Statistical Methods for Correlated Data

Marginal Inference via Generalized Estimating Equations

Michele Guindani

Department of Biostatistics
UCLA

Learning goals:

- GEE as a *marginal* approach
- Estimating Function and Estimating Equations
- ML and Quasi-likelihood estimation in GLMs as examples of EEs
- Fundamental Theorem of Estimating Equations:

  Consistency and Asymptotic Normality of EE estimators

Generalized Estimating Equations (GEE)

- Over the past 20 years, the GEE approach has proven to be an exceedingly useful method for the analysis of longitudinal data, especially when the response is discrete Here, we start by considering a continuous response.

- GEE defines a class of regression models that are known as *marginal* models

- The term marginal in this context is used to emphasize that the model for the mean response at each occasion depends only on the covariates of interest, and does not incorporate dependence on random effects or previous responses.

- For estimation of the regression model parameters, marginal models do not require distributional assumptions for the vector of longitudinal responses (semi-parametric modeling)

- This is in contrast to LMMs, where the mean response is modeled not only as a function of covariates but is conditional also on random effects.

Generalized Estimating Equations (GEE)

- The GEE approach has its basis in one particular type of extension of generalized linear models to longitudinal, or more generally, cluster-correlated data.

- The defining feature of marginal models is a regression model relating the mean response at each occasion to the covariates.

- With a marginal model, the main focus is on making inferences about population means.

- Marginal models for longitudinal data separately model the mean response and the within-subject association among the repeated responses.

- In a marginal model, the goal is to make inferences about the mean response, whereas the within-individual association is regarded as a nuisance characteristic of the data that must be taken into account in order to make correct inferences about changes in the population mean response over time.

Estimating Equations

- The GEE approach can be considered a multivariate extension of quasi-likelihood estimation, motivated by dependent data situations.

Estimating Equations

- The GEE approach can be considered a multivariate extension of quasi-likelihood estimation, motivated by dependent data situations.
- Let's review the main theoretical ideas behind estimating equations (see Sec 2.3 in the textbook)
- Let $\boldsymbol{Y} = [Y_1, \dots, Y_n]$ represent n observations from an unknown distribution. For now, we assume that the Y_i's are independent.
- An estimating function is a function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{G}\left(\boldsymbol{\theta}, Y_i\right)$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$\mathrm{E}\left[\boldsymbol{G}_n(\boldsymbol{\theta})\right] = \mathbf{0}$$

for all $\boldsymbol{\theta}$.

Estimating Equations

- The corresponding estimating equation is

$$\boldsymbol{G}_n\left(\widehat{\boldsymbol{\theta}}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}_n, Y_i\right) = \boldsymbol{0}$$

- The estimator $\widehat{\boldsymbol{\theta}}_n$ that solves is often unavailable in closed form
- The estimating function is sum of random variables, which provides the opportunity to evaluate its asymptotic properties via a central limit theorem since the first two moments are often easy to calculate
- The art of constructing estimating functions is to make them dependent on distribution-free quantities, for example, the first two moments of the data
- Robustness of inference to misspecification of higher moments often follows.

- One can view the score functions in ML estimation as an estimating function
$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{S}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}\log p\left(Y_i|\theta\right)$$
since, under suitable regularity conditions, $\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathrm{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$ and so $\mathrm{E}\left[\boldsymbol{G}_n(\boldsymbol{\theta})\right] = \frac{1}{n}\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0}$ and the MLE satisfies $G_n\left(\widehat{\theta}_n\right) = 0$
- See Sec 2.4.1 in your textbook for details

Theorem of Estimating Equations

- Suppose that $\hat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}\left(\boldsymbol{\theta}, Y_i\right) = \mathbf{0}$$

  that is, $\boldsymbol{G}_n\left(\widehat{\boldsymbol{\theta}}_n\right) = \mathbf{0}$.

- Then, $\widehat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}$(consistency) and

$$\left[\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\left(\boldsymbol{A}_n^{\mathrm{T}}\right)^{-1}\right]^{-1/2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \rightarrow_d \mathrm{N}_p\left(\mathbf{0}, \mathrm{I}_p\right)$$

  with

$$\boldsymbol{A}_n = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{G}_n(\boldsymbol{\theta})\right]$$

$$\boldsymbol{B}_n = \mathrm{E}\left[\boldsymbol{G}_n(\boldsymbol{\theta})\boldsymbol{G}_n(\boldsymbol{\theta})^{\mathrm{T}}\right] = \mathrm{var}\left[\boldsymbol{G}_n(\boldsymbol{\theta})\right]$$

Remarks

- Importance of the unbiasedness requirement: If the expectation of the estimating function is not zero, then an inconsistent estimator of $\boldsymbol{\beta}$ results.
- Estimators derived from an estimating function are invariant in the sense that if we are interested in a function, $\phi = g(\boldsymbol{\theta})$, then the estimator is $\widehat{\phi}_n = g\left(\widehat{\boldsymbol{\theta}}_n\right)$.

