# Marginal Models and Generalized Estimating Equations

## 1 Introduction

Here the focus is to estimate the population average effects, and it only relies on the marginal properties (mean and variance) of the data. We will consider subject-specific mean and variances, however, the free parameters will be shared by all. The focus is still on quantifying the effects of some predictors on the outcome, but marginally, i.e. in terms of population average effects.

Generalized Estimating Equations (GEE) is a statistical method used for analyzing longitudinal or clustered data for marginal models. It extends the generalized linear model (GLM) to account for correlated observations within clusters or repeated measures or longitudinal measurements. The formulation of GEE is motivated by the generalized least squares (GLS) loss, which gives us the estimating equation for correlated continuous data.

## 2 Generalized least square/MLE for correlated continuous data

For a continuous valued data $\mathbf{y}_i$ with $n_i$ observations, we consider the simple linear model, $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ where $\boldsymbol{\epsilon}_i \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_i)$ with $i = 1, \ldots, N$. In any case, given the predictors and observation time points, the other parameters in $\boldsymbol{\Sigma}_i$ need to be shared by all the subjects. The log-likelihood is given by,

$$-\frac{N}{2}\log\det(\boldsymbol{\Sigma}_i) - \frac{1}{2}\sum_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$$

When $\boldsymbol{\Sigma}_i$'s are given, we need to minimize $\sum_{i=1}^N(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ to estimate $\boldsymbol{\beta}$. Then the estimating equation for $\boldsymbol{\beta}$ is given by $\sum_{i=1}^N \mathbf{X}_i^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) = \mathbf{0}$.

Now if we replace $\mathbf{X}_i\boldsymbol{\beta}$ in GLS loss by $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, some pre-specified function of $\mathbf{x}_i^T\boldsymbol{\beta}$ as mean. The minimization problem becomes $\sum_{i=1}^N(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}))^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}))$. Then the estimating equation becomes, $\sum_{i=1}^N \left(\frac{\partial\boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$.

[**Vector-vector derivative:** The derivative of a vector function (a vector whose components are functions)

$\mu_i = \begin{bmatrix}\mu_{i,1} & \mu_{i,2} & \cdots & \mu_{i,n_i}\end{bmatrix}^\mathsf{T}$, with respect to an input vector,

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \beta_2 & \cdots & \beta_p\end{bmatrix}^\mathsf{T}, \; \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial \mu_{i,1}}{\partial \beta_1} & \frac{\partial \mu_{i,1}}{\partial \beta_2} & \cdots & \frac{\partial \mu_{i,1}}{\partial \beta_p} \\ \frac{\partial \mu_{i,2}}{\partial \beta_1} & \frac{\partial \mu_{i,2}}{\partial \beta_2} & \cdots & \frac{\partial \mu_{i,2}}{\partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mu_{i,n_i}}{\partial \beta_1} & \frac{\partial \mu_{i,n_i}}{\partial \beta_2} & \cdots & \frac{\partial \mu_{i,n_i}}{\partial \beta_p} \end{bmatrix}$$

# 3 Formulation

## 3.1 Mean part

Consider a dataset with $N$ subjects/clusters (e.g., households, schools) and $n_i$ observations within each subject/cluster $i$. Let $y_{i,j}$ denote the response variable for the $j$th observation in subject/cluster $i$. The GEE framework assumes the following form for the mean of $y_{i,j}$:

$$E(y_{i,j}) = \mu_{i,j} = g^{-1}(\mathbf{x}_{i,j}^T \boldsymbol{\beta}), \text{ for } j = 1, \ldots, n_i; \quad i = 1, \ldots, N,$$

where:

- $g(\cdot)$ is a link function.
- $\mathbf{x}_{i,j}^T$ is a vector of covariates for the $j$th observation in cluster $i$.
- $\boldsymbol{\beta}$ is the vector of regression coefficients.
- It can handle inherently unbalanced designs and missing data with ease (albeit making strong assumptions about missingness).

GEE focuses on estimating the population-average or marginal effects of covariates on the response, rather than the subject-specific effects.

**Example:** For binary data, we set $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \frac{1}{1+e^{-\mathbf{X}_i \boldsymbol{\beta}}}$, and for count data, the we set $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = e^{\mathbf{X}_i \boldsymbol{\beta}}$, where $\mathbf{X}_i$ is $n_i \times p$ dimensional design matrix for *i-th* subject.

## 3.2 Variance part

Here we set, $\boldsymbol{\Sigma}_i$ which is specified in two parts.

