

## Marginal Model for Categorical Data

### Outline

- Marginal model
  - Examples of marginal model
- GEE1
- Augmented GEE
  - GEE1.5
  - GEE2
  - Modeling the scale parameter  $\phi$
  - A note on modeling correlation of binary responses
  - Using marginal odds ratios to model association for binary responses

## Marginal Model

- Marginal models (or population-average model) distinguish from the mixed effects models (or subject-specific models) according to the interpretation of their regression coefficients. The marginal models are used to make inferences about population average.
- Marginal models emphasize the dependence of the mean response on the covariates of interest, but not on both the random effects and the covariates (mixed effects models), nor on previous responses (transition models).
- Marginal models do not require the joint distributional assumptions for the vector of responses, which may be difficult for the discrete data. The avoidance of distributional assumptions leads to a method of estimation of *generalized estimating equations* (GEE).
- For longitudinal data, the marginal model separately model the mean responses and within-subject association among the repeated responses.

## Notation of the marginal models

Let the random variable  $Y_{ij}$  denote the responses for the  $i$ th individual, measured at time  $t_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ . The response variables for the  $i$ th subject is an  $n_i \times 1$  vector,  $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})'$ .

Let  $\mathbf{X}_{ij}$  denote  $p \times 1$  vector of covariates  $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \dots, X_{ijp})'$  that are associated with each response,  $Y_{ij}$ .  $\mathbf{X}_{ij}$  may include covariates whose values do not change over time, i.e. time-stationary or between-subjects covariates, and covariates whose values change over time, i.e., time-varying or within-subject covariates. The covariates can be grouped into an  $n_i \times p$  matrix:

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{X}'_{i1} \\ \mathbf{X}'_{i2} \\ \vdots \\ \mathbf{X}'_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \dots & X_{i1p} \\ X_{i21} & X_{i22} & \dots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \dots & X_{in_ip} \end{pmatrix}, \quad i = 1, \dots, m$$

- A marginal model has the following components:

1. **Mean model:** the marginal mean of each response,  $E(Y_{ij} | \mathbf{X}_{ij}) = \mu_{ij}$ , depends on covariates via a known link function

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

2. **Correlation model** (nuisance):

$$\text{Var}(Y_{ij} | \mathbf{X}_i) = \phi V(\mu_{ij})$$

$$\text{Cor}(Y_{ij}, Y_{ik} | \mathbf{X}_i) = \rho_{ijk}$$

$$\text{Cov}(\mathbf{Y}_i | \mathbf{X}_i) = V_i(\phi, \boldsymbol{\alpha}) = \phi \mathbf{C}_i^{1/2} \mathbf{R}_i \mathbf{C}_i^{1/2}$$

where  $\mathbf{R}_i$  is the correlation matrix and  $\mathbf{C}_i = \text{diag}(V(\mu_{ij}))$  is a diagonal matrix of variances. The parameter  $\boldsymbol{\alpha}$  characterizes the correlation and  $\phi$  is a scale parameter for variances.

- Estimation and valid inference are achieved by constructing an unbiased estimating function without likelihood function.

### Example: marginal model for a continuous response

Consider  $Y_{ij}$  is a continuous response and we are interested in relating mean response to the covariates  $\mathbf{X}_i$ , The marginal mean model is:

$$\mu_{ij} = \eta_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta}.$$

The variance model is:

$$\begin{aligned} \text{Var}(Y_{ij} \mid \mathbf{X}_i) &= \phi V(\mu_{ij}) = \phi \\ \text{Cor}(Y_{ij}, Y_{ik} \mid \mathbf{X}_i) &= \alpha^{|k-j|} \end{aligned}$$

where  $0 \leq \alpha^{|k-j|} \leq 1$  is the pairwise correlation among the responses.

Note in this example we assume the variance are homogeneous over time, other variance functions are also allowed, e.g. a separate scale parameter,  $\phi_j$ . We specify a first-order autoregressive correlation pattern in this example, other correlation structure, e.g. exchangeable or unstructured are also possible.

### Example: marginal model for count data

Consider  $Y_{ij}$  is a count and we are interested to relate changes in the expected count to the covariates  $\mathbf{X}_i$ , The marginal mean model is:

$$\log(\mu_{ij}) = \eta_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta}.$$

The variance model is:

$$\text{Var}(Y_{ij} \mid \mathbf{X}_i) = \phi \mu_{ij}$$

$$\text{Cor}(Y_{ij}, Y_{ik} \mid \mathbf{X}_i) = \alpha_{jk}$$

Note this is a log-linear regression model with an extra-Poisson variance assumption which allows the variance to be inflated by a factor of  $\phi$  to deal with the issue of *over-dispersion*.