Remarks

- The variance of $\widehat{\boldsymbol{\theta}}_n$ is $\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\left(\boldsymbol{A}_n^{\mathrm{T}}\right)^{-1}$ where $\boldsymbol{B}_n$ is the covariance of the estimating function sandwiched by $\boldsymbol{A}_n$, which can be seen as the expectation of the inverse of the Jacobian matrix of the transformation from the estimating function to the parameter
- With independent and identically distributed observations $\boldsymbol{A}_n = n\boldsymbol{A}$ and $\boldsymbol{B}_n = n\boldsymbol{B}$ so that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \rightarrow_d \mathbf{N}_p\left[\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{A}^{\mathrm{T}}\right)^{-1}\right]$
- In practice, $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\theta})$ are replaced by $\boldsymbol{A}_n\left(\widehat{\boldsymbol{\theta}}_n\right)$ and $\boldsymbol{B}_n\left(\widehat{\boldsymbol{\theta}}_n\right)$ respectively, with asymptotic normality continuing to hold due to Slutsky's theorem.

Remarks

- More specifically, one can evaluate $\boldsymbol{A}$ and $\boldsymbol{B}$ empirically via $\widehat{\boldsymbol{A}}_n = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}, Y_i\right)$ and $\widehat{\boldsymbol{B}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}, Y_i\right) \boldsymbol{G}\left(\widehat{\boldsymbol{\theta}}, Y_i\right)^{\mathrm{T}}$
- By the weak law of large numbers, $\widehat{\boldsymbol{A}}_n \to_p \boldsymbol{A}$ and $\widehat{\boldsymbol{B}}_n \to_p \boldsymbol{B}$, and

$$\mathrm{var}\left(\widehat{\boldsymbol{\theta}}_n\right) = \frac{\widehat{\boldsymbol{A}}^{-1}\,\widehat{\boldsymbol{B}}\,\left(\widehat{\boldsymbol{A}}^{\mathrm{T}}\right)^{-1}}{n}$$

  is a consistent estimator of the variance.
- Sandwich estimation provides a consistent estimator of the variance in very broad situations.

Learning goals:

- Quasi-likelihood estimation in GLMs as examples of EEs
- Idea of assuming a working variance-covariance matrix
- Gauss-Markov Theorem for Dependent Data
  - What does it mean in terms of variance
- Generalized Estimating equaitons

Recap of Quasi-likelihood estimation in GLMs

- We assume N independent observations of a scalar response variable, Y_i. Associated with the response, Y_i, there are p covariates, $X_{i1}, \ldots, X_{ip}$
- The primary interest is relating the mean of $Y_i, \mu_i = E\left(Y_i | X_{i1}, \ldots, X_{ip}\right),$ to the covariates.
- So, we assume a transformation of the mean response, μ_i, is linearly related to the covariates via an appropriate link function, $h^{-1}\left(\mu_i\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$ where the link function $h^{-1}(\cdot)$ is a known function, such as $\log\left(\mu_i\right)$
- If Y is from an exponential family, $\operatorname{Var}\left(Y_i\right) = \phi v\left(\mu_i\right)$ where the scale parameter $\phi > 0.$ The variance function, $v\left(\mu_i\right),$ describes how the variance of the response is functionally related to the mean of Y_i.

Quasi-likelihood estimation

- The likelihood equations for generalized linear models depend only on the mean and variance of the response (and the link function).

⇨ Wedderburn (1974) suggested using them as "estimating equations" for any choice of link or variance function, even when the particular choice of variance function does not correspond to an exponential distribution, by solving the equations

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)' W_i^{-1} \{ Y_i - \mu_i(\boldsymbol{\beta}) \} = \mathbf{0}$$

- Wedderburn (1974) showed that for any choice of weights, W_i, the quasi-likelihood estimator of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}$, is consistent and asymptotically normal.
- If one chose weights $W_i = \mathrm{Var}(Y_i)$ then the resulting estimator would be the smallest variance among all estimators in this class.
- In GLMs, it is assumed that $W_i = \mathrm{Var}(Y_i) = \phi\, v(\mu_i)$, and this assumption is sufficient to characterize the distribution within the exponential family.

