Lecture 9 | 06.05.2024

Marginal models
for non-normal response

GLM extensions for the longitudinal data

- **Marginal models**
  - primary interest is given to the conditional mean structure
  - separate models for the mean and the covariance structure
- **Random effects models**
  - one equation used to account for both—the mean and the covariance
  - mostly used when some subject specific inference is of the main interest
- **Transition models**
  - primary interest again with respect to the mean structure
  - the correlation structure due to historical observations within the subject

GLM extensions for the longitudinal data

- **Marginal models**
  - primary interest is given to the conditional mean structure
  - separate models for the mean and the covariance structure

- **Random effects models**
  - one equation used to account for both—the mean and the covariance
  - mostly used when some subject specific inference is of the main interest

- **Transition models**
  - primary interest again with respect to the mean structure
  - the correlation structure due to historical observations within the subject

All three categories of the regression models for correlated (repeated) observations above lead to the same model (with the same interpretation) for the Gaussian type of the data but for the discrete data different models can produce different interpretations (due to the non-linearity involved n the models)

Marginal models in general

- For simplicity, let $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$ denote a vector of a correlated binary responses for some individual $i \in \{1, \dots, N\}$
- The idea is to model $P[\boldsymbol{Y}_i = \boldsymbol{y} | \mathbb{X}_i]$, for $\boldsymbol{y} \in \{0, 1\}^{\times n_i}$ by utilizing the marginals of the joint distribution (conditionally on $\mathbb{X}_i$) $P[\boldsymbol{Y}_i = \boldsymbol{y} | \mathbb{X}_i]$
- The saturated model has $2^{n_i} - 1$ parameters and different "marginals"
    - First order marginals $\mu_j = P[Y_{ij} = 1]$, for $j = 1, \dots, n_i$
    - Second order marginals $\mu_{jk} = P[Y_{ij} = 1, Y_{ik} = 1]$, for $j \neq k$
    - Third order marginals $\mu_{jkl} = P[Y_{ij} = 1, Y_{ik} = 1, Y_{ikl} = 1]$, for $j \neq k \neq l$
    - ...
    - The n_i^{th} order marginal $\mu_{1,\dots,N} = P[\boldsymbol{Y}_i = \mathbf{1}]$, where $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^{n_i}$

Marginal models in general

- For simplicity, let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ denote a vector of a correlated binary responses for some individual $i \in \{1, \ldots, N\}$
- The idea is to model $P[\boldsymbol{Y}_i = \boldsymbol{y} | \mathbb{X}_i]$, for $\boldsymbol{y} \in \{0, 1\}^{\times n_i}$ by utilizing the marginals of the joint distribution (conditionally on $\mathbb{X}_i$) $P[\boldsymbol{Y}_i = \boldsymbol{y} | \mathbb{X}_i]$
- The saturated model has $2^{n_i} - 1$ parameters and different "marginals"
  - First order marginals $\mu_j = P[Y_{ij} = 1]$, for $j = 1, \ldots, n_i$
  - Second order marginals $\mu_{jk} = P[Y_{ij} = 1, Y_{ik} = 1]$, for $j \neq k$
  - Third order marginals $\mu_{jkl} = P[Y_{ij} = 1, Y_{ik} = 1, Y_{ikl} = 1]$, for $j \neq k \neq l$
  - ...
  - The n_i^{th} order marginal $\mu_{1,\ldots,N} = P[\boldsymbol{Y}_i = \boldsymbol{1}]$, where $\boldsymbol{1} = (1, \ldots, 1)^\top \in \mathbb{R}^{n_i}$
- Which marginals should be used in the model and how should they explain the overall joint probability $P[\boldsymbol{Y}_i = \boldsymbol{y}]$ (always conditionally on $\mathbb{X}_i$)
  - full log-linear model
  - log-linear model for first and higher order marginals (GEE formulation)
  - Bahadur model for first order marginals and correlations
  - marginal model for μ_j and marginal odds ratios

