Missing Data: Theory & Methods

Chapter 3
MLE in Regression Models with Missing Covariates

Lu Mao lmao@biostat.wisc.edu

Contents

Generalized Linear Models

- In this chapter, we extend the problem of regression with missing covariates from the simple linear model considered in §2.4. to the generalized linear models (GLM) and longitudinal models.
- First we review some essential background of the GLM.
- The GLM generalizes the linear regression model to any model in which the response variable is in the exponential family (§2.3), such as binomial, Poisson, Gamma, Geometric, inverse Gaussian, and so on.
- We first consider the setting without covariates, that is, we consider a single outcome Y whose distribution belongs to the exponential family.

Generalized Linear Models

- To facilitate model building in the regression setting, we re-formulate the exponential family as follows.

  Definition 3.1 (Exponential Family)

  The distribution of Y belongs to an exponential family with canonical parameter γ and scale parameter ϕ, denoted as $Y \sim \mathcal{E}(\gamma, \phi)$, if its density can be written as

$$p_Y(y;\theta) = \exp\Big\{\phi[\gamma^{\mathrm{T}} y - a(\gamma)] - c(y,\phi)\Big\},$$

  where $\theta = (\gamma^{\mathrm{T}}, \phi)^{\mathrm{T}}$, and $a(\cdot)$ and $c(\cdot)$ are appropriate functions that make $p_Y(\cdot;\theta)$ a proper density function.

- In this formulation, the systematic part of Y is related only to γ, and ϕ only controls the dispersion Y.

Generalized Linear Models

- This is made precise in the following Proposition.

  Proposition 3.1 (Mean and Variance of $\mathcal{E}(\gamma, \phi)$)
  If $Y \sim \mathcal{E}(\gamma, \phi)$, *then*

$$EY = \dot{a}(\gamma), \quad \text{and } \mathrm{Var}Y = \phi^{-1}\ddot{a}(\gamma)$$

- One can show that the function $a(\cdot)$ is infinite times differentiable and is strict convex. So $\ddot{a}(\gamma) > 0$.

Generalized Linear Models

Proof. of Proposition 3.1 By $E\frac{\partial}{\partial\gamma}\log p_Y(Y;\theta) = 0$, we have

$$\phi E[Y - \dot{a}(\gamma)] = 0.$$

Thus, $EY = \dot{a}(\gamma)$. By

$$-E\left[\frac{\partial^2}{\partial\gamma^{\otimes 2}}\log p_Y(Y;\theta)\right] = E\left[\frac{\partial}{\partial\gamma}\log p_Y(Y;\theta)\right]^{\otimes 2},$$

we have

$$\phi\ddot{a}(\gamma) = \phi^2 E[Y - \dot{a}(\gamma)]^{\otimes 2}.$$

Hence, Var$Y = \phi^{-1}\ddot{a}(\gamma)$. □

Generalized Linear Models

Examples of Exponential Family

- **Normal**: $Y \sim N(\mu, \sigma^2)$, we have

$$\phi = \sigma^{-2}, \quad \gamma = \mu, \quad a(\gamma) = \frac{\gamma^2}{2}, \quad c(\phi, y) = \frac{\phi y^2 - \log(2\pi\phi^{-1})}{2}.$$

  So,

$$EY = \dot{a}(\gamma) = \gamma = \mu, \quad \mathsf{Var}Y = \phi^{-1}\ddot{a}(\gamma) = \phi^{-1} = \sigma^2.$$

- **Binomial**: $Y \sim \mathsf{Bin}(1, \pi)$, we have

$$\phi = 1, \quad \gamma = \log\left(\frac{\pi}{1-\pi}\right), \quad a(\gamma) = \log\left(1 + e^{\gamma}\right), \quad c(\phi, y) = 0.$$

  So,

$$EY = \dot{a}(\gamma) = \gamma = \mu, \quad \mathsf{Var}Y = \phi^{-1}\ddot{a}(\gamma) = \phi^{-1} = \sigma^2.$$

Generalized Linear Models

Examples of Exponential Family

- **Poisson**: $Y \sim \mathsf{Poisson}(\lambda)$, we have

$$\phi = 1, \quad \gamma = \log\lambda, \quad a(\gamma) = e^{\gamma}, \quad c(\phi, y) = \log y!.$$

  So,

$$EY = \dot{a}(\gamma) = e^{\gamma} = \lambda, \quad \mathsf{Var}Y = \phi^{-1}\ddot{a}(\gamma) = e^{\gamma} = \lambda.$$

- **Gamma**: $Y \sim \mathsf{Gamma}(\mu, \nu)$, that is,

$$p_Y(y;\theta) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp\left(-\frac{\mu}{\nu}y\right),$$

  we have

$$\phi = \nu, \gamma = -\mu^{-1}, a(\gamma) = -\log(-\gamma), c(\phi, y) = (1-\phi)\log y + \log\Gamma(\nu) - \nu\log\nu.$$

  So,

$$EY = \dot{a}(\gamma) = -\gamma^{-1} = \mu, \quad \mathsf{Var}Y = \phi^{-1}\ddot{a}(\gamma) = \phi^{-1}\gamma^{-2} = \nu^{-1}\mu^2.$$

Generalized Linear Models

- In the regression setting, suppose we have iid $(Y_i, Z_i), i = 1, \cdots, n$, where Y_i is the response and Z_i is a vector of covariates.
- The generalized linear models (GLM; McCullagh and Nelder, 1989) involve first specifying the conditional distribution of Y_i given Z_i as a member of the exponential family and then choosing a link function to relate the canonical parameter γ to the covariates.
- Specifically, define $\mu_i := E(Y_i|Z_i)$, then one models μ_i as a certain *function* of a *linear* combination of the components of Z_i.
- Because by Proposition , $\gamma_i = \dot{a}^{-1}(\mu_i)$, the mean model can be equivalently expressed as a model for the canonical parameter.

Generalized Linear Models

Definition 3.2 (Generalized Linear Models)

Generalized linear models are defined as follows:

(a) *[Distributional assumption] The conditional density of Y_i given Z_i is $\mathcal{E}(\gamma_i, \phi)$, which is a member of the exponential family (1.1.1), where γ_i is a function of Z_i and ϕ is a scale parameter shared by all subjects;*

(b) *[Structural assumption] The conditional mean μ_i is related to Z_i by*

$$\mu_i = \mu(\beta^{\mathrm{T}} Z_i),$$

where β is a regression parameter and $\mu(\cdot)$ is a known monotonic link function.

Generalized Linear Models

- To express the canonical parameter as a function of the linear predictor $\beta^{\mathrm{T}} Z_i$, we have

$$\gamma_i = \dot{a}^{-1}(\mu_i) = \dot{a}^{-1} \circ \mu(\beta^{\mathrm{T}} Z_i) =: \gamma(\beta^{\mathrm{T}} Z_i).$$

- Clearly, the choice of link function $\mu(x) = \dot{a}(x)$ leads to $\gamma(x) = x$, making $\gamma_i = \beta^{\mathrm{T}} Z_i$. Hence, $\mu(x) = \dot{a}(x)$ is called the **canonical link**.
- The canonical link gives rise to simple forms of conditional density functions, and, in many common situations, leads to very interpretable regression parameters.
- The familiar classical normal regression, logistic regression with binary response, and log-linear models for Poisson count data are all examples of GLM with canonical links.