Quasi-likelihood estimation

- In summary, Wedderburn (1974) proposed estimators of $\boldsymbol{\beta}$ that do not require distributional assumptions on the response.
- This allows more flexible models for variability, e.g., incorporating overdispersion.
- Example: Overdispersion violates the mean-variance relation induced from a proper probability model, which prohibits investigators from using a specific parametric distribution for the data. Overdispersion may emerge from different data collection procedures, one of which is that the response variable is recorded as an aggregation of dependent variables.
- Let us consider a Poisson log-linear regression model for count data. In a standard GLM analysis, count responses are typically assumed to follow a Poisson distribution with mean μ,

$$\log(\mu) = \mathbf{x}^T \boldsymbol{\beta}$$

Quasi-likelihood estimation

- The assumption of a Poisson distribution for the data implies the mean-variance relation of the following form

  $$\text{Var}(Y) = \mu$$

  since the dispersion parameter $\phi = 1$ and the unit variance function is $V(\mu) = \mu$.

Quasi-likelihood estimation

- The assumption of a Poisson distribution for the data implies the mean-variance relation of the following form
$$\mathrm{Var}(Y) = \mu$$
since the dispersion parameter $\phi = 1$ and the unit variance function is $V(\mu) = \mu$.
- One way to deal with over-dispersed count data is to introduce a dispersion parameter σ^2 that inflates the Poisson variance as given by choosing
$$W = \sigma^2\mu, \quad \sigma^2 > 1$$
Obviously, this response variable Y satisfying such a new mean-variance relation is no longer Poisson distributed. The choice W is assumed as describing the variance, but it may not be correspondent to the truth
- In the cases discussed above, the maximum likelihood estimation approach may not be applicable due to the unavailability of full density functions. However, it is possible to build a quasi-score function and obtain a quasi-likelihood estimator based on the assumptions for the first two moments.

Quasi-likelihood estimation

- Quasi-likelihood estimation only requires correct specification of the model for the mean to yield consistent and asymptotically normal estimators of $\boldsymbol{\beta}$, even when the variance of the response has been misspecified, that is, $W_i \neq \text{Var}(Y_i)$.

- More specifically, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$, satisfies

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(\mathbf{0}, C_\beta)$$

where

$$C_\beta = \lim_{N\to\infty} I_0^{-1} I_1 I_0^{-1}$$

with

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' W_i^{-1}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)$$

and

$$I_1 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' W_i^{-1}\,\text{Var}(Y_i)\,W_i^{-1}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)$$

Quasi-likelihood estimation

- Consistent estimators of the asymptotic covariance of the estimated regression parameters can be obtained using the empirical estimator of C_β first suggested by Cox (1961) and later by Huber (1982)
- The empirical variance estimator is obtained by evaluating $\partial \mu_i / \partial \boldsymbol{\beta}$ at $\widehat{\boldsymbol{\beta}}$ and substituting $\left(Y_i - \widehat{\mu}_i\right)^2$ for $\operatorname{Var}\left(Y_i\right)$.
- This is widely known as the sandwich variance estimator.
- It can be shown that the same asymptotic distribution holds when V_i is estimated rather than known, with V_i replaced by estimated weights, say $\widehat{V}_i$.

Generalized Estimating Equations, Sec 8.7 Wakefield

- Suppose we assume

$$\mathrm{E}\left[\boldsymbol{Y}_i|\boldsymbol{\beta}\right] = \boldsymbol{x}_i\boldsymbol{\beta}$$

and consider the $n_i \times n_i$ working variance-covariance matrix:

$$\mathrm{var}\left(\boldsymbol{Y}_i|\boldsymbol{\beta},\boldsymbol{\alpha}\right) = \boldsymbol{W}_i$$

To motivate GEE we begin by assuming that $\boldsymbol{W}_i$ is known. In this case the GLS estimator minimizes

$$\sum_{i=1}^{m}\left(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}\right)^{\mathrm{T}}\boldsymbol{W}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}\right)$$

and is given by the solution to the estimating equation

$$\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}\right) = 0$$

which is $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{Y}_i$

Generalized Estimating Equations

- We have

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{E}\left[\boldsymbol{Y}_i \right] = \boldsymbol{\beta}$$

  so long as the mean is correctly specified.