1. The variance of each $y_{i,j}$, given the covariates, depends on the mean according to $V(y_{i,j} \mid x_{i,j}) = \phi v(\mu_{i,j})$, where $v(\cdot)$ is a known "variance function" (i.e., a known function of the mean, $\mu_{i,j}$) and $\phi$ is a scale parameter that may be known or may need to be estimated. The choice of $v$ is based on GLM. For example, when the response is continuous, $\phi$ is a scale parameter that needs to be estimated. In contrast, with a binary response, $\phi$ is known and fixed at 1. For count data, $\phi$ is often estimated from the data at hand to allow for overdispersion relative to Poisson variability. **Example: Since $y_{i,j} \sim$ Bernoulli$(p_{i,j})$, then $\mu_{i,j} = p_{i,j}$ and $V(y_{i,j}) = p_{i,j}(1 - p_{i,j}) = \mu_{i,j}(1 - \mu_{i,j})$. Hence for binary data, this is the choice of $v(x) = x(1 - x)$. For Poisson, $V(y_{i,j}) = \mu_{i,j}$ and thus choice of $v$ is $v(x) = x$ for count data.**

2. The pairwise (or two-way) within-subject association among the vector of repeated responses, given the covariates, is assumed to be a function of the means, $\mu_{i,j}$, and an additional set of within-subject association parameters, $\boldsymbol{\alpha}$. For example, when the vector of parameters a represents the pairwise correlations among the responses, the covariances among the responses depend on $\mu_{i,j}$, $\phi$, and $\boldsymbol{\alpha}$. That is, given a model for the pairwise correlations, the corresponding covariance matrix can be constructed as the product of standard deviations and correlations $\boldsymbol{\Sigma}_i = \mathbf{A}_i^{1/2} Corr(\mathbf{y}_i) \mathbf{A}_i^{1/2}$, where $\mathbf{A}_i$ is a diagonal matrix with $V(y_{i,j} \mid \mathbf{x}_{i,j}) = \phi v(\mu_{i,j})$ along the diagonal (and $\mathbf{A}_i^{1/2}$ is a diagonal matrix with the standard deviations, $\sqrt{\phi v(\mu_{i,j})}$, along the diagonal), and $Corr(\mathbf{y}_i)$ is the correlation matrix (here a function of $\boldsymbol{\alpha}$). In the parlance of the GEE approach, $\boldsymbol{\Sigma}_i$ is known as a "working" covariance matrix to distinguish it from the true underlying covariance among the $\mathbf{y}_i$. That is, the term "working" acknowledges our uncertainty about the assumed model for the variances and within-subject associations; unless they have been correctly modeled, our model for the covariance matrix may not be correct.

## 4 Assumptions

The key assumptions of GEE include:

1. **Marginal Model Specification:** The mean structure is correctly specified.

2. **Independence:** Observations across subjects/clusters are uncorrelated.

3. **Correct Specification of Correlation Structure:** The correlation structure within subjects/clusters is correctly specified. (However this can be relaxed through sandwich estimator. The working correlation/covariance matrix should be "reasonably close" to the population structure.)

4. (**Sample size:** It should be reasonably large for asymptotic inference.)

## 5 General estimation steps based on [2]

Since both $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\beta}$ are unknown, we need to follow an iterative procedure for estimation. Let $\boldsymbol{\alpha}$ be a shared set of parameters that control $\boldsymbol{\Sigma}_i$. We compute iterative updates of $\boldsymbol{\Sigma}_i^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ for $t = 1, 2, \ldots$ until convergence. We need to start with some initial values, $\boldsymbol{\beta}^{(1)}, \boldsymbol{\Sigma}_i^{(1)}$.

- Step 0: Initialize $\boldsymbol{\beta}^{(1)}$ based on an appropriate `glm` model and use may use the Step 2 below to initialize $\boldsymbol{\Sigma}_i^{(1)}$'s.

- Step 1: Updating $\boldsymbol{\beta}^{(t+1)}$: GEE estimates the regression coefficients $\boldsymbol{\beta}$ by maximizing a quasi-likelihood function given $\boldsymbol{\Sigma}_i^{(t)}$. The quasi-likelihood function is specified based on the first two moments of the response variable and the working correlation structure. This leads to the following estimating equation: $\sum_{i=1}^{N} \left( \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$ The estimating equation is sum of subject-specific or cluster-specific score functions: $\sum_{i=1}^{N} U_i(\boldsymbol{\beta}) = 0$, where $U_i(\boldsymbol{\beta}) = \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}))$ is the score function for subject/cluster $i$.

- Step 2: Updating $\boldsymbol{\Sigma}_i^{(t+1)}$: This includes updating $\phi$ and $\boldsymbol{\alpha}$. Given the values of $\boldsymbol{\beta}^{(t+1)}$, we get $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\boldsymbol{\beta}^{(t+1)})$. We need to calculate standardize residuals which are $\hat{e}_{i,j} = \frac{y_{i,j} - \hat{\mu}_{i,j}}{\sqrt{v(\hat{\mu}_{i,j})}}$. Update of $\phi$ (if specified in the model. Like in binary case, it is not added, but if added, it's update will be) $\hat{\phi}^{(t+1)} = \frac{\sum_{i,j}(\hat{e}_{i,j}^{(t+1)})^2}{\sum_{i=1}^{N} n_i}$.