Generalized Linear Models

Examples of GLM with canonical links

- **Normal regression**: $\dot{a}(x) = x$. So under the canonical link,

$$\mu_i = \beta^{\mathrm{T}} Z_i.$$

- **Logistic regression**: $\dot{a}(x) = \frac{e^x}{1+e^x}$. So under the canonical link,

$$\mu_i = \frac{e^{\beta^{\mathrm{T}} Z_i}}{1 + e^{\beta^{\mathrm{T}} Z_i}}.$$

- **Poisson regression**: $\dot{a}(x) = e^x$. So under the canonical link,

$$\mu_i = e^{\beta^{\mathrm{T}} Z_i}.$$

Generalized Linear Models

- We review the MLE for (β, ϕ) under the setting of fully observed covariates Z_i.
- Note that

$$l(Y_i, Z_i; \theta) = \phi[Y_i \gamma(\beta^{\mathrm{T}} Z_i) - a \circ \gamma(\beta^{\mathrm{T}} Z_i)] - c(\phi, Y_i).$$

- For estimation of β, we have

$$\begin{aligned} \dot{l}_\beta(Y_i, Z_i; \theta) &= \phi[Y_i - \dot{a} \circ \gamma(\beta^{\mathrm{T}} Z_i)]\dot{\gamma}(\beta^{\mathrm{T}} Z_i) Z_i \\ &= \phi[Y_i - \mu(\beta^{\mathrm{T}} Z_i)]\dot{\gamma}(\beta^{\mathrm{T}} Z_i) Z_i. \end{aligned} \tag{3.1}$$

- Clearly, the solution for β does not depend on ϕ.
- As people usually work with the mean model $\mu(\cdot)$, it is worthwhile to explore the relationship between $\mu(\cdot)$ and $\gamma(\cdot)$ to have an alternative expression of (3.1).

Generalized Linear Models

- Since $\mu = \dot{a} \circ \gamma$, we have

$$\dot{\mu}(x) = \ddot{a} \circ \gamma(x) \dot{\gamma}(x).$$

- Hence, (3.1) can be re-written as

$$\begin{aligned} \dot{l}_\beta(Y_i, Z_i; \theta) &= \phi[Y_i - \mu(\beta^{\mathrm{T}} Z_i)] \frac{\dot{\mu}(\beta^{\mathrm{T}} Z_i)}{\ddot{a} \circ \gamma(\beta^{\mathrm{T}} Z_i)} Z_i \\ &= \frac{Y_i - \mu(\beta^{\mathrm{T}} Z_i)}{\phi^{-1} \ddot{a} \circ \gamma(\beta^{\mathrm{T}} Z_i)} \dot{\mu}(\beta^{\mathrm{T}} Z_i) Z_i. \\ &= \frac{Y_i - \mu(\beta^{\mathrm{T}} Z_i)}{\mathrm{Var}(Y_i | Z_i; \theta)} \dot{\mu}(\beta^{\mathrm{T}} Z_i) Z_i, \end{aligned}$$

  where the last equality follows from Proposition 3.1.

Generalized Linear Models

- Further, from (3.1), we have

$$-\ddot{l}_\beta(Y_i, Z_i; \theta) = \phi\dot{\mu}(\beta^{\mathrm{T}} Z_i)\dot{\gamma}(\beta^{\mathrm{T}} Z_i) Z_i^{\otimes 2} - \phi[Y_i - \mu(\beta^{\mathrm{T}} Z_i)]\ddot{\gamma}(\beta^{\mathrm{T}} Z_i) Z_i^{\otimes 2}.$$

- The second term on the right hand side of the above equation is negligible because it has expectation zero (why?).
- Thus, we can work

$$\begin{aligned} -\ddot{l}^*_\beta(Y_i, Z_i; \theta) &:= \phi\dot{\mu}(\beta^{\mathrm{T}} Z_i)\dot{\gamma}(\beta^{\mathrm{T}} Z_i) Z_i^{\otimes 2} \\ &= \frac{\dot{\mu}(\beta^{\mathrm{T}} Z_i)^2}{\mathsf{Var}(Y_i | Z_i; \theta)} Z_i^{\otimes 2}. \end{aligned} \tag{3.2}$$

  Hence, the $(t+1)$th updated in the Newton-Raphson algorithm is given by

$$\beta^{(t+1)} = \beta^{(t)} - \left\{ \sum_{i=1}^n \ddot{l}^*_\beta(Y_i, Z_i; \theta^{(t)}) \right\}^{-1} \sum_{i=1}^n \dot{l}_\beta(Y_i, Z_i; \theta^{(t)}).$$

Generalized Linear Models

- After obtaining $\widehat{\beta}$, we can calculate the maximum likelihood estimator $\widehat{\phi}$ of the dispersion parameter ϕ.
- After simple calculations, we have

$$\dot{l}_\phi(Y_i, Z_i; \widehat{\beta}, \phi) = Y_i \gamma(\widehat{\beta}^{\mathrm{T}} Z_i) - a \circ \gamma(\widehat{\beta}^{\mathrm{T}} Z_i) - \dot{c}(\phi, Y_i),$$

  and

$$-\ddot{l}_\phi(Y_i, Z_i; \widehat{\beta}, \phi) = \ddot{c}(\phi, Y_i).$$

- The Newton-Raphson algorithm can be carried out accordingly.
- Note that β and ϕ are orthogonal parameters in the sense that

$$E\dot{l}_\beta(Y_i, Z_i; \beta, \phi)\dot{l}_\phi(Y_i, Z_i; \beta, \phi) = 0.$$

- Thus, the variance of $\widehat{\beta}$ can be estimated by

$$-\left\{ \sum_{i=1}^n \ddot{l}^*_\beta(Y_i, Z_i; \widehat{\theta}) \right\}^{-1}.$$

Generalized Linear Models

Examples of GLM with canonical links

- For GLMs with canonical links, we have

$$\dot{l}_\beta(Y_i, Z_i; \theta) = \phi[Y_i - \mu(\beta^{\mathrm{T}} Z_i)]Z_i,$$

and

$$-\ddot{l}^*_\beta(Y_i, Z_i; \theta) = \phi\dot{\mu}(\beta^{\mathrm{T}} Z_i)Z_i^{\otimes 2}.$$

- **Normal regression:**

$$\dot{l}_\beta(Y_i, Z_i; \theta) = \sigma^{-2}(Y_i - \beta^{\mathrm{T}} Z_i)Z_i,$$

and

$$-\ddot{l}^*_\beta(Y_i, Z_i; \theta) = \sigma^{-2} Z_i^{\otimes 2}.$$

Generalized Linear Models

Examples of GLM with canonical links

- **Logistic regression:**

$$\dot{l}_\beta(Y_i, Z_i; \theta) = \left(Y_i - \frac{e^{\beta^{\mathrm{T}} Z_i}}{1 + e^{\beta^{\mathrm{T}} Z_i}}\right) Z_i,$$

and

$$-\ddot{l}^*_\beta(Y_i, Z_i; \theta) = \frac{e^{\beta^{\mathrm{T}} Z_i}}{(1 + e^{\beta^{\mathrm{T}} Z_i})^2} Z_i^{\otimes 2}.$$

- **Poisson regression:**

$$\dot{l}_\beta(Y_i, Z_i; \theta) = \left(Y_i - e^{\beta^{\mathrm{T}} Z_i}\right) Z_i,$$

and

$$-\ddot{l}^*_\beta(Y_i, Z_i; \theta) = e^{\beta^{\mathrm{T}} Z_i} Z_i^{\otimes 2}.$$

Contents

EM by the Method of Weights

History of using EM for GLM with missing covariates

- The EM algorithm has been a popular technique for obtaining MLEs in GLM's with missing covariate data.
- Fuchs (1982, JASA) uses the EM algorithm to get MLE for log-linear models with incomplete data. Little and Schluchter (1985, Biometrika) use the EM algorithm to obtain estimates in a regression model with ignorably missing categorical and continuous covariates. Schluchter and Jackson (1989, JASA) use EM to find parameter estimates in log-linear models with ignorable missing data.
- A general method for estimation in the presence of missing covariates has been proposed by Ibrahim (1990), who uses the EM by the **method of weights** to find the MLEs.
- The method of weights can be used for categorical and continuous covariates, and for ignorably and non-ignorably missing data.