- If the information about $\boldsymbol{\beta}$ grows with increasing m, then $\boldsymbol{\beta}$ is consistent.
- The variance, $\mathrm{var}(\widehat{\boldsymbol{\beta}})$, is given by

$$\left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \, \mathrm{var}\left(\boldsymbol{Y}_i \right) \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right) \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1}$$

  Notice that if the assumed variance-covariance matrix is correct, i.e. $\mathrm{var}\left(\boldsymbol{Y}_i \right) = \boldsymbol{W}_i$, then

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1}$$

  If m is large then a multivariate central limit theorem shows that $\widehat{\boldsymbol{\beta}}$ is asymptotically normal.

Gauss-Markov Theorem for Dependent Data, Prob. 8.1

Suppose $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{x}\boldsymbol{\beta}$ and $\mathrm{var}(\boldsymbol{Y}) = \boldsymbol{V}$, with $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^{\mathrm{T}}, \ldots, \boldsymbol{Y}_m^{\mathrm{T}}\right]^{\mathrm{T}}$ and where $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]^{\mathrm{T}}$ and $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^{\mathrm{T}}$ is $N \times (k+1)$ with $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}], \boldsymbol{x}_{ij} = [1, x_{ij1}, \ldots, x_{ijk}]^{\mathrm{T}}, N = \sum_i n_i$ and $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector of regression coefficients.
Consider linear estimators of the form

$$\tilde{\boldsymbol{\beta}}_{\mathrm{w}} = \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{Y}$$

where $\boldsymbol{W}$ is symmetric and positive definite. Then,

(a) $\mathrm{E}\left[\widetilde{\boldsymbol{\beta}}_{\mathrm{w}}\right] = \boldsymbol{\beta}$

(b) $\mathrm{var}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{v}}\right) \leq \mathrm{var}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{w}}\right)$

Generalized Estimating Equations

Generalized Estimating Equations

- We consider the multivariate extension of the quasi-likelihood approach to the setting of marginal models for longitudinal responses
- We now suppose that $\mathrm{var}\left(\boldsymbol{Y}_i\right) = \boldsymbol{W}_i(\boldsymbol{\alpha})$ where α are unknown parameters in the variance-covariance model.
- Typically, we assume that

  $$\boldsymbol{W}_i(\boldsymbol{\alpha}) = \alpha_1 \boldsymbol{R}_i\left(\boldsymbol{\alpha}_2\right)$$

  where $\alpha_1 = \mathrm{var}\left(Y_{ij}\right)$ denotes the variance of the response for all i and j (typically, we assume some form of the variances, e.g., that they are a function of the mean, as well as homogeneity or heteroskedasticity), and $\boldsymbol{R}_i\left(\boldsymbol{\alpha}_2\right)$ is a working correlation matrix that depends on parameters α_2
- GEE uses a working second moment assumption; "working" refers to the choice of a variance model that may not necessarily correspond to exactly the form we believe to be true but rather to be a choice that is statistically convenient

GEE

- Several choices for $\boldsymbol{R}_i$ (independence, exchangeable and AR(1) models)
- If $R_i(\boldsymbol{\alpha}) = I_{n_i}$, then the GEE reduces to the QL-estimating equations for a GLM that assumes that the repeated measures are independent
- More generally, one can specify

$$\boldsymbol{W}_i = \phi A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2}$$

where $A_i = \text{diag}\left\{v\left(\mu_{ij}\right)\right\}$ is a diagonal matrix with diagonal elements $v\left(\mu_{ij}\right)$ which are specified entirely by the marginal means (i.e., by $\boldsymbol{\beta}$).

GEE

- For known $\alpha, \hat{\boldsymbol{\beta}}$ is the root of the estimating equation

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}) \left(\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta}\right) = \mathbf{0}$$

  which is reminiscent of the GLS minimization.

- When α is unknown, we require a "well-behaved" estimator $\widehat{\alpha}$, so to have a stable weighting matrix $W(\widehat{\boldsymbol{\alpha}})$
- The sandwich variance estimator is

$$\widehat{\operatorname{var}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \operatorname{var}\left(\boldsymbol{Y}_i\right) \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right) \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right)^{-1}$$

  where $\widehat{\boldsymbol{W}}_i = \boldsymbol{W}_i(\widehat{\boldsymbol{\alpha}})$ and $\operatorname{var}\left(\boldsymbol{Y}_i\right)$ is estimated by the variance-covariance matrix of the residuals:

$$\left(\boldsymbol{Y}_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}\right) \left(\boldsymbol{Y}_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}$$

GEE

- The sandwich variance estimator is a consistent estimator of $\text{var}(\widehat{\boldsymbol{\beta}})$, so long as we have independence between units, that is, $\text{cov}\left(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}\right) = 0$ for $i \neq i'$

- For inference, we use the asymptotic distribution

$$\sqrt{m}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathbf{N}_{k+1}(\mathbf{0}, \widehat{\text{var}}(\widehat{\boldsymbol{\beta}}))$$

  where the asymptotic is in the number of units, m.