### Example: marginal model for a binary response

- Consider the effect of vitamin A deficiency (Xerophthalmia,  $X$ ) on respiratory infection (RI,  $Y$ ) in the Indonesian Children's Health Study. Let  $i$  indicate the child and  $j$  the visit. The marginal mean model is:

$$\text{logit}(\mu_{ij}) = \log \frac{\Pr(Y_{ij} = 1)}{\Pr(Y_{ij} = 0)} = \beta_0 + \beta_1 I_{X_{ij}=1}.$$

The variance model is:

$$\text{Var}(Y_{ij}) = \mu_{ij}(1 - \mu_{ij}),$$

$$\text{Cor}(Y_{ij}, Y_{ik}) = \alpha.$$

- The parameter of interest is  $\beta_1$ ,

$$\exp(\beta_1) = \frac{\text{Odds of RI among vitamin A deficient children}}{\text{Odds of RI among non-deficient children}}.$$

- When the prevalence of RI is low, the odds ratio (OR) is approximately the same as relative risk

(RR).

- The risk may be different for different children with the same covariates, so the parameter is a population average (assuming random sample).
- The correlation between two binary variables  $Y_1$  and  $Y_2$  has a constrained range that depends on  $\mu_1$  and  $\mu_2$ . So it might be desirable to model the correlation differently, for example, using the marginal odds ratio.

$$\log \text{OR}(Y_{ij}, Y_{ik}) = \alpha_{jk}.$$

where

$$\text{OR}(Y_{ij}, Y_{ik}) = \left\{ \frac{Pr(Y_{ij} = 1, Y_{ik} = 1)Pr(Y_{ij} = 0, Y_{ik} = 0)}{Pr(Y_{ij} = 1, Y_{ik} = 0)Pr(Y_{ij} = 0, Y_{ik} = 1)} \right\}$$



## GEE1

### Estimating function

- When  $(\phi, \boldsymbol{\alpha})$  are known, then the estimator  $\hat{\boldsymbol{\beta}}$  is defined by the estimating equation:

$$\mathbf{0} = \sum_{i=1}^m D_i^T V_i^{-1} \{Y_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\},$$

where

$$D_i(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}, D_i(j, k) = \frac{\partial \mu_{ij}}{\partial \beta_k},$$

$$V_i(\boldsymbol{\beta}, \phi, \boldsymbol{\alpha}) = \phi C_i^{1/2} R_i(\boldsymbol{\alpha}) C_i^{1/2}.$$

where  $V_i$  is known as a “working” covariance to distinguish it from the true underlying covariance among the  $Y_{ij}$ ’s and  $R_i$  is a “working” correlation.

- For example, for logistic model with one covariate and exchangeable correlation,

$$\mu_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{ij})}{1 + \exp(\beta_0 + \beta_1 x_{ij})}$$

$$D_i(j) = \left( \frac{\partial \mu_{ij}}{\partial \beta_0}, \frac{\partial \mu_{ij}}{\partial \beta_1} \right)$$

$$C_i = \begin{pmatrix} \mu_{i1}(1 - \mu_{i1}) & 0 & \cdots & 0 \\ 0 & \mu_{i2}(1 - \mu_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_{in_i}(1 - \mu_{in_i}) \end{pmatrix}$$

$$R_i = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha & 1 \end{pmatrix}$$

## GEE1 — estimation

An iterative algorithm is used to find  $(\hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\boldsymbol{\alpha}})$ :

1. Starting with an initial estimate of  $\boldsymbol{\beta}$  with an ordinary generalized linear model assuming independence.
2. Given  $\hat{\boldsymbol{\beta}}^{(l)}$ , calculate method-of-moments estimates for  $\phi$ ,  $\boldsymbol{\alpha}$ , and the estimate of covariance of  $\mathbf{V}_i$ :

$$\hat{V}_i = \hat{\phi} \hat{\mathbf{C}}_i^{1/2} R_i(\hat{\boldsymbol{\alpha}}) \hat{\mathbf{C}}_i^{1/2}$$

3. Given estimates for  $\phi$ ,  $\boldsymbol{\alpha}$  and  $\mathbf{V}_i$ , solve the estimating equation using Fisher's scoring algorithm:

$$\hat{\boldsymbol{\beta}}^{(l+1)} = \hat{\boldsymbol{\beta}}^{(l)} + \left( \sum_{i=1}^m \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \sum_{i=1}^m \hat{D}_i^T \hat{V}_i^{-1} \{\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i\}.$$