Full log-linear model

- the joint probability $P[\boldsymbol{Y}_i = \boldsymbol{y}]$ for $\boldsymbol{y} = (y_1, \ldots, y_{n_i})^\top \in \{0,1\} \times \ldots \{0,1\}$ can be expressed as $P[\boldsymbol{Y}_i = \boldsymbol{y}] = P[Y_{i1} = y_1 \wedge Y_{i2} = y_2 \wedge \cdots \wedge Y_{in_i} = y_{n_i}]$
- there are many options how to define a saturated model (with 2^n total parameters but only $2^n - 1$ free parameters)
- the model commonly used model (by Bishop et al. 1975) is a log-linear model which can be formulated as

$$P[\boldsymbol{Y}_i = \boldsymbol{y}] = c(\boldsymbol{\theta}_i) \exp\left\{ \sum_{j=1}^{n} \theta_{ij}^{(1)} y_j + \sum_{j_1 < j_2} \theta_{ij_1j_2}^{(2)} y_{j1} y_{j2} + \cdots + \theta_{i1\ldots n_i}^{(n_i)} y_1 \ldots y_{n_i} \right\}$$

  for the vector $\boldsymbol{\theta}_i = (\theta_{i1}^{(1)}, \ldots, \theta_{in_i}^{(1)}, \theta_{i12}^{(2)}, \ldots, \theta_{i1\ldots n_i}^{(n)})^\top \in \mathbb{R}^{2^{n_i}-1}$ which is the canonical vector of the unknown model parameters ($\theta_{ij}^{(1)}$ are conditional log odds and $\theta_{i\ldots}^{(k)}$ for $k \geq 2$ are conditional log odds ratios)
- however,the association between Y_j and Y_k depends on all other values of Y_l for $l \neq j, k$, respectively

$$\log\left[\frac{P[Y_{ij} = 1 | Y_{ik} = y_k, Y_{il} = 0 \forall l \neq j,k]}{P[Y_{ij} = 0 | Y_{ik} = y_k, Y_{il} = 0 \forall l \neq j,k]}\right] = \theta_{ij}^{(1)} + \theta_{ijk}^{(2)} y_k$$

Full log-linear model

- the joint probability $P[\boldsymbol{Y}_i = \boldsymbol{y}]$ for $\boldsymbol{y} = (y_1, \ldots, y_{n_i})^\top \in \{0,1\} \times \ldots \{0,1\}$ can be expressed as $P[\boldsymbol{Y}_i = \boldsymbol{y}] = P[Y_{i1} = y_1 \wedge Y_{i2} = y_2 \wedge \cdots \wedge Y_{in_i} = y_{n_i}]$
- there are many options how to define a saturated model (with 2^n total parameters but only $2^n - 1$ free parameters)
- the model commonly used model (by Bishop et al. 1975) is a log-linear model which can be formulated as

$$P[\boldsymbol{Y}_i = \boldsymbol{y}] = c(\boldsymbol{\theta}_i) \exp\left\{ \sum_{j=1}^{n} \theta_{ij}^{(1)} y_j + \sum_{j_1 < j_2} \theta_{ij_1j_2}^{(2)} y_{j1} y_{j2} + \cdots + \theta_{i1\ldots n_i}^{(n_i)} y_1 \ldots y_{n_i} \right\}$$

for the vector $\boldsymbol{\theta}_i = (\theta_{i1}^{(1)}, \ldots, \theta_{in_i}^{(1)}, \theta_{i12}^{(2)}, \ldots, \theta_{i1\ldots n_i}^{(n)})^\top \in \mathbb{R}^{2^{n_i}-1}$ which is the canonical vector of the unknown model parameters ($\theta_{ij}^{(1)}$ are conditional log odds and $\theta_{i\ldots}^{(k)}$ for $k \geq 2$ are conditional log odds ratios)
- however,the association between Y_j and Y_k depends on all other values of Y_l for $l \neq j, k$, respectively

$$\log\left[\frac{P[Y_{ij} = 1 | Y_{ik} = y_k, Y_{il} = 0 \forall l \neq j, k]}{P[Y_{ij} = 0 | Y_{ik} = y_k, Y_{il} = 0 \forall l \neq j, k]}\right] = \theta_{ij}^{(1)} + \theta_{ijk}^{(2)} y_k$$

- it would be more interesting to model $P[Y_j = 1 | \boldsymbol{X}] = E[Y_j | \boldsymbol{X}] = \mu_j$