- The sandwich estimator is also called “robust”, because it provides valid standard errors when the assumed model for the covariance is not correct, for large sample sizes. Another term is “empirical”.

- Heuristically, using results from the method of moments, $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ regardless of whether W_i is the true covariance matrix of $\boldsymbol{Y}_i$ because

$$\begin{aligned} E\left\{u_\beta(\boldsymbol{\beta})\right\} &= E\left[\sum_{i=1}^{N} D_i' \boldsymbol{W}_i^{-1}\left\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right\}\right] \\ &= \sum_{i=1}^{N} D_i' \boldsymbol{W}_i^{-1} E\left\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right\} = \mathbf{0} \end{aligned}$$

  with $D_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and we are solving $u_\beta(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ for $\widehat{\boldsymbol{\beta}}$.

- While the variance V_i does not need have to be correctly specified in order to obtain a consistent estimator of $\boldsymbol{\beta}$, the result is of course conditional on the mean response to be correctly specified.

Apparent coolness (or paradox) of GEE

- It looks like we can obtain a valid estimate of β and its sampling variability, even if we have not modeled the within-subject correlation correctly (in LDA!)

- However:
  - better estimates of $\boldsymbol{V}_i$

  ⇨ allow greater precision in small samples

  - If the number of subjects is small (m) or the number of repeated measures (n) is small, or the design severely unbalanced

  ⇨ the sandwich estimator is biased (for ↓ m, the variance of $\hat{\boldsymbol{\beta}}$ is underestimated)

Interpretation of the parameters

- The GEE approach is constructed to carry out marginal inference, and so we cannot perform individual-level inference.
- For a linear model, marginalizing a LMM produces a marginal model identical to that used in a GEE approach
- The interpretation of the fixed effects parameters as "population-wide parameters" is valid in both paradigms
- When nonlinear models are considered there is no equivalence and the differences between the conditional and marginal approaches to inference becomes more pronounced.

- In the most general case of working variance model specification, we may allow the working variance model to depend on $\boldsymbol{\beta}$ also, so that we have $\boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to allow mean-variance relationships.
- For example, in a longitudinal setting, the variance may depend on the square of the marginal mean μ_{ij} with an autoregressive covariance model:

$$\begin{aligned} \operatorname{var}(Y_{ij}) &= \alpha_1 \mu_{ij}^2 \\ \operatorname{cov}(Y_{ij}, Y_{ik}) &= \alpha_1 \alpha_2^{|t_{ij}-t_{ik}|} \mu_{ij}\mu_{ik} \\ \operatorname{cov}(Y_{ij}, Y_{i'k}) &= 0, \quad i \neq i' \end{aligned}$$

  with $j = 1, \ldots, n_i, k, k' = 1, \ldots, n_{i'}$ and where t_{ij} is the time associated with response Y_{ij}.
- α_1 is the component of the variance that does not depend on the mean (and is assumed constant across time and across individuals)
- α_2 is the correlation between responses on the same individual which are one unit of time apart
- $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]$

- The roots of the estimating equation

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta})\,(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = \mathbf{0}$$

are not available in closed form when $\boldsymbol{\beta}$ appears in $\boldsymbol{W}$.

- The roots of the estimating equation

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \left(\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta}\right) = \mathbf{0}$$

are not available in closed form when $\boldsymbol{\beta}$ appears in $\boldsymbol{W}$.

- We can write the $(k+1) \times 1$ estimating function in a variety of forms,

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})$$

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta}\right)$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \boldsymbol{x}_{ij} W_i^{jk} \left(Y_{ik} - \boldsymbol{x}_{ik} \boldsymbol{\beta}\right)$$

where W_i^{ij} denotes entry (i, j) of $\boldsymbol{W}_i^{-1}$. We will often use the middle form, since this emphasizes that the basic unit of replication (upon which the asymptotic properties depend) is indexed by i.

Estimation of Variance Parameters

- A second set of (moment) estimating equations can be used to estimate $\boldsymbol{\alpha}$.