EM by the Method of Weights

GLM with categorical MAR covariates

- We first consider GLM with categorical covariates that are assumed to be MAR.
- Let η denote the density function of Z. With fully observed covariates Z, the joint density of (Y, Z) can be factorized as

$$p(Y, Z; \theta, \eta) = p(Y|Z; \theta)\eta(Z).$$

  So the inference of θ has nothing to do with η.
- With $M(Z)$, which is a coarsened version of Z, the joint density of $(Y, M(Z))$ under MAR is a multiple of

$$p(Y, M(Z); \theta, \eta) = \int_{M(z)=M(Z)} p(Y|z; \theta)\eta(z)d\nu(z),$$

  in which η no longer factorizes with the conditional density containing θ.

EM by the Method of Weights

GLM with categorical MAR covariates

- So, in order to make inference on θ, it is necessary that we have a model for η, say, $p(Z; \alpha)$, where α is a Euclidean parameter.
- Hence the parameters consists of $\theta = (\beta^{\mathrm{T}}, \phi, \alpha^{\mathrm{T}})^{\mathrm{T}}$.
- The full data are (Y_i, Z_i), $i = 1, \cdots, n$, where the Z_i take values in $\mathcal{Z} := \{z_1, \cdots, z_m\}$, and the observed data are $(Y_i, M_i, M_i(Z_i)), i = 1, \cdots, n$.
- The full data log-likelihood is

$$\sum_{i=1}^{n} \Big\{ l(Y_i, Z_i; \beta, \phi) + l_\alpha(Z_i; \alpha) \Big\},$$

  where

$$l(Y_i, Z_i; \beta, \phi) = \phi[Y_i\gamma(\beta^{\mathrm{T}} Z_i) - a \circ \gamma(\beta^{\mathrm{T}} Z_i)] - c(\phi, Y_i),$$

  as in §3.1 and $l_\alpha(Z_i; \alpha) = \log p(Z_i; \alpha)$.

EM by the Method of Weights

GLM with categorical MAR covariates

- Therefore, at the $(j+1)$th iteration, the Q function takes the form

$$\begin{aligned} Q(\theta|\theta^{(j)}) &= \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} l(Y_i, z_k; \beta^{(j)}, \phi^{(j)}) + \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} l_\alpha(z_k; \alpha^{(j)}) \\ &= Q_1(\beta, \phi|\theta^{(j)}) + Q_2(\alpha|\theta^{(j)}), \end{aligned} \tag{3.3}$$

where $\mathcal{C}_i = \{z \in \mathcal{Z} : M_i(z) = M_i(Z_i)\}$ and

$$\begin{aligned} w_{ik}^{(j)} &= E[I(Z_i = z_k)|Y_i, M_i, M_i(Z_i); \theta^{(j)}] \\ &= E[I(Z_i = z_k)|Y_i, M_i(Z_i); \theta^{(j)}] \qquad \text{(MAR)} \\ &= \frac{p(Y_i|z_k; \beta^{(j)}, \phi^{(j)}) p(z_k; \alpha^{(j)})}{\sum_{z \in \mathcal{C}_i} p(Y_i|z; \beta^{(j)}, \phi^{(j)}) p(z; \alpha^{(j)})}. \end{aligned} \tag{3.4}$$

EM by the Method of Weights

GLM with categorical MAR covariates

- The M step thus involves separate maximizations regarding (β, ϕ) and α, respectively.
- Both maximizations are equivalent to doing full-data MLE with each incomplete observation replaced by a set of weighted observations.
- Thus, the implementation of the EM by the method of weights for missing categorical covariates is quite straightforward in SAS or R.

EM by the Method of Weights

GLM with categorical MAR covariates

- Specifically, the maximization of $Q_1(\beta, \phi|\theta^{(j)})$ can be done with respect to β and ϕ separately similar to the full-data scenario.
- By the results from §3.1,

$$\begin{aligned}\dot{Q}_{1\beta}(\beta, \phi|\theta^{(j)}) &:= \frac{\partial}{\partial\beta} Q_1(\beta, \phi|\theta^{(j)}) \\ &= \phi \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} [Y_i - \mu(\beta^{\mathrm{T}} z_k)] \dot{\gamma}(\beta^{\mathrm{T}} z_k) z_k \\ &= \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} \frac{Y_i - \mu(\beta^{\mathrm{T}} z_k)}{\mathsf{Var}(Y|Z = z_k; \theta)} \dot{\mu}(\beta^{\mathrm{T}} z_k) z_k.\end{aligned}$$

EM by the Method of Weights

GLM with categorical MAR covariates

- Similarly, by ignoring the zero-mean term, $-\ddot{Q}_{1\beta}(\beta, \phi|\theta^{(j)}) := -\frac{\partial}{\partial\beta}\dot{Q}_{1\beta}(\beta, \phi|\theta^{(j)})$ can be approximated by

$$\begin{aligned}-\ddot{Q}^{*}_{1\beta}(\beta, \phi|\theta^{(j)}) &= \phi \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} \dot{\mu}(\beta^{\mathrm{T}} z_k) \dot{\gamma}(\beta^{\mathrm{T}} z_k) z_k^{\otimes 2} \\ &= \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} \frac{\dot{\mu}(\beta^{\mathrm{T}} z_k)^2}{\mathsf{Var}(Y|Z = z_k; \theta)} z_k^{\otimes 2}.\end{aligned}$$

- Thus, the Newton-Raphson algorithm for the M step of β, which does not involve ϕ, can be easily constructed.

EM by the Method of Weights

GLM with categorical MAR covariates

- After obtaining $\beta^{(j+1)}$, $\phi^{(j+1)}$ can be calculated using the Newton-Raphson algorithm on

$$\begin{aligned}\dot{Q}_{1\phi}(\beta^{(j+1)}, \phi|\theta^{(j)}) &:= \frac{\partial}{\partial\phi} Q_1(\beta^{(j+1)}, \phi|\theta^{(j)}) \\ &= \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} [Y_i \gamma(\beta^{(j+1)^{\mathrm{T}}} z_k) - a \circ \gamma(\beta^{(j+1)^{\mathrm{T}}} z_k)] \\ &\quad - \sum_{i=1}^{n} \dot{c}(\phi, Y_i),\end{aligned}$$

with

$$-\ddot{Q}_{1\phi}(\beta^{(j+1)}, \phi|\theta^{(j)}) := -\frac{\partial}{\partial\phi} \dot{Q}_{1\phi}(\beta^{(j+1)}, \phi|\theta^{(j)}) = \sum_{i=1}^{n} \ddot{c}(\phi, Y_i).$$

EM by the Method of Weights

GLM with categorical MAR covariates

- The M step for α, as usual, solves the weighted estimating equation

$$\sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} \dot{l}_\alpha(z_k; \alpha) = 0.$$

- Variance estimator for the MLE $\widehat{\theta}$ can be easily derived using the Louis formula since all conditional expectations have closed forms.