Marginal models towards GEE

- **Mean structure**
  The marginal (conditional) expectation of the response depends (non-linearly) on a linear combination of the explanatory variables (i.e., linear predictor)

$$h(\mu_{ij}) = \boldsymbol{X}_{ij}^{\top}\beta, \qquad \text{for } \mu_{ij} = E[Y_{ij}|\boldsymbol{X}_{ij}] \text{ and } \boldsymbol{\beta} \in \mathbb{R}^{p}$$

  for a known, strictly monotone, and twice continuously differentiable function h

- **Variance structure**
  The marginal (conditional) variance of the response depends on the marginal mean (and, optionally, some other overdispersion parameter $\phi > 0$) as

$$Var(Y_{ij}|\boldsymbol{X}_{ij}) = v(\mu_{ij})\phi, \quad \text{for } \phi > 0$$

  for a known positive and continuously differentiable function v

- **Covariance structure**
  The correlation between two observations Y_{ij} and Y_{ik} (within the same subject $i \in \{1, \ldots, N\}$) is assumed to be modeled as

$$Cor(Y_{ij}, Y_{ik}|\boldsymbol{X}_{ij}, \boldsymbol{X}_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \boldsymbol{\alpha}), \quad \text{for } \boldsymbol{\alpha} \in \mathbb{R}^{q}$$

  for a known covariance function ρ

Key pivots of the marginal models

- Instead of specifying the whole distribution (i.e., the exponential family of distributions) which is required, for instance, for the likelihood based estimation in GLM, only a specification of the first two moments (and their mutual relationship) is provided (quasi-likelihood and GEE instead)
- For $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$ and $\mathbb{X}_i = (\boldsymbol{X}_{i1}, \dots, \boldsymbol{X}_{in_i})^\top$ we separately specify the mean $E[\boldsymbol{Y}_i|\mathbb{X}_i] = \mathbb{X}_i\beta$ and the covariance $Var[\boldsymbol{Y}_i|\mathbb{X}_i] = \mathbb{V}_i(\mathbb{X}_i, \beta, \phi, \boldsymbol{\alpha})$
- As the distribution is not provided the likelihood based on the data can not be constructed. Therefore, in a sense, this is not a parametric model (where the parameters specify the whole distribution) but rather a semi-parametric one (the parameters only specify the first two moments)
- In a log-linear model (with the first (β) and higher order ($\boldsymbol{\alpha}$) marginals) the score function for $\beta \in \mathbb{R}^p$ leads to a GEE formulation with the equations $(\partial\boldsymbol{\mu}/\partial\beta)^\top[Var(\boldsymbol{Y})]^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0}$
- Therefore, in order to use the quasi-likelihood estimation approach appropriately, one has to correctly specify both, the mean and the variance-covariance function

Maximum likelihood for GLM

- for some generic random vector $(Y, \boldsymbol{X}^\top)^\top \sim F_{(Y,\boldsymbol{X})}$ we assume that $f(y|\boldsymbol{X}) = \exp\{\phi^{-1}(y\theta - \psi(\theta)) + c(y, \phi)\}$ is an exponential family
- the linear predictor $\theta = \boldsymbol{X}^\top\beta$ is associated with the conditional mean $E[Y|\boldsymbol{X}] = \mu$ via the link function g, such that $g(\mu) = \theta \equiv \theta(\beta)$
- as far as $f(y|\boldsymbol{X})$ is a probability density function, it holds that $\int(y|\boldsymbol{X})\mathrm{d}y = 1$ (integrating with respect to the appropriate measure)
- first and second partial derivatives of $\int(y|\boldsymbol{X})\mathrm{d}y = 1$ with respect to θ yields the following:

$$\frac{\partial}{\partial\theta} : \int [y - \psi'(\theta)]f(y|\boldsymbol{X})\mathrm{d}y = 0$$

and

$$\frac{\partial^2}{\partial\theta^2} : \int [\psi^{-1}(y - \psi'(\theta))^2 - \psi''(\theta)]f(y|\boldsymbol{X})\mathrm{d}y = 0$$

which gives $\mu = E[Y|\boldsymbol{X}] = \psi'(\theta)$ and $Var[Y|\boldsymbol{X}] = \phi\psi''[(\psi')^{-1}(\mu)]$

Score equations under MLE

- for the random sample $\mathcal{D}_S = \{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, N\}$ we have $f(y|\boldsymbol{X}_i) = \exp\{\phi^{-1}(y\theta_i - \psi(\theta_i)) + c(y, \phi)\}$ where $\theta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} \equiv \theta_i(\boldsymbol{\beta})$
- as $\theta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ and we aim to estimate the unknown parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we need partial derivatives with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$

- **Log-likelihood**

$$\ell(\boldsymbol{\beta}, \phi, \mathcal{D}_S) = \frac{1}{\phi} \sum_{i=1}^{N} [Y_i\theta_i - \psi(\theta_i)] + \sum_{i=1}^{N} c(Y_i, \phi)$$