Estimation of Variance Parameters

- A second set of (moment) estimating equations can be used to estimate $\boldsymbol{\alpha}$.
- Let
$$U_{ist}(\boldsymbol{\beta}) = \frac{\left(Y_{is} - \mu_{is}\right)\left(Y_{it} - \mu_{it}\right)}{\phi\left\{v\left(\mu_{is}\right)v\left(\mu_{it}\right)\right\}^{1/2}}$$
which is an unbiased estimate of the correlation ρ_{ist} of the residuals between occasions s and t for individual i.
- The U_{ist} can be grouped together to form the $n_i\left(n_i - 1\right)/2 \times 1$ vector $\boldsymbol{U}_i(\boldsymbol{\beta}) = \left(U_{i12}, U_{i13}, \ldots, U_{in_{i-1}n_i}\right)'$
- Also, let $\boldsymbol{\rho}_i(\boldsymbol{\alpha}) = E\left(\boldsymbol{U}_i; \boldsymbol{\alpha}\right) = \left(\rho_{i12}, \rho_{i13}, \ldots, \rho_{in_{i-1}n_i}\right)'$.

Estimation of Variance Parameters

- In addition to the EEs for $\boldsymbol{\beta}$, $u_\beta(\boldsymbol{\beta}) = \mathbf{0}$, we can then consider the following EEs for $\boldsymbol{\alpha}$,

$$u_\alpha(\boldsymbol{\alpha}) = \sum_{i=1}^{N} E_i' \tilde{W}_i^{-1} \left\{ \boldsymbol{U}_i(\boldsymbol{\beta}) - \boldsymbol{\rho}_i(\boldsymbol{\alpha}) \right\} = \mathbf{0}$$

  where $\tilde{W}_i \approx \operatorname{Cov}(\boldsymbol{U}_i)$ and $E_i = \partial \boldsymbol{\rho}_i(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$.

- The working covariance matrix for $\boldsymbol{U}_i$ is typically specified as $\operatorname{diag}\left\{\operatorname{Var}(U_{ist})\right\}$.

- In their original paper on GEE, Liang and Zeger (1986) let W_i be the $(n_i \times n_i - 1)/2 \times (n_i \times n_i - 1)/2$ identity matrix, whereas Prentice suggested letting W_i be a diagonal matrix with the approximate variances of U_{ist} along the diagonal.

GEE algorithm

- Obtaining GEE estimates requires an iterative algorithm. The structure of the GEE suggests the use of a specific iterative scheme:
- Estimate $\boldsymbol{\beta}$ (given the current estimate of $\boldsymbol{\alpha}$)
- Estimate α (given the current estimate of $\boldsymbol{\beta}$)
- The solution $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$ of the two system of equations can be obtained by a Fisher scoring algorithm.

GEE algorithm

- Given a starting value for $\boldsymbol{\beta}$, say under the naive assumption of independence, the solution $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$ can be obtained by iterating between

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} + \left[\sum_{i=1}^{N} D_i^{(m)'} \left\{W_i^{(m)}\right\}^{-1} D_i^{(m)}\right]^{-1} \sum_{i=1}^{N} D_i^{(m)'} \left\{W_i^{(m)}\right\}^{-1} \left\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i\left(\widehat{\boldsymbol{\beta}}^{(m)}\right)\right\}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^{(m+1)} =& \widehat{\boldsymbol{\alpha}}^{(m)} + \left[\sum_{i=1}^{N} E_i^{(m)'} \left\{\tilde{W}_i^{(m)}\right\}^{-1} E_i^{(m)}\right]^{-1} \\ & \times \sum_{i=1}^{N} E_i^{(m)'} \left\{\tilde{W}_i^{(m)}\right\}^{-1} \left[\boldsymbol{U}_i\left\{\widehat{\boldsymbol{\beta}}^{(m)}\right\} - \boldsymbol{\rho}_i\left\{\widehat{\boldsymbol{\alpha}}^{(m)}\right\}\right]\end{aligned}$$

until $\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)}$ and $\widehat{\boldsymbol{\alpha}}^{(m+1)} = \widehat{\boldsymbol{\alpha}}^{(m)}$, where $D_i^{(m)} = D_i\left\{\widehat{\boldsymbol{\beta}}^{(m)}\right\}, W_i^{(m)} = W_i\left\{\widehat{\boldsymbol{\beta}}^{(m)}, \widehat{\boldsymbol{\alpha}}^{(m)}\right\}$, $E_i^{(m)} = E_i\left\{\widehat{\boldsymbol{\beta}}^{(m)}, \widehat{\boldsymbol{\alpha}}^{(m)}\right\}$, and $\tilde{W}_i^{(m)} = \tilde{W}_i\left\{\widehat{\boldsymbol{\beta}}^{(m)}, \widehat{\boldsymbol{\alpha}}^{(m)}\right\}$