4. Iterate the above two steps until convergence is achieved.

## GEE1 — working correlation

- The model chosen for  $R_i(\boldsymbol{\alpha})$  is called the “working correlation” since it needs not to be the true correlation to obtain a valid point estimate  $\hat{\boldsymbol{\beta}}$  (consistent and asymptotically normal).
- If  $R_i(\boldsymbol{\alpha})$  is the correct correlation, then the model-based estimates of the standard errors for  $\hat{\boldsymbol{\beta}}$  can be used. Otherwise we use the empirical estimates of standard errors.
- Replacing  $\boldsymbol{\alpha}$  with a (any) consistent estimator does not affect the large sample properties of  $\hat{\boldsymbol{\beta}}$ . The asymptotic variance of  $\hat{\boldsymbol{\beta}}$  would be the same as if  $\boldsymbol{\alpha}$  is known. (Liang and Zeger, 1986).
- Some working correlation examples
  - Independent correlation: assumes all pairwise correlation coefficients are zero:

$$\text{Cor}(Y_{ij}, Y_{ik}) = 0, j \neq k.$$

- Exchangeable correlation: assumes all pairwise correlation coefficients are equal and hence the components in the response vector  $\mathbf{Y}$  are exchangeable (not ordered):

$$\text{Cor}(Y_{ij}, Y_{ik}) = \alpha, j \neq k.$$

- AR(1) correlation: the autoregressive correlation structure of order 1 assumes the correlation coefficients decay exponentially over time, and the responses are ordered in time and more correlated if they are closer to each other in time than if they are more distant:

$$\text{Cor}(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}, j \neq k.$$

- Unstructured correlation: assumes all pairwise correlation coefficients are different parameters:

$$\text{Cor}(Y_{ij}, Y_{ik}) = \alpha_{jk}, j \neq k.$$

## GEE1 — variance of $\hat{\beta}$

- The solution  $\hat{\beta}$  is consistent and asymptotically normal.
- If the correlation model is correct, then, the *model-based* estimate for the variance of  $\hat{\beta}$  is  $\widehat{\text{Var}}(\hat{\beta}) = A^{-1}$ , where

$$A = \sum_{i=1}^m D_i^T(\hat{\beta}) V_i^{-1}(\hat{\beta}, \phi, \alpha) D_i(\hat{\beta}).$$

- If the correlation model is *not* correct, then we can use the *empirical* (so called “sandwich”) variance estimate:

$$\begin{aligned} \tilde{\text{Var}}(\hat{\beta}) &= A^{-1} B A^{-1}, \\ B &= \sum_{i=1}^m U_i U_i^T \\ &= \sum_{i=1}^m D_i^T(\hat{\beta}) V_i^{-1}(\hat{\beta}, \phi, \alpha) \hat{\text{Cov}}(\mathbf{Y}_i) V_i^{-1}(\hat{\beta}, \phi, \alpha) D_i(\hat{\beta}) \end{aligned}$$

where

$$\hat{\text{Cov}}(\mathbf{Y}_i) = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i) (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)^T.$$

- $\hat{\text{Cov}}(\mathbf{Y}_i)$  is a poor estimator for  $\text{Cov}(\mathbf{Y}_i)$ . However, we do not need a good estimator for each  $\text{Cov}(\mathbf{Y}_i)$ . With sufficient independent replication ( $m$  large), the average covariance can be well estimated (consistency).
- What if  $(\phi, \boldsymbol{\alpha})$  are unknown? How can we estimate them and what is the impact on the estimation of  $\boldsymbol{\beta}$ ?
- Liang and Zeger (1986) proposed to use moment estimators for the unknown parameters (GEE1).

## GEE1 — estimating $\alpha$ using method of moments

- Let  $N = \sum_{i=1}^m n_i$ .
- Recall that  $\text{Var}(Y_{ij} \mid \mathbf{X}_i) = \phi V(\mu_{ij})$  where  $V$  is a known function.
- The scale parameter  $\phi$ , if exists, can be estimated by

$$\hat{\phi} = \frac{1}{N - p} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{V(\hat{\mu}_{ij})},$$

where  $p$  is the dimension of  $\beta$ .

- Binomial:  $V(\hat{\mu}_{ij}) = \hat{\mu}_{ij}(1 - \hat{\mu}_{ij})$ .
- Poisson:  $V(\hat{\mu}_{ij}) = \hat{\mu}_{ij}$ .



- Define the residuals:

$$r_{ij} = \frac{Y_{ij} - \hat{\mu}_{ij}(\hat{\boldsymbol{\beta}})}{\left[\hat{\text{Var}}(Y_{ij})\right]^{1/2}},$$

where  $\hat{\text{Var}}(Y_{ij}) = \hat{\phi}V(\hat{\mu}_{ij})$ .