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM

- When the covariates are continuous rather than categorical, the E step consists of an integral rather than a weighted sum.
- Assume without loss of generality that the covariate Z_i can be represented as $(M_i(Z_i), \widetilde{Z}_i)$ (This can always be achieved by re-arranging the components or doing an invertible transformation of Z_i).
- So the Q function can be written as

$$Q(\theta|\theta^{(j)}) = \sum_{i=1}^{n} \int l\Big(Y_i, (M_i(Z_i), \widetilde{z}_i); \beta, \phi\Big) \times p(\widetilde{z}_i|Y_i, M_i(Z_i); \theta^{(j)}) d\widetilde{z}_i + \sum_{i=1}^{n} \int l_\alpha\Big(M_i(Z_i), \widetilde{z}_i; \alpha) p(\widetilde{z}_i|Y_i, M_i(Z_i); \theta^{(j)}) d\widetilde{z}_i. \quad (3.5)$$

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM

- This integral generally does not have a closed form. Ibrahim and Weisberg (1992, Australian Journal) consider approximating (3.5) using Gaussian quadrature, which discretizes the integral into a weighted complete data form so that the EM by the method of weights can be used to find parameter estimates.
- However, Gaussian quadrature may not perform well in small samples, for large missing data fractions, or when the distribution of the covariates is not approximately normal.
- Alternatively, one can evaluate (3.5) using the Monte Carlo EM algorithm in conjunction with the Gibbs sampler.

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM

- Since the integrals in (3.5) involve the conditional expectation of the form

$$E\Big[g\Big(Y_i, M_i(Z_i), \widetilde{Z}_i\Big)\Big|Y_i, M_i(Z_i); \theta^{(j)}\Big], \tag{3.6}$$

  we are essentially dealing with the conditional distribution of $\widetilde{Z}_i$ given Y_i and $M_i(Z_i)$. Denote this conditional distribution as

$$[\widetilde{z}_i|Y_i, M_i(Z_i); \theta^{(j)}]. \tag{3.7}$$

- If we can simulate a large number of observations $\widetilde{z}_{i1}^{(j)}, \cdots, \widetilde{z}_{iK_i}^{(j)}$ from (3.7), then, by the law of large numbers, the conditional expectation (3.6) can be approximated by

$$K_i^{-1} \sum_{k=1}^{K_i} g\Big(Y_i, M_i(Z_i), \widetilde{z}_{ik}^{(j)}\Big).$$

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM

- Using this Monte-Carlo integration, the Q function at the $(j+1)$th iteration can be approximated by

$$Q(\theta|\theta^{(j)}) = \sum_{i=1}^{n} K_i^{-1} \sum_{k=1}^{K_i} \Big\{ l\big(Y_i, (M_i(Z_i), \widetilde{z}_{ik}^{(j)}); \beta, \phi\big) + l_\alpha\big(M_i(Z_i), \widetilde{z}_{ik}^{(j)}; \alpha\big) \Big\}.$$

- The Q function takes the same form as that for categorical covariates. In this case, each full-data log-likelihood function is weighted by K_i^{-1}. So, the method of weights readily applies here.
- There is only one question remaining - how to sample from the conditional distribution (3.7).

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM Gibbs Sampling

- Suppose we have a collection of J random variables denoted by $U = (U_1, \cdots, U_J)$.
- We assume that the full conditional distributions

$$[U_j | (U_{j'} : j' \neq j, j' = 1, \cdots, J)], j = 1, \cdots, J,$$

  are available for sampling. Here, "available" means that samples may be generated by some method.
- Under mild conditions (see Besag, 1974), the one-dimensional conditional distributions uniquely determine the full joint distribution $[U_1, \cdots, U_J]$, and hence all marginal distributions $[U_j], j = 1, \cdots, J$.

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM Gibbs Sampling

- The Gibbs sampler proceeds as follows

(0) Suppose we have a set of arbitrary starting values $(U_1^{(0)}, U_2^{(0)} \cdots, U_J^{(0)})$;
(1) Draw $U_1^{(1)}$ from $[U_1 | U_2^{(0)}, U_3^{(0)} \cdots, U_J^{(0)}]$;
(2) Draw $U_2^{(1)}$ from $[U_2 | U_1^{(1)}, U_3^{(0)} \cdots, U_J^{(0)}]$;
    ...
(J) Draw $U_J^{(1)}$ from $[U_J | U_1^{(1)}, U_2^{(1)} \cdots, U_{J-1}^{(1)}]$.

This completes one iteration of Gibbs Sampling.

- After t such iterations, one obtains $(U_1^{(t)}, U_2^{(t)} \cdots, U_J^{(t)})$.

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM Gibbs Sampling

- For the Gibbs sampling algorithm outlined above

(a) $(U_1^{(t)}, U_2^{(t)} \cdots, U_J^{(t)}) \to_d [U_1, U_2 \cdots, U_J]$

(b) The convergence in (a) is exponential in t.

- Sampling from the one-dimensional conditional distributions can be done using some acceptance-rejection method such as the adaptive rejection sampling (Gilks and Wild, 1992), which generally requires knowing the kernel of the conditional densities.
- Hence if we can find an analytic function $f(u_1, \cdots, u_J)$ to which the joint density of $[U_1, U_2 \cdots, U_J]$ is proportional, then every conditional density is proportional to f.
- Then, we can use the Gibbs sampler to sample from the joint distribution where each one-dimensional draw is completed by the adaptive rejection algorithm.

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM Gibbs Sampling

- Back to our E step without continuous covariates, note that the conditional density of $\widetilde{Z}_i$ given Y_i, $M_i(Z_i)$, and $\theta^{(j)}$, is proportional to

$$p(Y_i|M_i(Z_i), \widetilde{z}_i; \beta^{(j)}, \phi^{(j)}) p(M_i(Z_i), \widetilde{z}_i; \alpha^{(j)}).$$

- So, we can use this kernel and the Gibbs sampler to draw $\widetilde{z}_{i1}^{(j)}, \cdots, \widetilde{z}_{iK_i}^{(j)}$ from the conditional distribution (3.7).
- More about Gibbs sampling and rejection algorithms is to be covered in later chapters.

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM
Specifying the covariate model

- An important issue with missing covariate data is the specification of a parametric model for the missing covariates.
- In missing data problems, one should consider strategies for reducing the number of nuisance parameters in the covariate distribution.
- One such strategy is to model the joint distribution of the covariates as a product of one dimensional conditional distributions (Lipsitz and Ibrahim, 1996; Ibrahim et al., 1999).

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM
Specifying the covariate model

- Suppose that we write the joint distribution of the p-dimensional covariate vector $(z_1, \cdots, z_p)$ as

$$p(z_1, \cdots, z_p; \alpha) = p(z_p|z_1, \cdots, z_{p-1}; \alpha_p)p(z_{p-1}|z_1, \cdots, z_{p-2}; \alpha_{p-1}) \times \cdots p(z_1; \alpha_1), \quad (3.8)$$

  where α_j is a vector of parameters for the jth conditional distribution and $\alpha = (\alpha_1^{\mathrm{T}}, \cdots, \alpha_p^{\mathrm{T}})^{\mathrm{T}}$.
- It is important to note that a model needs to be specified only for the covariates that are not completely observed. When one or more covariates are completely observed for all n observations, then those covariates can be conditioned upon when constructing the distribution of the missing covariates.

EM by the Method of Weights

GLM with continuous MAR covariates using Monte Carlo EM
Specifying the covariate model

- For example, suppose we have (z_1, z_2, z_3) and $((z_1, z_2)$ are missing for some subjects and z_3 is observed for all subjects. Then, we can take as our covariate distribution

$$p(z_1, z_2|z_3) = p(z_1|z_2, z_3)p(z_2|z_3).$$

- The specification of covariate distribution in (3.8) has a number of desirable features, including easing the computational burden in the Gibbs sampling algorithm required for sampling from (3.7).
- The EM by the method of weights for MAR data is implemented by the commercially available software XMISS, which is commonly used to analyze clinical trials with missing data.