Estimator of $\boldsymbol{\alpha}$ in Liang and Zeger (1986)

- The estimator of $\boldsymbol{\alpha}$ originally proposed by Liang and Zeger (1986) is non-iterative.
- For example, suppose an "exchangeable" correlation pattern is assumed, in which $\rho_{ist} = \alpha$ for all $s < t$.
- Then, Liang and Zeger (1986) proposed estimating α, given the current estimate of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}$, by

$$\widehat{\alpha} = \frac{1}{N^*} \sum_{i=1}^{N} \sum_{s<t}^{n_i} \frac{\left(Y_{is} - \widehat{\boldsymbol{\mu}}_{is}\right)\left(Y_{it} - \widehat{\boldsymbol{\mu}}_{it}\right)}{\widehat{\phi}\left\{v\left(\widehat{\mu}_{is}\right) v\left(\widehat{\mu}_{it}\right)\right\}^{1/2}}, \quad \text{where } N^* = \sum_{i=1}^{N} \frac{n_i\left(n_i - 1\right)}{2}$$

- Alternatively, various degree-of-freedom corrections to account for the estimation of $\boldsymbol{\beta}$ have been suggested for the denominator N^*, e.g., $N^* - p$.

Inferences on $\boldsymbol{\beta}$

- For making inferences about $\boldsymbol{\beta}$, since the GEE approach is not likelihood-based, likelihood ratio tests are not available for hypotheses testing;
- This also has ramifications for inferences when there are missing data, a topic that will be discussed later
- Inferences typically rely on Wald test statistics based on quadratic forms.
- For a null hypothesis $H_0 : L\boldsymbol{\beta} = 0$, versus the alternative $H_A : L\boldsymbol{\beta} \neq 0$, for an $r \times p$ matrix L of full rank $r \leq p$, the Wald test statistic is

$$X^2 = (L\widehat{\boldsymbol{\beta}})' \left\{ L\widehat{\text{Cov}}(\widehat{\boldsymbol{\beta}})L' \right\}^{-1} L\widehat{\boldsymbol{\beta}} \sim \chi^2_r$$

Implicit assumption in GEE

- The key property for (asymptotically) unbiased estimators using GEE is

$$E\left[D_i' V_i^{-1}\left\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right\}\right] = E_{x_i}\left[D_i' V_i^{-1} E_{y_i|x_i}\left\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\right\}\right] = \mathbf{0}$$

  where $E_{x_i}(\cdot)$ denotes expectation with respect to the marginal distribution of X_i and $E_{y_i|x_i}(\cdot)$ denotes expectation with respect to the conditional distribution of $\boldsymbol{Y}_i$ given X_i.

- Note that the j th element of the vector $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ is $\mu_{ij} = E\left(Y_{ij}|\boldsymbol{X}_{ij}\right)$, but the elements of $E_{y_i|x_i}\left(\boldsymbol{Y}_i\right)$ are $E\left(Y_{ij}|\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}\right)$.

- Thus, an implicit assumption is

$$E\left(Y_{ij}|X_i\right) = E\left(Y_{ij}|\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}\right) = E\left(Y_{ij}|\boldsymbol{X}_{ij}\right)$$

- With time-stationary covariates, this assumption poses no difficulties; it necessarily holds because $\boldsymbol{X}_{ij} = \boldsymbol{X}_{ik}$ for all occasions $k \neq j$

- With time-stationary covariates, this assumption poses no difficulties; it necessarily holds because $\boldsymbol{X}_{ij} = \boldsymbol{X}_{ik}$ for all occasions $k \neq j$
- Also, with time-varying covariates that are fixed by design of the study (e.g., time since baseline, treatment group indicator in a crossover trial), the assumption also holds because values of the covariates at any occasion are determined *a* priori by study design and in a manner completely unrelated to the longitudinal response.

- With time-stationary covariates, this assumption poses no difficulties; it necessarily holds because $\boldsymbol{X}_{ij} = \boldsymbol{X}_{ik}$ for all occasions $k \neq j$
- Also, with time-varying covariates that are fixed by design of the study (e.g., time since baseline, treatment group indicator in a crossover trial), the assumption also holds because values of the covariates at any occasion are determined *a* priori by study design and in a manner completely unrelated to the longitudinal response.
- However, when a time-varying covariate varies randomly over time the assumption may not hold, e.g. the current value of the response, say Y_{ij} the current covariates $\boldsymbol{X}_{ij}$, predicts the subsequent value of $\boldsymbol{X}_{i,j+1}$
- Pepe and Anderson (1994) recommend using GEE with a "working independence" assumption. Under a "working independence" assumption, the weight matrix is diagonal and the corresponding estimating equations simplify and are unbiased regardless of whether or not the assumption is satisfied.