- The correlation parameter  $\boldsymbol{\alpha}$  can be estimated as simple functions of  $r_{ij}$ , e.g.
  - Exchangeable correlation:

$$\hat{\alpha} = \frac{1}{\left(\frac{1}{2} \sum_i n_i(n_i - 1) - p\right)\hat{\phi}} \sum_{i=1}^m \sum_{j < k} r_{ij} r_{ik}.$$

- AR(1) correlation:

$$\hat{\alpha} = \frac{1}{\left(\sum_i (n_i - 1) - p\right)\hat{\phi}} \sum_{i=1}^m \sum_{j \leq n_i - 1} r_{ij} r_{i,j+1}.$$

- Unstructured correlation:

$$\hat{\alpha}_{jk} = \frac{1}{(m - p)\hat{\phi}} \sum_{i=1}^m r_{ij} r_{ik}.$$

## GEE1 — more about $\alpha$

- Recall that we use simple MoM estimators to estimate  $(\phi, \alpha)$  in GEE1. The scale parameter  $\phi$  is often considered to be a nuisance and it does not affect the estimates of  $\beta$ . But what about  $\alpha$ ?
  1. Shouldn't we consider the parameter as  $(\beta, \alpha)$ ?
  2. Can't we improve upon the estimation of  $\alpha$ ?
  3. Would “better” estimation of  $\alpha$  help us to “better” estimate  $\beta$ ?
- Shouldn't we consider the parameter as  $(\beta, \alpha)$ ?

**Answer:** It depends.

- Is  $\alpha$  a nuisance? If the covariance structure is of secondary interest (often the case) then GEE1 is usually fine. However, if the covariance matrix *is* of primary interest then GEE1 is not ideal.
- Are you willing to sacrifice some model robustness in order to let  $(\beta, \alpha)$  be the target parameter?

Note that in GEE1, the estimate  $\hat{\beta}$  is consistent even if the model for  $\alpha$  is wrong. Other approaches

that treat  $\beta$  and  $\alpha$  on equal ground may not have this property.

- Can't we improve upon the estimation of  $\alpha$ ?

**Answer:** Yes!

- We can adopt alternative association (dependence) models that are more suitable for categorical data.
- We can create estimators that are targeted at  $(\beta, \alpha)$  jointly and are efficient for both.
- Would “better” estimation of  $\alpha$  help us to “better” estimate  $\beta$ ?

**Answer:** Model for  $\alpha$  is important for the efficiency of  $\hat{\beta}$ .

## GEE1 — more about the “sandwich” estimator of $\text{Cov}(\hat{\beta})$

- It is a robust estimator as it provides valid standard error when the assumed model for the covariance among the repeated measures, i.e. “working” covariance, is misspecified.
- With the sandwich estimator, why bother expending effort to model the within-subject association in marginal model?
  - Efficiency or precision could be improved with a “working” covariance more closer to the true underlying covariance.
  - The robustness property depends on the large sample property, i.e., relative large number of subjects and relative small number of repeated measures.
- The sandwich estimator may be problematic for severely unbalanced data. The model-based variance estimator should be used instead.

## GEE1 — hypothesis testing

- **Generalized Wald Test:**

- $H_0 : \beta_j = 0$

$$\frac{\hat{\beta}_j}{\text{s.e.}} \sim \mathcal{N}(0, 1).$$

- Write  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ ,  $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$

$$\hat{\boldsymbol{\beta}}_1^T \hat{\text{Var}}_1(\hat{\boldsymbol{\beta}})^{-1} \hat{\boldsymbol{\beta}}_1 \sim \chi_r^2,$$

where  $r$  is the dimension of  $\boldsymbol{\beta}_1$  and  $\hat{\text{Var}}_1(\hat{\boldsymbol{\beta}})$  is the estimated variance matrix corresponding to  $\hat{\boldsymbol{\beta}}_1$  ( $r \times r$  subprincipal matrix of  $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$ ).

- $H_0 : L\boldsymbol{\beta} = 0$ , where  $L$  is a  $r \times p$  matrix,

$$\hat{\boldsymbol{\beta}}^T L^T (L \hat{\text{Var}}(\hat{\boldsymbol{\beta}}) L^T)^{-1} L \hat{\boldsymbol{\beta}} \sim \chi_r^2,$$

where  $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$  is the estimated variance matrix for  $\hat{\boldsymbol{\beta}}$ .

- **Generalized Score Test:**  $H_0 : L\boldsymbol{\beta} = \mathbf{0}$  where  $L$  is a  $r \times p$  matrix.

$$T = \mathbf{S}(\tilde{\boldsymbol{\beta}})^T \mathbf{A} L^T (L \mathbf{B} L^T)^{-1} L \mathbf{S}(\tilde{\boldsymbol{\beta}}) \sim \chi_r^2.$$

where  $\mathbf{A}$  and  $\mathbf{B}$  are model-based and empirical covariance estimates,  $\mathbf{S}(\tilde{\boldsymbol{\beta}})$  is the GEE evaluated at the regression parameter estimate  $\tilde{\boldsymbol{\beta}}$  under  $H_0$ . Rotnitzky and Jewell (1990).

