of estimation.
- Likelihood-based methods are NOT available for testing fit, comparing models, and conducting inferences about parameters.
- Empirical estimators more variable than the parametric ones.
- Sandwich based standard errors underestimate the true ones, unless sample size has several hundred subjects or more.
- More ideal for balanced data.