EM by the Method of Weights

GLM with nonignorably missing covariates

- The EM by the method of weights can also be easily adapted to nonignorably missing covariates.
- If the missing covariates are nonignorable, it is typical to specify a parametric model for the missing data mechanism, i.e., a selection model, and incorporate it into the full-data log-likelihood.
- If missingness involves only missing components of covariates (not other types of coarsening), the missingness mechanism can be defined as the distribution of p random vector R_i, whose kth component, R_{ik}, equals 1 if Z_{ik} is observed for subject i, and 0 if Z_{ik} is missing.
- The conditional distribution of R_i given (Y_i, Z_i), denoted $p(R_i|Y_i, Z_i; \psi)$, is indexed by the parameter vector ψ, and is a multinomial distribution with 2^p cell probabilities.

EM by the Method of Weights

GLM with nonignorably missing covariates

- The full-data density of (R_i, Y_i, Z_i) for subject i is then given by

$$p(Y_i, Z_i, R_i|\beta, \phi, \alpha, \psi) = p(Y_i|Z_i; \beta, \phi)p(Z_i; \alpha)p(R_i|Y_i, Z_i; \psi),$$

  which leads to the full-data log-likelihood

$$\sum_{i=1}^{n} \Big\{ l(Y_i, Z_i; \beta, \phi) + l_\alpha(Z_i; \alpha) + l_\psi(R_i, Y_i, Z_i; \psi) \Big\}.$$

- The whole parameter consists of $\theta = (\beta^{\mathrm{T}}, \phi, \alpha^{\mathrm{T}}, \psi^{\mathrm{T}})^{\mathrm{T}}$. The main interest is in the estimation of β, with ϕ, α, and ψ viewed as nuisance parameters.
- One can choose any multinomial model for $p(R_i|Y_i, Z_i; \psi)$, provided that the whole model is identifiable. One obvious choice is a log-linear model.

EM by the Method of Weights

GLM with nonignorably missing covariates

- Assuming that the missing covariates are categorical, at the $(j+1)$th iteration

$$\begin{aligned} Q(\theta|\theta^{(j)}) &= \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} l(Y_i, z_k; \beta^{(j)}, \phi^{(j)}) + \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} l_\alpha(z_k; \alpha^{(j)}) \\ &\quad + \sum_{i=1}^{n} \sum_{z_k \in \mathcal{C}_i} w_{ik}^{(j)} l_\psi(R_i, Y_i, z_k; \psi^{(j)}) \\ &=: Q_1(\beta, \phi|\theta^{(j)}) + Q_2(\alpha|\theta^{(j)}) + Q_2(\psi|\theta^{(j)}), \end{aligned}$$

  where, (use $\cdot$ to denote component-wise product)

$$\begin{aligned} w_{ik}^{(j)} &= E[I(Z_i = z_k)|Y_i, R_i, R_i \cdot Z_i; \theta^{(j)}] \\ &= \frac{p(R_i|Y_i, z_k; \psi^{(j)})p(Y_i|z_k; \beta^{(j)}, \phi^{(j)})p(z_k; \alpha^{(j)})}{\sum_{z \in \mathcal{C}_i} p(R_i|Y_i, z; \psi^{(j)})p(Y_i|z; \beta^{(j)}, \phi^{(j)})p(z; \alpha^{(j)})}. \end{aligned}$$

EM by the Method of Weights

GLM with nonignorably missing covariates

- In the continuous covariate case, the E step can be written as

$$\begin{aligned} Q(\theta|\theta^{(j)}) = &\sum_{i=1}^{n} \int l(Y_i, Z_i; \beta, \phi) p(Z_{i,\text{mis}}|Y_i, R_i, Z_{i,\text{obs}}; \theta^{(j)}) dZ_{i,\text{mis}} \\ &+ \sum_{i=1}^{n} \int l_\alpha(Z_i; \alpha) p(Z_{i,\text{mis}}|Y_i, R_i, Z_{i,\text{obs}}; \theta^{(j)}) dZ_{i,\text{mis}} \\ &+ \sum_{i=1}^{n} \int l_\psi(R_i, Y_i, Z_i; \psi) p(Z_{i,\text{mis}}|Y_i, R_i, Z_{i,\text{obs}}; \theta^{(j)}) dZ_{i,\text{mis}}, \end{aligned}$$

  where $Z_{i,\text{mis}}$ is the missing part and $Z_{i,\text{obs}}$ is the observed part of Z_i.
- This time, we have that

$$p(Z_{i,\text{mis}}|Y_i, R_i, Z_{i,\text{obs}}; \theta^{(j)}) \propto p(Y_i|Z_i; \beta^{(j)}, \phi^{(j)}) p(Z_i; \alpha^{(j)}) p(R_i|Y_i, Z_i; \psi^{(j)}).$$

EM by the Method of Weights

GLM with nonignorably missing covariates

- For each observation i, we take a sample $z_{i1}^{(j)}, \cdots, z_{i,K_i}^{(j)}$ from $p(Z_{i,\text{mis}}|Y_i, R_i, Z_{i,\text{obs}}; \theta^{(j)})$ using the Gibbs sampler along with the adaptive rejection algorithm.
- Using Monte Carlo EM, the E step can be written as

$$\begin{aligned} Q(\theta|\theta^{(j)}) = &\sum_{i=1}^{n} K_i^{-1} \sum_{j=1}^{K_i} l(Y_i, Z_{i,\text{obs}}, Z_{i,\text{mis}} = z_{ik}^{(j)}; \beta, \phi) \\ &+ \sum_{i=1}^{n} K_i^{-1} \sum_{j=1}^{K_i} l_\alpha(Z_{i,\text{obs}}, Z_{i,\text{mis}} = z_{ik}^{(j)}; \alpha) \\ &+ \sum_{i=1}^{n} K_i^{-1} \sum_{j=1}^{K_i} l_\psi(R_i, Y_i, Z_{i,\text{obs}}, Z_{i,\text{mis}} = z_{ik}^{(j)}; \psi). \end{aligned}$$

The M step proceeds in a similar fashion by the method of weights.

EM by the Method of Weights

GLM with nonignorably missing covariates

- The ideas in modeling the covariate distribution can be used to specify a model for the missing data mechanism for nonignorably missing covariates. One possible model is a joint log-linear model for $p(R_i|Y_i, Z_i; \psi)$.
- Alternatively, one may model the missing data mechanism by a sequence of one-dimensional conditional distributions to obtain $p(R_i|Y_i, Z_i; \psi)$, leading to

$$\begin{aligned} p(R_{i1}, \cdots, R_{ip}|Y_i, Z_i; \psi) = & \, p(R_{ip}|R_{i1}, \cdots, R_{i,p-1}, Y_i, Z_i; \psi_p) \\ & \times p(R_{i,p-1}|R_{i1}, \cdots, R_{i,p-2}, Y_i, Z_i; \psi_{p-1}) \\ & \times \cdots \times p(R_{i1}|Y_i, Z_i; \psi_1), \end{aligned}$$

where ψ_j is a vector of indexing parameters for the jth conditional distribution and $\psi = (\psi_1^{\mathrm{T}}, \cdots, \psi_p^{\mathrm{T}})^{\mathrm{T}}$.