- In the generalized Wald test statistic,  $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$  using the sandwich form can be unstable if  $n_i$ 's are large and  $m$  is small. If the correlation structure has been carefully modelled, the model-based (or “working”)  $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$  can be used.
- The generalized score test is invariant under any differentiable transformation of the parameters,  $\boldsymbol{\tau} = \boldsymbol{\tau}(\boldsymbol{\beta})$ .

## Augmented GEE

### GEE1.5 — Prentice (1988)'s approach for binary data

- **Idea:** Estimation of both parameters vectors,  $\beta$  and  $\alpha$ , in the marginal response model and the correlation model, respectively.
- GEE1 uses an estimating function  $U_1$  based on the centered first moments  $(\mathbf{Y}_i - \boldsymbol{\mu}_i)$  for the estimation of  $\beta$ . We can add a second estimation function based on the centered *second* moments to estimate  $\alpha$ :

$$\frac{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})}{[\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})]^{1/2}} - \sigma_{ijk}$$

where  $\sigma_{ijk} = E \left[ \frac{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})}{[\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})]^{1/2}} \right]$  is the mean of the sample correlation.

- Now, the joint estimating functions are:

$$\mathbf{U}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^m D_i^T(\boldsymbol{\beta}) V_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} \quad (1)$$

$$\mathbf{U}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^m E_i^T(\boldsymbol{\beta}, \boldsymbol{\alpha}) W_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \{\mathbf{S}_i - \boldsymbol{\sigma}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})\}, \quad (2)$$

where

$$\mathbf{S}_i = (S_{i1}S_{i2}, S_{i1}S_{i3}, S_{i1}S_{in_i}, \dots, S_{in_i-1}S_{in_i})^T$$

$$S_{ij} = (Y_{ij} - \mu_{ij})/[\mu_{ij}(1 - \mu_{ij})]^{1/2}$$

$$\boldsymbol{\sigma}_i = E(\mathbf{S}_i)$$

$$E_i = \frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\alpha}}$$

$$W_i = \text{diag}(\text{Var}(S_{i1}S_{i2}), \dots, \text{Var}(S_{in_i-1}S_{in_i}))$$

$$\text{Var}(S_{ij}S_{ik}) = 1 + (1 - 2\mu_{ij})(1 - 2\mu_{ik})[\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})]^{-1/2}\sigma_{ijk} - \sigma_{ijk}^2 \quad (\text{Check this})$$



- Note that the above  $\mathbf{W}_i$  is an  $n_i(n_i - 1)/2$ -dimensional working independent variance matrix for  $\mathbf{S}_i$ . Other working matrices could readily be substituted. To model  $\text{Cov}(\mathbf{S}_i)$  properly, we need specify models for higher moments (with more parameters), which is typically difficult.
- Iterative method can be used for solving the joint estimating equations

$$\mathbf{0} = \mathbf{U}_1(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

$$\mathbf{0} = \mathbf{U}_2(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

Given  $(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\boldsymbol{\alpha}}^{(l)})$ :

1. Fixed  $\hat{\boldsymbol{\alpha}}^{(l)}$ , solve  $\mathbf{0} = \mathbf{U}_1(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}^{(l)})$  to get  $\hat{\boldsymbol{\beta}}^{(l+1)}$ .
  2. Fixed  $\hat{\boldsymbol{\beta}}^{(l+1)}$ , solve  $\mathbf{0} = \mathbf{U}_2(\hat{\boldsymbol{\beta}}^{(l+1)}, \boldsymbol{\alpha})$  to get  $\hat{\boldsymbol{\alpha}}^{(l+1)}$ .
- $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$  is consistent and asymptotically normal under correct model specification. Similar to GEE1,  $\hat{\boldsymbol{\beta}}$  is consistent even if the model for  $\boldsymbol{\alpha}$  is misspecified.

## GEE2 — Prentice and Zhao (1991)'s approach

- They considered a systematic approach for generating estimating functions for  $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ .
- Combine the response vector  $\mathbf{Y}_i$  and the pairwise crossproducts,  $\mathbf{S}$ , into one outcome vector

$$\mathbf{T}_i^T = (\mathbf{Y}_i^T, \mathbf{Z}_i^T),$$

where  $\mathbf{Y}_i^T = (y_{i1}, \dots, y_{in_i})$ , and  $\mathbf{Z}_i^T = (y_{i1}^2, y_{i1}y_{i2}, \dots, y_{i2}^2, y_{i2}y_{i3}, \dots, y_{in_i}^2)$ .