Efficiency

- Since all GEE estimators of $\boldsymbol{\beta}$ are consistent and asymptotically normal, it is of interest to consider their efficiency under various working covariance assumptions.
- Claim: Setting $R_i = I$, the leads to an estimator with nearly the same efficiency as the true variance covariance matrix
- For a balanced longitudinal design, with no missing data, there are cases where this claim has some justification.
- However, for the case of a within-subject effect when the covariate design on time is not the same for all subjects (i.e., $X_{ijk} \neq X_{ij'k}$ for some occasions $j \neq j'$ and $X_{ijk} \neq X_{i'jk}$ for some subjects $i \neq i'$) , there can be a very discernible loss of efficiency under the non-optimal "working independence" assumption for the covariance, especially when the true correlations are moderately large (see, for example, Lipsitz et al.,1994)
- If V_i is correctly specified, a consistent estimator for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $\left(\sum_{i=1}^{N} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1}$, which is often referred as the model-based variance estimator.

- With small sample sizes, the traditional GEE with the classic "sandwich" variance estimator does not perform satisfactorily and considerable downward bias is exhibited in turn leading to inflated type I errors and lower coverage rates of the resulting confidence intervals.

  Intuition behind this point: The $Var(Y_i)$ in the formula for the $Cov(\hat{\beta})$ is estimated using the sample variance, a point we have made before. With small samples, the fitted values of the means $\hat{\boldsymbol{\mu}}_i$ tend to be closer to the observed values than the true values $\boldsymbol{Y}_i$; hence, the sample variances (computed through the residuals) tend to be biased, and in particular to *underestimate* the true variability of $Var(\hat{\beta})$.

- Underestimation of the variability of the coefficients results in tests that are too liberal and tend to reject more often than not

- For a more technical discussion of these points (and some interesintg Theorems), see: Kauermann and Carroll, A Note on the Efficiency of Sandwich Covariance Matrix Estimation, JASA, 2001

- Several remedy strategies on modifications of variance estimators have been proposed to improve the finite small-sample performance

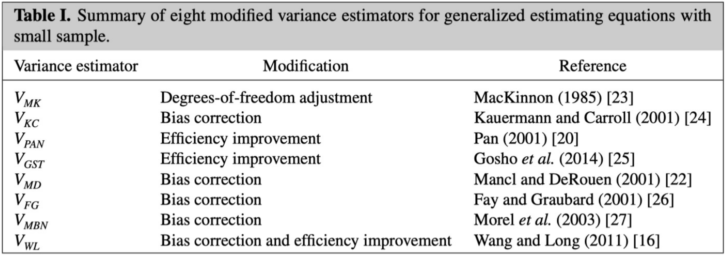

Table I. Summary of eight modified variance estimators for generalized estimating equations with small sample.

Variance estimator	Modification	Reference
V_{MK}	Degrees-of-freedom adjustment	MacKinnon (1985) [23]
V_{KC}	Bias correction	Kauermann and Carroll (2001) [24]
V_{PAN}	Efficiency improvement	Pan (2001) [20]
V_{GST}	Efficiency improvement	Gosho *et al.* (2014) [25]
V_{MD}	Bias correction	Mancl and DeRouen (2001) [22]
V_{FG}	Bias correction	Fay and Graubard (2001) [26]
V_{MBN}	Bias correction	Morel *et al.* (2003) [27]
V_{WL}	Bias correction and efficiency improvement	Wang and Long (2011) [16]

Wang et al (2016) *Statist. Med*. 2016,3 5, 1706-1721

- Here, we just present a simple modification that resembles the one adopted in linear regression: V_{MK} the *degrees-of-freedom corrected* "sandwich" variance estimator proposed by MacKinnon (1985)
- This estimator incorporates the simplest adjustment by adopting an adjustment factor of $\frac{m}{m-p}$, Which is shown by

  $$V_{MK} = \frac{m}{m-p} V_{LZ}$$

  where V_{LZ} denotes the original sandwich estimator (LZ stands for Liang and Zeger, 1986)
- When $N \to \infty, V_{MK} \to_p V_{LZ}$ and since $V_{sandwich}$ is consistent, so is V_{MK}.
- V_{MK} allows to adjust the sandwich estimator for the bias, but this is obtained at the expense of the efficiency of the estimator itself, since

  $$Cov(vec(V_{MK})) > Cov(vec(V_{LZ}))$$