Contents

A Logistic Regression Example

- Consider the Six-Cities data described in §1.1. To summarize the data
  - $Y = 1, 0$: Wheezing status of child
  - $Z_1 = 1, 0$: City of residence (Kingston-Harriman, TN vs Portage, WI)
  - $Z_2 = 0, 1, 2, \cdots$: Daily number of cigarettes smoked by mother
- Total number of subjects is $n = 390$; Y is fully observed; 17 cases has missing Z_1; 30 cases has missing Z_2; one case misses both Z_1 and Z_2. We assume the data are MAR.
- A logistic regression is used to model the wheezing status

$$\text{Logit}\Pr(Y = 1|Z_1, Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2.$$

- For covariate modeling, we plan to model

$$[Z_1, Z_2] = [Z_2|Z_1] \cdot [Z_1].$$

A Logistic Regression Example

- Though Z_2 is count data, an exploratory analysis shows that Z_2 is a mixture of smokers and nonsmokers, and thus the Poisson is inadequate.

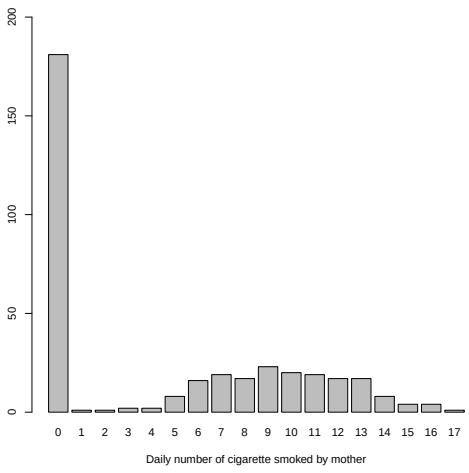

Figure 3.1: Bar plot for Z_2 in the Six-Cities data.

A Logistic Regression Example

- Instead, we use Zero-Inflated Poisson (ZIP; see Problem 1.6) to model Z_2.
- We let

$$\Pr(Z_1 = k) = p_k, \;\; k = 0, 1,$$

$$Z_2 | Z_1 = k \sim \mathsf{ZIP}(\rho_k, \lambda_k).$$

- Recall that $X \sim \mathsf{ZIP}(\rho, \lambda)$ if

$$X|(G = 0) \sim \mathsf{Poisson}(\lambda), \quad X|(G = 1) \sim 0, \quad \Pr(G = 1) = \rho.$$

- So, the whole parameter consists of $\theta := (\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$, where $\beta = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$ and $\alpha = (\rho_0, \lambda_0, \rho_1, \lambda_1, p_1)^{\mathrm{T}}$.

A Logistic Regression Example

- Under the current specification of joint distribution of (Z_1, Z_2),

$$\begin{aligned} q_{kj}(\theta) &:= \Pr(Z_1 = k, Z_2 = j|\alpha) \\ &= p_k \left(\rho_k + (1 - \rho_k)e^{-\lambda_k}\right)^{I(j=0)} \left((1 - \rho_k)e^{-\lambda_k} \frac{\lambda_k^j}{j!}\right)^{I(j>0)}, \end{aligned}$$

  where $k = 0, 1$, and $j = 0, 1, 2, \cdots$.
- Under the logistic regression model

$$\begin{aligned} s_{kj}(y; \beta) &:= \Pr(Y = y|Z_1 = k, Z_2 = j; \beta) \\ &= \frac{e^{y\beta^{\mathrm{T}} Z}}{1 + e^{(\beta^{\mathrm{T}} Z)}}, \end{aligned}$$

  where $y = 0, 1$.

A Logistic Regression Example

- There are four missing patterns on Z, denoted as $m_1, \cdots, m_4$, which are

$$m_1(Z) = \emptyset, \qquad m_2(Z) = Z_1,$$

$$m_3(Z) = Z_2, \qquad m_4(Z) = (Z_1, Z_2).$$

- So M_i may be represented by

$$M_i = \sum_{r=1}^{4} I(R_i = r) m_r.$$

- We will compute the weights

$$w_{ikj}(\theta) := E[I(Z_{i1} = k, Z_{i2} = j) | Y_i, M_i(Z_i); \theta]$$

  for these four scenarios separately.

A Logistic Regression Example

- $R_i = 1$:

$$w_{ikj}(\theta) = \frac{s_{kj}(Y_i; \beta) q_{kj}(\alpha)}{\sum_{k'=0}^{1} \sum_{j'=0}^{\infty} s_{k'j'}(Y_i; \beta) q_{k'j'}(\alpha)};$$

- $R_i = 2$:

$$w_{ikj}(\theta) = \frac{I(Z_1 = k) s_{kj}(Y_i; \beta) q_{kj}(\alpha)}{\sum_{j'=0}^{\infty} s_{kj'}(Y_i; \beta) q_{kj'}(\alpha)};$$

- $R_i = 3$:

$$w_{ikj}(\theta) = \frac{I(Z_2 = j) s_{kj}(Y_i; \beta) q_{kj}(\alpha)}{\sum_{k'=0}^{1} s_{k'j}(Y_i; \beta) q_{k'j}(\alpha)};$$

- $R_i = 4$:

$$w_{ikj}(\theta) = I(Z_1 = k, Z_2 = j).$$

- Here and in the sequel, the upper bound ∞ can be replaced with a large integer J in actual implementation. (For the current dataset, take, say, $J = 20$.)

A Logistic Regression Example

- Now, at the $(t+1)$th iteration, the M step for β can be easily done with a Newton-Raphson algorithm with

$$\dot{Q}_1(\beta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=0}^{1}\sum_{j=0}^{\infty} w_{ikj}^{(t)} \begin{pmatrix} 1 \\ k \\ j \end{pmatrix} \left(Y_i - \frac{e^{\beta_0+k\beta_1+j\beta_2}}{1+e^{\beta_0+k\beta_1+j\beta_2}} \right),$$

and

$$-\ddot{Q}_1(\beta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=0}^{1}\sum_{j=0}^{\infty} w_{ikj}^{(t)} \frac{e^{\beta_0+k\beta_1+j\beta_2}}{(1+e^{\beta_0+k\beta_1+j\beta_2})^2} \begin{pmatrix} 1 \\ k \\ j \end{pmatrix}^{\otimes 2},$$

where $w_{ikj}^{(t)} = w_{ikj}(\theta^{(t)})$.

A Logistic Regression Example

- For the M step for α, we could use a Newton-Raphson algorithm based on the joint distribution of (Z_1, Z_2).
- We could also treat G as part of the full data, in which case the M step has a closed-form solution.
- Note that when doing the M step for β, it does not matter whether we treat G as part of the full data, since by construction

$$[Y|Z_1, Z_2, G] = [Y|Z_1, Z_2].$$

So the full-data log-likelihood for the regression model does not change, giving rise to the same $Q_1(\beta|\theta^{(t)})$.

A Logistic Regression Example

- With G part of the full data, we are dealing with full-data covariate distribution

$$\Pr(Z_1 = k, G = g, Z_2 = j|\alpha) = I\Big(j = (1-g)j\Big) p_k \rho_k^g \left((1-\rho_k)e^{-\lambda_k}\frac{\lambda_k^j}{j!}\right)^{1-g},$$

  where the indicator $I\Big(j = (1-g)j\Big)$ restricts j to be zero when $g = 1$.
- After standard derivation similar to that given in the solution to Problem 1.6, we find that

$$p_k^{(t+1)} = n^{-1}\sum_{i=1}^{n}\sum_{j=0}^{\infty} w_{ikj}^{(t)}, \qquad k = 0, 1.$$

A Logistic Regression Example

- And,

$$\rho_k^{(t+1)} = \xi_k^{(t)} \frac{\sum_{i=1}^{n} w_{ik0}^{(t)}}{\sum_{i=1}^{n}\sum_{j=0}^{\infty} w_{ikj}^{(t)}},$$

$$\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{\infty} w_{ikj}^{(t)} j}{\sum_{i=1}^{n}\left(w_{ik0}^{(t)}(1-\xi_k^{(t)}) + \sum_{j=1}^{\infty} w_{ikj}^{(t)}\right)}, \qquad k = 0, 1,$$

  where

$$\xi_k^{(t)} = \frac{\rho_k^{(t)}}{\rho_k^{(t)} + (1-\rho_k^{(t)})e^{-\lambda_k^{(t)}}}.$$

- The details are relegated to the Addendum.