- The estimating equations for  $(\boldsymbol{\beta}, \boldsymbol{\alpha})$  can be generated under a **Quadratic Exponential Family** (QEF) model, where the loglikelihood

$$\Pr_i(\mathbf{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \propto \boldsymbol{\theta}_{1i}^T \mathbf{Y}_i + \boldsymbol{\theta}_{2i}^T \mathbf{Z}_i + c_i(\mathbf{Y}_i),$$

where the canonical parameters  $(\boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i})$  are functions of  $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ .

- Optimal estimating equations for  $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ :

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\delta}) &= \sum_{i=1}^m \boldsymbol{D}_i^T(\boldsymbol{\delta}) \boldsymbol{V}_i^{-1}(\boldsymbol{\delta}) \boldsymbol{f}_i(\boldsymbol{\delta}) \\ &= \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\beta}} \\ 0 & \frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\alpha}} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_i(1, 1) & \boldsymbol{V}_i(1, 2) \\ \boldsymbol{V}_i(1, 2)^T & \boldsymbol{V}_i(2, 2) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{s}_i - \boldsymbol{\sigma}_i \end{pmatrix} \end{aligned}$$

$$\boldsymbol{V}_i(1, 1) = \text{Cov}(\boldsymbol{Y}_i)$$

$$\boldsymbol{V}_i(1, 2) = \text{Cov}(\boldsymbol{Y}_i, \boldsymbol{S}_i)$$

$$\boldsymbol{V}_i(2, 2) = \text{Cov}(\boldsymbol{S}_i)$$

$$S_{ijk} = (Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})$$

$$\boldsymbol{\sigma}_i = \text{E}[\boldsymbol{S}_i], \sigma_{ijk} = \text{Cov}(Y_{ij}, Y_{ik}).$$

- Third and fourth order moments are contained in  $\boldsymbol{V}_i$ , specifically  $\boldsymbol{V}_i(1, 2)$  and  $\boldsymbol{V}_i(2, 2)$ .

- One can specify a working variance matrix in  $V_i$ , in which the third and fourth moments are expressed as functions of  $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$ .

- Independence working models

$$V_i(1, 2) = \mathbf{0}$$

$$V_i(2, 2) = \text{diagonal matrix}$$

- Gaussian working models

$$V_i(1, 2) = \mathbf{0}$$

$$V_i(2, 2) : \text{Cov}(S_{ijk}, S_{ij'k'}) = \sigma_{ijj'}\sigma_{ikk'} + \sigma_{ijk'}\sigma_{ikj'}$$

- Estimation can be done using the Fisher scoring procedure

$$\begin{pmatrix} \boldsymbol{\beta}^{(l+1)} \\ \boldsymbol{\alpha}^{(l+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(l)} \\ \boldsymbol{\alpha}^{(l)} \end{pmatrix} + \left( \sum_i D_i^T V_i^{-1} D_i \right)^{-1} \left( \sum_i D_i^T V_i^{-1} \mathbf{f}_i \right)$$

- Sandwich variance estimator can be used to protect against the 3rd/4th moment model misspecification.
- Consistency of both  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\alpha}}$  depends on the **correct modeling of both mean and covariance**.
- The matrix  $V_i$  has dimension  $M_i \times M_i$  where  $M_i = n_i + n_i(n_i + 1)/2$ , and its inverse is required!
- GEE1.5 is a special case of GEE2. The efficiency gain of GEE2 comparing with GEE1.5 depends on the correct specification of the 3rd/4th moments.
- Conclusion: GEE2 may not be worthwhile after all. If we want to specify higher order moments, why not use the likelihood?

## Modeling the scale parameter $\phi$ — Yan and Fine (2004)'s approach

Remember that  $\text{Cov}(\mathbf{Y}_i) = \phi \mathbf{C}_i^{1/2} \mathbf{R}_i \mathbf{C}_i^{1/2}$ . When the scale parameter  $\phi$  (i.e the over-dispersion parameter for heteroscedasticity) is important, Yan and Fine (2004) extended GEE1.5 by adding a third model for the scale parameter :

$$g_3(\phi_{ij}) = \mathbf{X}_{3ij}^T \boldsymbol{\gamma},$$

where  $g_3$  is the link function, e.g. log. The estimating equation is

$$\mathbf{U}_\phi = \sum_{i=1}^m \mathbf{D}_{3i}^T \mathbf{V}_{3i}^{-1} (\mathbf{T}_i - \boldsymbol{\phi}_i) = \mathbf{0}, \quad (3)$$

where

$$\begin{aligned} \mathbf{T}_i &= (T_{i1}, \dots, T_{in_i})^T, T_{ij} = \frac{(Y_{ij} - \mu_{ij})^2}{V(\mu_{ij})}, \\ \mathbf{D}_{3i} &= \frac{\partial \boldsymbol{\phi}_i}{\partial \boldsymbol{\gamma}}, \\ \boldsymbol{\phi}_i &= (\phi_{i1}, \dots, \phi_{in_i})^T. \end{aligned}$$

- $V_{3i}$  also contains third and fourth moments.
- $\hat{\beta}$  is consistent with mis-specification of the correlation and scale parameters;  $\hat{\gamma}$  is consistent if the mean and the scale structures are correctly specified, regardless of mis-specification of the correlation.
- The method is implemented in R package **geepack**.