- **First order derivatives wrt. $\boldsymbol{\beta}$**

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi, \mathcal{D}_S)}{\partial \boldsymbol{\beta}} = \frac{1}{\phi} \sum_{i=1}^{N} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} [Y_i - \psi'(\theta_i)]$$

- **Score equations for $\boldsymbol{\beta}$**

$$\mathcal{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} [Y_i - \psi'(\theta_i)] = \boldsymbol{0}$$

Score equations under MLE – continuation

- since $\mu_i = \psi'(\theta_i)$ a also $v_i = v(\mu_i) = \psi''(\theta_i)$ the score equations become

$$\mathcal{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v_i^{-1} [Y_i - \psi'(\theta_i)] = \mathbf{0}$$

  and, thus, the estimation of $\boldsymbol{\beta} \in \mathbb{R}^p$ depends on the exponential distribution only through the mean μ_i and the variance function $v_i = v(\mu_i)$

- the solution is obtained numerically (e.g., Newton-Raphson, iterative re-weighted LS, Fisher scoring)
- the inference about $\boldsymbol{\beta} \in \mathbb{R}^p$ is based on classical maximum likelihood theory (i.e., asymptotic Wald tests, likelihood ratio tests, score tests)
- the estimation of the over-dispersion parameter can be estimated from the residuals

$$\widehat{\phi} = \frac{1}{N-p} \sum_{i=1}^{N} \frac{[Y_i - \widehat{\mu}_i]^2}{v_i(\widehat{\mu}_i)}$$

General Estimating Equations (GEE)

- for independent observations $Y_1, \dots, Y_N$ (within the GLM framework) the corresponding score equations for estimating $\beta \in \mathbb{R}^p$ are

$$\mathcal{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v_i^{-1}[Y_i - \mu_i] = \mathbf{0}$$

- for longitudinal observations $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ (within the GEE framework) the score equations for $\beta \in \mathbb{R}^p$ can be seen as multivariate extensions

$$\mathcal{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} v_{ij}^{-1}[Y_{ij} - \mu_{ij}] = \mathbf{0}$$

- which can be also expressed in a more common (as a sum of independent subjects) matrix notation

$$\mathcal{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbb{D}_i^\top [\mathbb{V}_i(\boldsymbol{\alpha})]^{-1}[\boldsymbol{Y}_i - \boldsymbol{\mu}_i] = \mathbf{0}$$

where $\mathbb{D}_i = \left(\frac{\partial \mu_{ij}}{\partial \beta_k}\right)_{j,k=1}^{n_i,p}$ and $\mathbb{V}_i(\boldsymbol{\alpha}) \equiv \mathbb{V}_i(\mathbb{X}_i, \boldsymbol{\beta}, \phi, \boldsymbol{\alpha})$

Correlation structure within $\boldsymbol{Y}_i$

- Note, that the variance matrix $Var[\boldsymbol{Y}_i|\mathbb{X}_i]$ is relatively complex and specific structural decomposition is typically used to model the variance-covariance structure more carefully

$$Var[\boldsymbol{Y}_i|\mathbb{X}_i] = \mathbb{V}_i(\mathbb{X}_i, \boldsymbol{\beta}, \phi, \boldsymbol{\alpha}) = \phi \mathbb{A}_i^{1/2}(\boldsymbol{\beta})\mathbb{R}_i(\boldsymbol{\alpha})\mathbb{A}_i^{1/2}(\boldsymbol{\beta})$$

where the matrix $\mathbb{A}_i^{1/2}(\boldsymbol{\beta})$ models the covariance of the repeated observations for the given subject $i \in \{1, \dots, N\}$

$$\mathbb{A}_i^{1/2}(\boldsymbol{\beta}) = \begin{pmatrix} \sqrt{v_{i1}(\mu_{i1})} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{v_{in_i}(\mu_{in_i})} \end{pmatrix}$$

and the correlation of the repeated observations is modeled by $\mathbb{R}_i(\boldsymbol{\alpha})$

Correlation structure within $\boldsymbol{Y}_i$

- Note, that the variance matrix $Var[\boldsymbol{Y}_i|\mathbb{X}_i]$ is relatively complex and specific structural decomposition is typically used to model the variance-covariance structure more carefully

$$Var[\boldsymbol{Y}_i|\mathbb{X}_i] = \mathbb{V}_i(\mathbb{X}_i, \boldsymbol{\beta}, \phi, \boldsymbol{\alpha}) = \phi\mathbb{A}_i^{1/2}(\boldsymbol{\beta})\mathbb{R}_i(\boldsymbol{\alpha})\mathbb{A}_i^{1/2}(\boldsymbol{\beta})$$

  where the matrix $\mathbb{A}_i^{1/2}(\boldsymbol{\beta})$ models the covariance of the repeated observations for the given subject $i \in \{1, \dots, N\}$

$$\mathbb{A}_i^{1/2}(\boldsymbol{\beta}) = \begin{pmatrix} \sqrt{v_{i1}(\mu_{i1})} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{v_{in_i}(\mu_{in_i})} \end{pmatrix}$$

  and the correlation of the repeated observations is modeled by $\mathbb{R}_i(\boldsymbol{\alpha})$