Contents

Generalized Linear Mixed Models

- In a longitudinal study, each experimental or observational unit is measured at baseline and repeatedly over time. Incomplete data are not unusual under such designs, as many subjects are not available to be measured at all time points.
- In addition, a subject can be missing at one follow-up time and then measured at the next, resulting in nonmonotone missing data patterns. Such data present a considerable modeling challenge for the statistician.
- When nonresponse is unrelated to the values of the missing variables, i.e., ignorable missing, software is available for analyzing unbalanced longitudinal data (SAS Proc Mixed, BMDP5V, SAS GLIMMIX Macro).

Generalized Linear Mixed Models

- These tools eliminate complete-case bias by incorporating all available information when the data are MAR.
- However, nonignorable missing data are very common in longitudinal studies.
  - In many cancer and AIDS clinical trials, the side-effects of the treatment may affect participation, and missingness can depend on the outcome as well as the treatment covariate.
  - In quality of life studies, compliance is not compulsory, and those with a poor prognosis may be more likely not to complete the questionnaire at every visit.
  - In environmental studies, geographic location or environmental factors may influence the response. Examples of nonignorable missingness can also be found in longitudinal psychiatric studies (see Molenberghs, et. al., 1997; Little and Wang, 1996).

Generalized Linear Mixed Models

- Methods for handling missing data often depend on the pattern of missingness and the mechanism that generates the missing values.
- For the present, missing outcomes will be of primary interest. So we consider the covariates to be fixed.
- To illustrate the various missingness patterns and mechanisms in a regression setting, consider a data set that consists of a vector of responses $Y_i = (Y_{i1}, \cdots, Y_{i,n_i})^{\mathrm{T}}$ that may contain missing values, and an $n_i \times p$ matrix $X_i = (X_{i1}, \cdots, X_{i,n_i})$ of covariates, where each X_{ij} is a p-dimensional vector.

Generalized Linear Mixed Models

- **Monotone Missing**: Once a subject is unobserved, they are never observed again. Thus, monotone missing data is also termed dropout.
- For example, missing values in the vector of responses, Y_i, occur as dropouts if whenever Y_{ij} is missing, so are Y_{ik}, for all $k \geq j$. Likelihoods are easier to evaluate with monotone patterns of missing data.
- **Nonmonotone Missing**: A subject is observed again after a missing value occurs.
- For example, if Y_i contains missing values, and Y_{ij} may be missing while Y_{ik} is observed, for some $k > j$. Likelihoods are more difficult to evaluate with nonmonotone patterns of missing data.

Generalized Linear Mixed Models

Random Effects Models without Missing Values

- The normal random effects model, also known as the Laird-Ware model (1982), is described as follows.
- For a given subject i with n_i repeated measurements, the outcome vector Y_i is given by

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \qquad i = 1, \cdots, n,$$

  where β is a p vector of unknown regression parameters, commonly referred to as fixed effects, Z_i is a known $n_i \times q$ matrix of covariates for the q vector of random effects b_i, and ϵ_i is an n_i vector of errors.
- The presence of b_i introduces correlations between the repeated measurements Y_{ij} within subject i
- The columns of Z_i are usually a subset of X_i, allowing for fixed effects as well as random effects.

Generalized Linear Mixed Models

Random Effects Models without Missing Values

- It is typically assumed that the ϵ_i are independent, the b_i are iid, the b_i are independent of the ϵ_i, and

$$\xi_i \sim N_{n_i}(0; \sigma^2 I_{n_i}), \qquad b_i \sim N_q(0, D),$$

  where I_{n_i} is the $n_i \times n_i$ identity matrix and $N_q(0, D)$ denotes the q-dimensional multivariate normal distribution with mean 0 and covariance matrix D.
- The positive definite matrix D is the covariance matrix of the random effects and is typically assumed to be unstructured and unknown.
- Under the model assumptions,

$$Y_i | b_i \sim N(X_i\beta + Z_i b_i, \sigma^2 I_{n_i}).$$

Generalized Linear Mixed Models

Random Effects Models without Missing Values

- Note that

$$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} \sim N\left\{ \begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} Z_i D Z_i^{\mathrm{T}} + \sigma^2 I_{n_i} & Z_i D \\ D Z_i^{\mathrm{T}} & D \end{pmatrix} \right\},$$

  so that $E[b_i | Y_i; \beta, \sigma^2, D]$ and $\mathrm{Var}[b_i | Y_i; \beta, \sigma^2, D]$ have closed form expressions (see A1.2).
- The facilitates the computing of MLE by treating the b_i as missing data and using the EM.
- The major advantage of the EM over direct maximization (by Newton-Raphson) is in fitting covariance matrices with large numbers of parameters.

Generalized Linear Mixed Models

Random Effects Models without Missing Values

- Let $\theta = (\beta, \sigma^2, D)$. The full-data log-likelihood is

$$l_n(D;\theta) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{-\frac{(Y_{ij} - \beta^{\mathrm{T}}X_{ij} - b_i^{\mathrm{T}}Z_{ij})^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2\right\}$$
$$-\sum_{i=1}^{n}\frac{b_i^{\mathrm{T}}D^{-1}b_i}{2} - \frac{n}{2}\log\det D.$$

- By inspection of full-data log-likelihood, we have that, at the $(t+1)$th iteration,

$$\beta^{(t+1)} = \left(\sum_{i=1}^{n}X_i^{\mathrm{T}}X_i\right)^{-1}\sum_{i=1}^{n}X_i^{\mathrm{T}}(Y_i - Z_i\widehat{b}_i^{(t)}),$$

Generalized Linear Mixed Models

Random Effects Models without Missing Values

$$\sigma^{2(t+1)} = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{n_i}(Y_{ij} - \beta^{\mathrm{T}}X_{ij} - \widehat{b}_i^{(t)\mathrm{T}}Z_{ij})^2 + Z_i\widehat{C}_i^{(t)}Z_i^{\mathrm{T}}\right)}{\sum_{i=1}^{n}n_i},$$

$$D^{(t+1)} = n^{-1}\sum_{i=1}^{n}\left(\widehat{b}_i^{(t)\otimes 2} + \widehat{C}_i\right),$$

where

$$\widehat{b}_i^{(t)} = E[b_i|Y_i;\theta^{(t)}], \quad \text{and } \widehat{C}_i^{(t)} = \mathrm{Var}[b_i|Y_i;\theta^{(t)}].$$

- When there are missing values in Y_i and/or X_i that are MAR, the EM algorithm can be similarly constructed by adding another layer of E step (possibly via Monte Carlo methods).
- When some components of X_i are missing, we need to postulate a model for X.

Generalized Linear Mixed Models

- The generalized linear model (GLM) with random effects, also known as the generalized linear mixed model (GLMM), is the GLM generalization of the normal linear random effects model described by Laird and Ware (1982).
- For a given individual i with $j = 1, \cdots, ni$ repeated measurements, outcome Y_{ij} is modeled as

$$p(Y_{ij}|X_i, b_i; \beta, \phi) = \exp[\phi\{Y_{ij}\gamma(\eta_{ij}) - a \circ \gamma(\eta_{ij})\} - c(\phi, Y_{ij})],$$

where Y_i is an n_i-dimensional response, ϕ is a scalar dispersion parameter, γ is the link function, $\eta_{ij} = \beta^{\mathrm{T}} X_{ij} + b_i^{\mathrm{T}} Z_{ij}$ is the linear predictor.