## A Note on Modeling Correlation of Binary Responses

- Correlation for binary data are constrained by their means.

Let  $E(Y_1) = \mu_1$ ,  $E(Y_2) = \mu_2$ ,  $\rho_{12} = \text{Cor}(Y_1, Y_2)$ , and  $\pi_{12} = E(Y_1 Y_2)$ , then  $\pi_{12} < \min(\mu_1, \mu_2)$ , and hence,

$$\rho_{12}^2 \leq \min \left\{ \frac{\mu_1(1 - \mu_2)}{\mu_2(1 - \mu_1)}, \frac{\mu_2(1 - \mu_1)}{\mu_1(1 - \mu_2)} \right\},$$

because the pairwise correlation

$$\rho_{12} = \frac{\pi_{12} - \mu_1 \mu_2}{[\mu_1(1 - \mu_1)\mu_2(1 - \mu_2)]^{1/2}},$$

and  $\pi_{12}$  is constrained to satisfy

$$\max(0, \mu_1 + \mu_2 - 1) < \pi_{12} < \min(\mu_1, \mu_2).$$



- Modeling odds ratios:

$$\begin{aligned}\Psi_{ijk} &= \frac{\Pr(Y_{ij} = 1, Y_{ik} = 1) \Pr(Y_{ij} = 0, Y_{ik} = 0)}{\Pr(Y_{ij} = 1, Y_{ik} = 0) \Pr(Y_{ij} = 0, Y_{ik} = 1)}, \\ &= \frac{\Pr(Y_{ij} = 1 \mid Y_{ik} = 1) / \Pr(Y_{ij} = 0 \mid Y_{ik} = 1)}{\Pr(Y_{ij} = 1 \mid Y_{ik} = 0) / \Pr(Y_{ij} = 0 \mid Y_{ik} = 0)}.\end{aligned}$$

- The log odds ratios  $\log \Psi \in (-\infty, \infty)$ , are symmetric about 0 and not constrained by the marginal means.
- Interpretation:  $\Psi_{ijk} = 1$  or  $\log \Psi_{ijk} = 0$  implies  $(Y_{ij}, Y_{ik})$  are uncorrelated.
- The odds ratio,  $\Psi_{ijk}$  and the marginal means  $\mu_{ij}$ ,  $\mu_{ik}$  determine  $\pi_{ijk} = E(Y_{ij}Y_{ik})$ , hence determine the correlation  $\rho_{ijk}$ , and variance  $V_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ .

The odds ratio of  $\Psi_{ijk}$  can be expressed as function of  $\mu_{ij}$ ,  $\mu_{ik}$  and  $\pi_{ijk}$ :

$$\Psi_{ijk} = \frac{\pi_{ijk}(1-\mu_{ij}-\mu_{ik}+\pi_{ijk})}{(\mu_{ij}-\pi_{ijk})(\mu_{ik}-\pi_{ijk})} \quad (4)$$

Thus,  $\pi_{ijk}$  can be solved from the above and expressed as function of  $\mu_{ij}$ ,  $\mu_{ik}$  and  $\Psi_{ijk}$ :

$$\pi_{ijk} = \frac{A - [A^2 - 4(\Psi_{ijk} - 1)\Psi_{ijk}\mu_{ij}\mu_{ik}]^{1/2}}{2(\Psi_{ijk} - 1)} \quad \text{when } \Psi_{ijk} \neq 1$$

$$\pi_{ijk} = \Psi_{ijk}\mu_{ij}\mu_{ik} \quad \text{when } \Psi_{ijk} = 1$$

where

$$A = 1 - (\mu_{ij} + \mu_{ik})(1 - \Psi_{ijk})$$

Lipsitz et al (1991)

## Using Marginal Odds Ratios to Model Association for Binary Responses

### Lipsitz et al (1991)'s approach

- They modified the approach of Prentice (1988) using odds ratios.
- The joint models (Lipsitz et al, 1991) are:

$$\text{Mean model} \quad \text{logit}(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}$$

$$\text{Correlation model} \quad \log(\Psi_{ijk}) = \mathbf{Z}_{ijk}^T \boldsymbol{\alpha}$$

- Lipsitz et al. noted that  $\boldsymbol{\beta}$  estimates from odds ratio model are slightly efficient comparing to that from correlation model by simulation and these two models are computationally similar.