  The pairwise association parameters, $\boldsymbol{\alpha}$, can be estimated similarly, depending on the model for the within-subject association in the third component of the marginal model. For example, in a balanced design ($n_i = n$ for all $i$) with unstructured correlation, we set $Corr(\mathbf{y}_i) = \boldsymbol{\alpha}$, a $n \times n$ matrix with $\alpha_{j,k} = Corr(y_{i,j}, y_{i,k})$ can be estimated by $\hat{\alpha}_{j,k}^{(t+1)} = \frac{1}{\hat{\phi}^{(t+1)} N} \sum_{i=1}^{N} \hat{e}_{i,j}^{(t+1)} \hat{e}_{i,k}^{(t+1)}$. When other structures like autoregressive or Gaussian-kernel-based covariances are assumed, this step needs to be formulated accordingly.

  [1] provides alternative specification for $Corr(\mathbf{y}_i)$ based on log of odds.

- Iterate Steps 1 and 2 until convergence.

### 5.1 Properties

- $\hat{\boldsymbol{\beta}}$ is consistent estimator of $\boldsymbol{\beta}$.

- In large samples, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution

- $Cov(\hat{\boldsymbol{\beta}}) = \mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1}$ where $\mathbf{B} = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)$

  $\mathbf{M} = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_i^{-1} Cov(\mathbf{y}_i) \boldsymbol{\Sigma}_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)$ $\mathbf{B}$ and $\mathbf{M}$ can be estimated by replacing $\boldsymbol{\alpha}$, $\phi$, and $\boldsymbol{\beta}$ by their estimates, and replacing $Cov(\mathbf{y}_i)$ by $(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T$. This is the sandwich variance estimator, which is consistent even under misspecification.

## 6 Interpretation

The interpretation of GEE coefficients $\boldsymbol{\beta}$ is similar to that of coefficients in GLMs. They represent the change in the expected value of the response variable for a unit change in the corresponding covariate, holding other covariates constant. However, these interpretations are at the population-averaged level rather than the subject-specific level.

## 7 Conclusion

Generalized Estimating Equations (GEE) provide a flexible approach for analyzing correlated data, particularly longitudinal or clustered data. But it only estimates the population average effect. Thus it is called the ‘marginal model’. In that sense, GEE allows for robust inference even when the correlation structure is misspecified. However, it’s essential to assess model assumptions and choose an appropriate correlation structure for reliable results. It also requires large enough sample size.

## 8 Basic difference with GLMM

First, GEE is a nonparametric method, as it makes no probabilistic assumption on the data. The specific steps are designed based on the data type (binary or count or categorical etc). The GEE-based fixed effect estimates are consistent as long as the model for the mean has been correctly specified, even under model misspecification for other parameters. But, the trade-off is that we do not get individual-specific effects in GEE.

Also, estimates of fixed effect terms in GLMM will not match with the estimated effect in GEE. In GEE, say for binary data, we model the marginal expectation of the data $\mathbf{y}_i$ as $\mu_i(\boldsymbol{\beta}) = \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}} = E(\mathbf{y}_i \mid \mathbf{X}_i)$. In GLMM, we model $\mathbf{y}_i \sim$Bernoulli($\mathbf{p}_i$) with logit($\mathbf{p}_i$) $= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$. So, under GLMM, $E(\mathbf{y}_i \mid \mathbf{X}_i) = E_{\mathbf{b}_i}E(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{b}_i) = E_{\mathbf{b}_i}\left(\frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i}}\right) = \int\left(\frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i}}\right) f(\mathbf{b}_i)\partial\mathbf{b}_i$, where $f(\mathbf{b}_i)$ is the density for the random effect $\mathbf{b}_i$.

Since, $\int\left(\frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i}}\right) f(\mathbf{b}_i)\partial\mathbf{b}_i \neq \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}}$ or for any link $g^{-1}$, the fixed effect estimates from GLMM will not match with GEE estimates even under identical assumptions on the correlation.

**Due to the above inequality, the interpretation of $\beta$ in GLMM is the population average effect on the log-odds (for binary data with a logistic link) but not on the outcome. But $\beta$ in GEE is the population average effect on the outcome itself.**

## 9 R implementation

https://library.virginia.edu/data/articles/getting-started-with-generalized-estimating-equations

## References

[1] Vincent Carey, Scott L Zeger, and Peter Diggle. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526, 1993.

[2] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.