Generalized Linear Mixed Models

- The link is said to be the canonical link when $\gamma(x) = x$.
- For simplicity, we assume that $\phi = 1$, as in logistic and Poisson regressions.
- Again, we assume that $b_i \sim N_q(0, D)$, where D is a $q \times q$ unknown covariance matrix.
- Treating the random effects b as missing data, the full-data likelihood based on n subjects for the GLMM is given by

$$\prod_{i=1}^{n} p(Y_i, b_i|X_i; \beta, D) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(Y_{ij}|X_i, b_i; \beta)p(b_i; D).$$

And the observed-data likelihood is

$$\prod_{i=1}^{n} p(Y_i|X_i; \beta, D) = \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} p(Y_{ij}|X_i, b_i; \beta)p(b_i; D)db_i.$$

Generalized Linear Mixed Models

- Unlike the case with linear mixed model, the likelihood $\int \prod_{j=1}^{n_i} p(Y_{ij}|X_i, b_i; \beta)p(b_i; D)db_i$ does not have a closed form.
- When using the EM algorithm based on the full-data likelihood, the M step is similar to that of GLM with missing covariates, but the E step does not have closed form. So one generally has to resort to Monte Carlo EM.
- When some components of Y and/or X are missing, an additional layer of E step over Y and X needs to be added. The procedures are necessarily more complex, but conceptually standard. So is the case when Y (and/or X) is non-ignorably missing.
- The difficult part for non-ignorably missing longitudinal data lies in modeling the missing data mechanism.

Generalized Linear Mixed Models

GLMM with non-ignorable missing response

- As always, when the missing data mechanism is nonignorable, it is typical to specify a parametric model for it and incorporate it into the full data log-likelihood.
- We consider the case of GLMM with missing response.
- The missing data mechanism can be defined as the distribution of the n_i-dimensional random vector R_i, whose jth component, R_{ij}, equals 1 if Y_{ij} is missing, and 0 if Y_{ij} is observed.
- The distribution of R_i is a multinomial distribution with 2^{n_i} cell probabilities, indexed by ψ.

Generalized Linear Mixed Models

GLMM with non-ignorable missing response

- The goal is to have a full data likelihood $p(Y_i, b_i, R_i|\beta, \sigma^2, D, \psi)$. Note that since there is no missing value in X, we consider them to be fixed.
- Little (1993, 1995) identified two ways of factoring this joint distribution:
  - **Selection model**: postulate a model for $[R_i|Y_i]$, so that

$$p(Y_i, b_i, R_i|\beta, \sigma^2, D, \psi) = p(Y_i|b_i; \beta, \sigma^2)p(b_i; D)p(R_i|Y_i; \phi).$$

  - **Pattern mixture model**: postulate a model for $[Y_i|R_i, b_i]$, so that

$$p(Y_i, b_i, R_i|\beta, \sigma^2, D, \psi) = p(Y_i|R_i, b_i; \beta, \sigma^2)p(b_i; D)p(R_i; \phi),$$

    where we have assumed that $R_i \perp\!\!\!\perp b_i$.

Generalized Linear Mixed Models

GLMM with non-ignorable missing response
Selection models

- Diggle and Kenward (1994) propose a product Bernoulli model for the missing data mechanism under the selection modeling approach. That is

$$p(R_i|Y_i; \phi) = \prod_{j=1}^{N_i} \Pr(R_{ij} = 1|Y_{i1}, \cdots, Y_{ij}; \psi)^{R_{ij}} \times \left[(1 - \Pr(R_{ij} = 1|Y_{i1}, \cdots, Y_{ij}; \psi))\right]^{1-R_{ij}},$$

  where R_{ij} is modeled via a logistic regression involving all of the previous outcomes as well as the current outcome.
- This model has the form

$$\text{logit}\Pr(R_{ij} = 1|Y_{i1}, \cdots, Y_{ij}; \psi) = \psi_0 + \psi_1 Y_{i1} + \cdots + \psi_j Y_{ij}.$$

Generalized Linear Mixed Models

GLMM with non-ignorable missing response
Selection models

- The model can be extended to permit possible relationships between the missing data process and covariates X_{ij}, including time t_j. This would allow, for example, the missing response rate to depend on treatment/exposure.
- A more general multinomial missing data model which incorporates a general correlation structure can be constructed by specifying the joint distribution of $R_1 = (R_{i1}, \cdots, R_{i,n_i})$ through a sequence of one-dimensional conditional distributions discussed in §3.2 (see Ibrahim, Chen, and Lipsitz, 2001).
- But this model is not as parsimonious as the product Bernoulli model, which seems quite reasonable in practice.

Generalized Linear Mixed Models

GLMM with non-ignorable missing response
Pattern mixture models

- Pattern-mixture models are based on an alternative factorization of $p(Y_i, b_i, R_i|\beta, \sigma^2, D, \psi)$ based on $[Y_i|R_i, b_i]$, $[b_i]$, and $[R_i]$.
- Assuming $[Y_i|R_i, b_i]$ is normal, since the distribution of Y_i depends on R_i, this model implies that the marginal distribution of Y_i is a mixture of normal distributions rather than normal.
- Speficially,

$$[Y_i|R_i = r_k, b_i; \beta, \sigma^2] \sim N_{n_i}(X_i\beta_{(k)} + Z_i b_i, \sigma^2_{(k)} I_{n_i}).$$

- Therefore, the pattern mixture model is **not** consistent with the original regression model that we put forth for the full data. The interpretation of (pattern-specific) parameters is also questionable.

Generalized Linear Mixed Models

GLMM with non-ignorable missing response

- Some references in GLMM with non-ignorable missing data using selection models include Wu and Carroll (1988), Diggle and Kenward (1994), Little (1995), Ibrahim, Chen, and Lipsitz (2001), and Stubbendick and Ibrahim (2003). Approaches based on pattern-mixture models include Little (1995), Little and Wang (1996), and Hogan and Laird (1997). Troxel, Lipsitz, and Harrington (1998).
- A nice review of missing data methods in longitudinal studies is given in Ibrahim and Molenberghs (2009).

References

- Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. Applied statistics, 49-93.
- Hogan, J. W. and Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. Statistics in Medicine, 16, 239-257.
- Ibrahim, J. G. (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association, 85, 765-769,
- Ibrahim, JG, Chen, MH, Lipsitz, SR, and Herring, AH (2005). Missing Data Methods for Generalized Linear Models: A Comparative Review, Journal of the American Statistical Association, 100, 332âĂŞ346.
- Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing Responses in Generalized Linear Mixed Models When the Missing Data Mechanism Is Nonignorable, Biometrika, 88, 551-564.
- Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. Test, 18, 1-43.
- Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association, 90, 1112-1121.
- Little, R. J. A. and Wang, Y. (1996). Pattern-mixture Models for Multivariate Incomplete Data with Covariates. Biometrics, 52, 98-111.
- McCulloch, C. E. and Nelder, J. A. (1989). Generalized Linear Models. New York: Chapman and Hall.
- Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum Likelihood Methods for Nonignorable Responses and Covariates in Random Effects Models. Biometrics, 59, 1140-1150.
- Troxel, A. B., Lipsitz, S. R., and Harrington, D. P. (1998). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. Biometrika, 661-672.
- Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics, 175-188.