## Carey, Zeger, and Diggle (1993)'s alternating logistic regression (ALR) approach

- Let  $\gamma_{ijk} = \log \Psi_{ijk}$ . Note that the pairwise conditional expectations:

$$\text{logit } E(Y_{ij} | Y_{ik}, \mathbf{X}_i) = \gamma_{ijk} Y_{ik} + \Delta_{ijk}, \text{ where}$$

$$\Delta_{ijk} = \log \left( \frac{\mu_{ij} - \pi_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \pi_{ijk}} \right).$$

(Check this using the fact in (4).)

- Suppose that all the odds ratios are the same,  $\gamma_{ijk} = \gamma$ , then an estimate for  $\gamma$  could be obtained by a logistic regression of  $y_{ij}$  on  $y_{ik}$ ,  $1 \leq j < k \leq n_i, i = 1, \dots, m$ , using  $\Delta_{ijk}$  as an offset.
- More generally, we assume  $\log(\Psi_{ijk}) = \mathbf{Z}_{ijk}^T \boldsymbol{\alpha}$ , where  $\mathbf{Z}_{ijk}$  is a set of covariates characterizing the log-odds ratio between observations  $j$  and  $k$ . For example, in family data,  $\mathbf{Z}_{ijk}$  would encode the type of family relationship for  $Y_{ij}$  and  $Y_{ik}$ : husband-wife, parent-sib, or sib-sib. We then estimate the vector  $\boldsymbol{\alpha}$  by a logistic regression of  $y_{ij}$  on the product  $\mathbf{z}_{ijk}^T y_{ik}$  with the offset  $\Delta_{ijk}$ .

- Note that the offset depends on both  $\boldsymbol{\beta}$  and  $\boldsymbol{\alpha}$  so that iterations are required. We alternate the following two steps until convergence.

1. Given the current values of  $(\boldsymbol{\beta}^{(l)}, \boldsymbol{\alpha}^{(l)})$ , calculate  $\hat{V}^{(l)}$  and solve the  $\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(l)}) = 0$  for an updated  $\hat{\boldsymbol{\beta}}^{(l+1)}$ , where

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^m \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T \mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i).$$

2. Given  $\hat{\boldsymbol{\beta}}^{(l+1)}$  and  $\hat{\boldsymbol{\alpha}}^{(l)}$ , evaluate the offset and perform the offset logistic regression of  $Y_{ij}$  on  $Z_{ijk}Y_{ik}$  with a total of  $\sum_{i=1}^m n_i(n_i - 1)/2$  observations to obtain  $\hat{\boldsymbol{\alpha}}^{(l+1)}$ .

- Formally, the ALR uses the same model as in Lipsitz et al. (1991) but uses this estimating equation for  $\boldsymbol{\alpha}$ :

$$\boldsymbol{U}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^m \boldsymbol{F}_i^T(\boldsymbol{\beta}, \boldsymbol{\alpha}) \tilde{\boldsymbol{W}}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \boldsymbol{T}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}), \text{ where}$$

$$\boldsymbol{F}_i = \frac{\partial \boldsymbol{\zeta}_i}{\partial \boldsymbol{\alpha}}$$

$$\zeta_{ijk} = E(Y_{ij} | Y_{ik})$$

$$\boldsymbol{T}_i \text{'s } (j, k) \text{ entry: } T_{ijk} = Y_{ij} - \zeta_{ijk}$$

$$\tilde{\boldsymbol{W}}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \text{diag}(\text{Var}(Y_{ij} | Y_{ik}))$$

$$= \text{diag}(\zeta_{ijk}(1 - \zeta_{ijk})).$$

- No working assumption about the 3rd/4th-order odds ratios are required.
- The efficiency is comparable to GEE2 but more computationally efficient for large clusters (does not require the inverse of large matrices).

## Further Reading

- Chapters 7.1, 7.5, and 8 of DHLZ.

## References

- Cavey V, Zeger SL, and Diggle P (1993). Modelling multivariate binary data with alternating logistic regression. *Biometrika* **80**:517-26.
- Ling KY and Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**:13-22.
- Liang KY, Zeger SL, and Qaquish B (1992). Multivariate regression analysis for categorical data (with discussion). *JRSSB* **54**:3-40.
- Lipsitz SR, Laird NM, and Harrington DP (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* **78**:153-60.
- Prentice RL (1988). Correlated binary regression with covariates specific to each binary observations. *Biometrics* **44**:1033-48.
- Prentice RL and Zhao LP (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**:825-39.
- Rotnitzky, A. and Jewell, N.P., (1990), Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data. *Biometrika*, **77**:485 - 497.
- Yan J and Fine J (2004). Estimating equations for association structures. *Statistics in Medicine* **23**:859-74.