- Recall, that the covariances and variances follow from the mean structure... for modeling purposes we specify the mean structure and the correlations (i.e., the working correlation matrix)

Statistical properties and inference

- the GEE estimates of β are consistent even if the working correlation matrix is incorrect

$$\sqrt{N}(\widehat{\beta}_N - \beta) \underset{as.}{\sim} N_p(\mathbf{0}, \mathbb{I}_0^{-1}\mathbb{I}_1\mathbb{I}_0^{-1})$$

where $\mathbb{I}_0$ is the limit matrix of $\sum_{i=1}^N \mathbb{D}_i^\top [\mathbb{V}_i(\beta)]^{-1}\mathbb{D}_i$ and, analogously also $\mathbb{I}_1$ is the limit matrix of $\sum_{i=1}^N \mathbb{D}_i^\top [\mathbb{V}_i(\beta)]^{-1} Var(\boldsymbol{Y}_i)[\mathbb{V}_i(\beta)]^{-1}\mathbb{D}_i$

- note, that if the variance matrix $Var(\boldsymbol{Y}_i)$ is correctly specified, the asymptotic variance only reduces to $\mathbb{I}_0^{-1}$ (the likelihood variance)
- an estimate for $\mathbb{I}_1$ can be obtained by replacing $Var(\boldsymbol{Y}_i)$ by $(\boldsymbol{Y}_i - \widehat{\mu}_i)(\boldsymbol{Y}_i - \widehat{\mu}_i)^\top$ which usually leads to a good estimate of $\mathbb{I}_1$ even if it is a bad estimate for $Var(\boldsymbol{Y}_i)$

Alternatives and extensions

- **Prentice's two sets of GEE**
  (two sets of GEE equations—one to obtain the estimates for β and the other one to get the estimates for α)

- **GEE based on linearization**
  (linearization in a form of $Y_{ij} = \mu_{ij} + \varepsilon_{ij}$, where $\varepsilon_{ij} = \mu_{ij}$ with probability $1 - \mu_{ij}$ and $\varepsilon_{ij} = 1 - \mu_{ij}$ with probability μ_{ij})

- **GEE2 up to GGE*k* generalizations**
  (extended marginal mean structure considering the first and second (possibly up to k^{th} order) marginals and pairwise associations)

- **Alternating Logistic Regression (ALR)**
  (parameters β and α are estimated in two separate alternating regression formulations that are iterated until convergence)

Alternating logistic regression

- when the response variable only takes two possible values $Y_{ij} \in \{0,1\}$ (i.e., logistic regression) the mean and the correlation structure can be estimated using two alternating regression models
- first order marginals are used to model the conditional mean structure using the equation $logit(E[Y_{ij}|\boldsymbol{X}_{ij}]) = \boldsymbol{X}_{ij}^\top \beta$ and the marginal odds ratios are used to model the associations

$$logit\ P[Y_{ij} = 1 | Y_{ik} = y_{ik}] = \alpha_{ijk} y_{ik} + \ln \left(\frac{P[Y_{ij} = 1, Y_{ik} = 1] P[Y_{ij} = 0, Y_{ik} = 0]}{P[Y_{ij} = 0, Y_{ik} = 1] P[Y_{ij} = 1, Y_{ik} = 0]} \right)$$

  where $\alpha_{ijk} y_{ik} \in \mathbb{R}$ is modeled as a predictor variable and some unknown parameter and the odds ratio is an offset parameter (intercept)
- the alternating logistic regression is (almost) as efficient as GEE2 and (almost) as computationally easy as GGE

Summary

- Marginal models for correlated observations are specifically suitable for population interpretation and population based inference
- Different strategies can be used to build the model using the marginals of the joint distribution $P[\boldsymbol{Y}_i = \boldsymbol{y} | \mathbb{X}_i]$
- Different models imply different interpretation of the estimated parameters and also different limitations for a practical utilization
- For a subject specific interpretation another models need to be used – for instance, models with random effects